

Online Appendix

The Impacts of a Prototypical Home Visiting Program on Child Skills

Jin Zhou, James Heckman, Bei Liu, and Mai Lu

February 10, 2022

Contents

A	Conducting the Experiment	1
B	Baseline Comparisons	2
C	Denver II Test	7
D	Linear Model Estimates on Raw Scores	8
E	Wild Bootstrap Procedure	10
F	Data, Attrition, and Nonresponse	10
G	Measures of Interactions and Extracted Factors	14
	G.1 Measures of Interactions	14
	G.2 Factor Model of Interaction	14
H	Robustness Check for Factor Normalization	17
I	Consistency and Asymptotic Normality of Individual Factors (Wang, 2020) and a Factor Estimation Procedure from Chen, Fernández-Val, and Weidner (2021)	20
J	Fit of Estimated Models to Sample Data	20
K	Stochastic Dominance	21
L	The Monthly Age Distribution Comparison	24

A Conducting the Experiment

Field efforts included recruiting and training interviewers, hiring county-based Ph.D level project directors, and engaging support by the local government and health system. In early 2015, local home visitors were recruited and trained. At the same time, in January 2015, the baseline data were collected, which covered all of the presented households with children under 2 years of age in Huachi county. The information includes the Infant-Toddler HOME Inventory for 0-36 months, a household demographic survey, and village-level registration data. Fifty-six villages are selected into the treatment group, and fifty-five villages are in the control group. Figure A.2 shows the locations of treatment and control group villages in Huachi County.

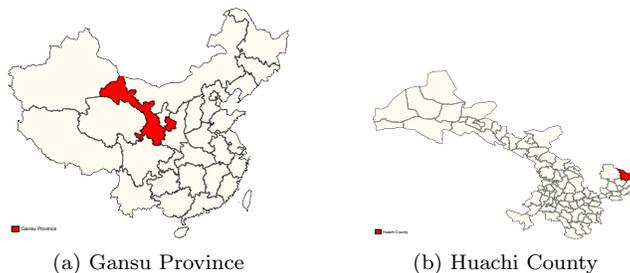


Figure A.1: The Location of Huachi County

In January 2016, there was a sample enlargement, including children born between February 9th, 2015, and April 30th, 2015, for both control and treatment groups. Newly-enrolled children were assessed on their health (height, weight, head circumferences, hemoglobin), the Denver II test, the Infant-Toddler HOME Inventory for 0-36 months, and a household demographic survey before the home visiting intervention began.

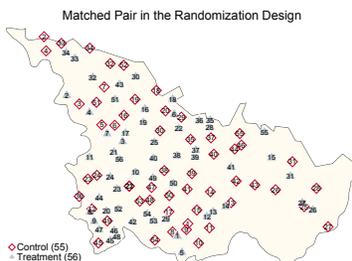


Figure A.2: The Locations of Randomized Paired Villages in Huachi County

From September 2015 to November 2017, home visitors provided weekly one-hour home visits to the treated children. Home visitors taught caregivers to encourage infants' and toddlers' play-based learning during daily family activities and routine events. Culturally adapted learning materials such as picture books, games, and toys were employed to demonstrate specific interactions.

Between September 2015 and November 2017, two rounds of follow-ups were conducted in July 2016 and July 2017, respectively, which include early childhood development assessments: anthropometrics (height, weight, head circumference, and hemoglobin), the Denver Development Screening Test (2nd ed) for 6-72 months, the Infant-Toddler HOME Inventory for 0-36 months, Early Childhood HOME (36-60 months), and a household demographic survey.

January 2015 Before home visit interventions, baseline information were collected for children with birth dates between April 1, 2013, and November 30, 2014.

September 2015 Initiation of the home visit.

January 2016 Enlarging the enrollment:¹

- Including 180 children (born between February 9, 2015, and April 30, 2015) for both treatment and control groups.
- Assessed children's health and Denver II scores before home visit intervention.

July 2016 Conducted a family survey and assessed children's health and Denver II test scores for both control and treatment groups.

¹By January 2016, the data sample included 1566 children: 247 attrited and 76 were newly enrolled. In Section F, we document how we deal with the missing data problem.

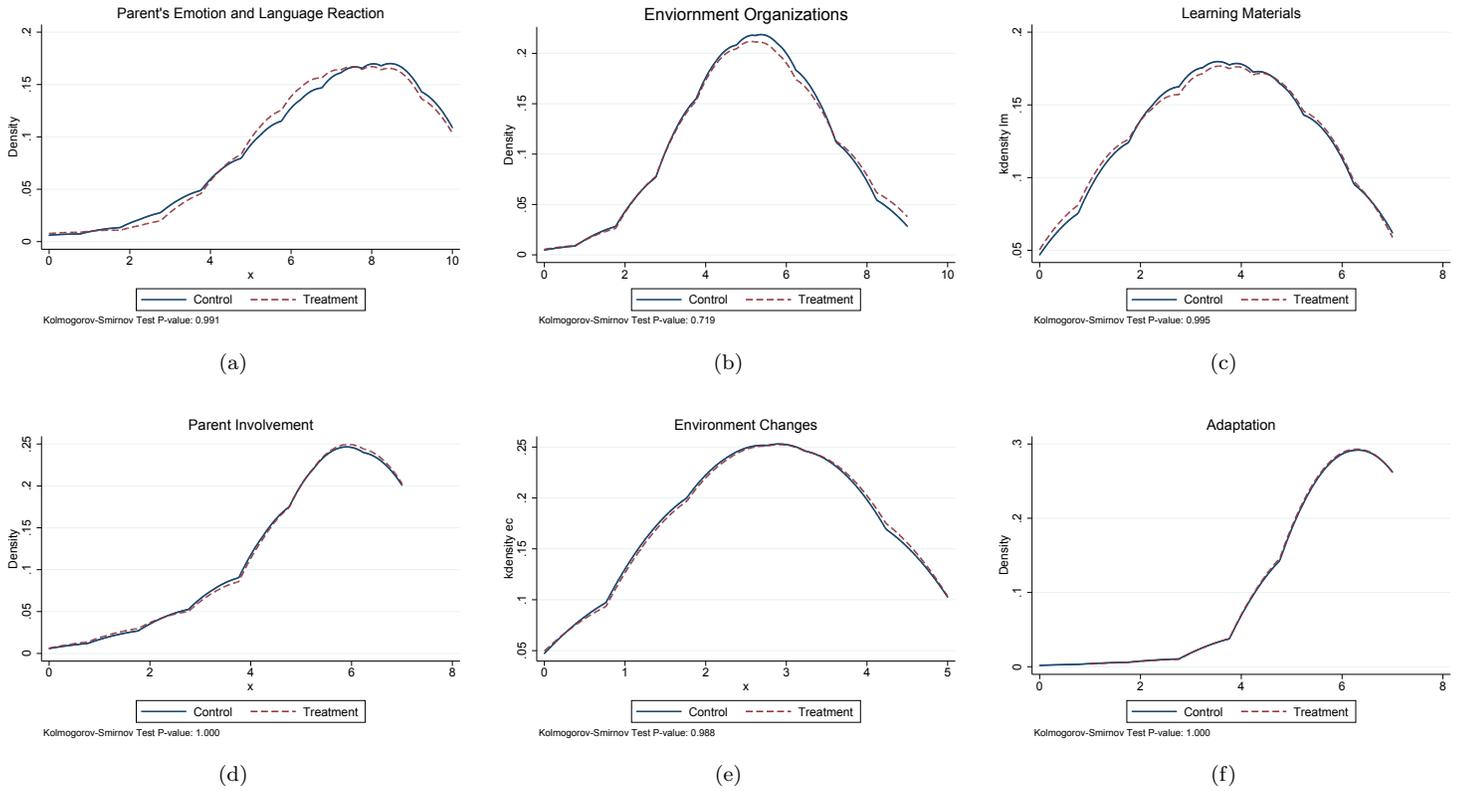


Figure B.1: Distributions of Outcomes Used in Designing Matched Village Pairs

July 2017 Conducted a family survey and assessed children’s health and Denver II scores for both control and treatment groups.

November 2017 End of home visit intervention.

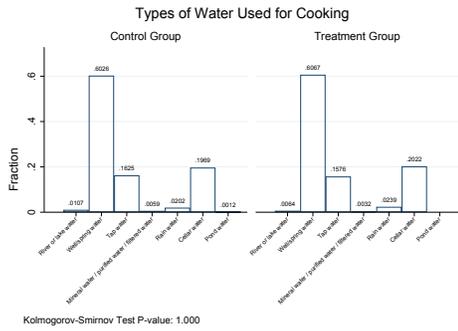
B Baseline Comparisons

In order to examine the quality of the randomization, in this section, we compare both the targeted and untargeted moments or distributions in the randomization design between the treatment and control groups. In Figure B.1, we give a comparison of the variables which are used in designing the matched pair. We can find that the control group and the treatment group have very similar distributions for the variables used for the randomization design. The Kolmogorov-Smirnov test p -values are all above 0.7, which indicates that we cannot reject the hypothesis that the distributions of the control and treatment groups are identical.

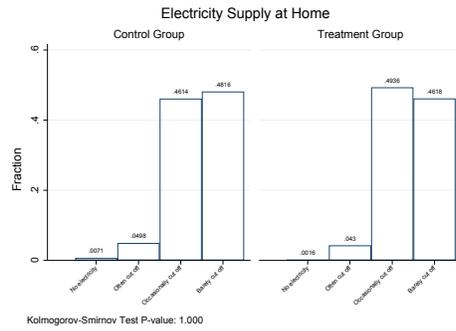
Next, we compare the variables which are not considered in the randomization estimation. We conduct this comparison for living conditions; family education levels, family structure, and economic conditions; pregnancy knowledge, pregnancy behavior, and the situations in pregnancy; children’s health and development measures; and parent-child interaction.

Figure B.2 shows that well or spring water is the main water source for cooking in both control and treatment groups. About 95% of households have stable electricity for daily life. In Huachi county, there is a kind of traditional cave dwelling housing (Yaodong). 70% of households are still living in this kind of traditional housing. Figure B.3 shows the outward appearances of Yaodong in Huachi county. The fractions of different types of durable goods owned by each household are presented in Figure B.4. All t -test p -values are above 0.05, and there is no significant difference between treatment and control group households in terms of the ownership of durable goods. Almost every household has a cell phone and television. One notable fact is that about 70% of households have at least a motorcycle and above 20% of households own their cars. The ownership of an automobile is higher than 14%, which is the car parc rate in China. Most residents are living in a mountainous area. Cars or motorcycles are important tools to connect to places outside of the village.

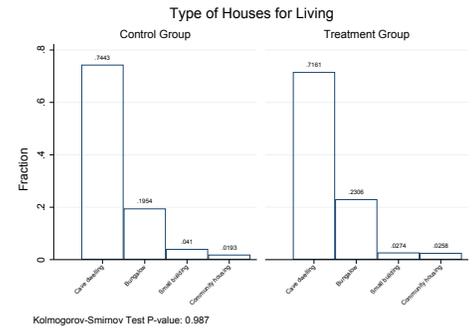
In Huachi county, the family structure is quite stable. For example, in Figure B.5 more than 98% of children’s fathers and mothers are married or cohabitating. Figure B.6 shows the education distribution for different household members. The children’s fathers and mothers have higher education levels than the grandfather-mother generation. More than 62% of children’s fathers finished at least nine years of compulsory education. About 19% of the fathers graduate from high school or above. For the children’s mothers, about 55% of them finished at least nine years of mandatory education, and about 11%



(a)



(b)

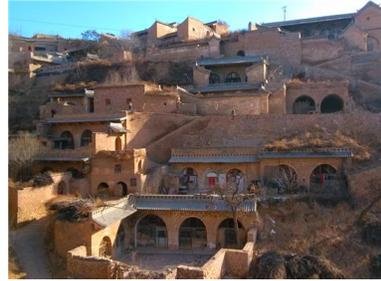


(c)

Figure B.2: Living Conditions

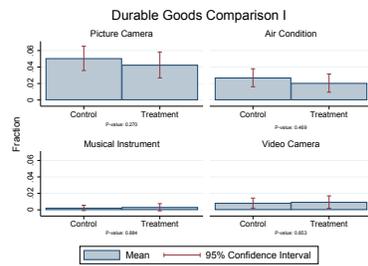


(a)

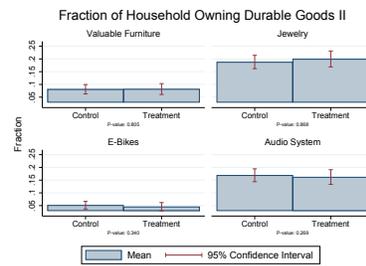


(b)

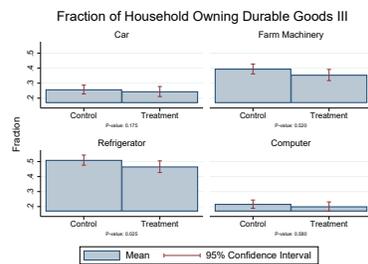
Figure B.3: Yaodong in Huachi County



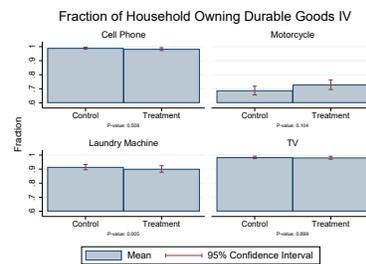
(a)



(b)



(c)



(d)

Figure B.4: Fraction of Households Owning Durable Goods

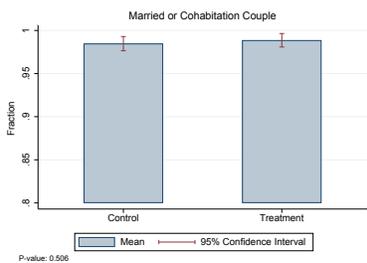


Figure B.5: Family Structure

are high school graduates or above. The fractions of mothers who have graduated primary school or middle school are similar to the fractions of the fathers. For the grandfather-mother generation, it is clear that grandmothers are less educated. More than 40% of them do not have any formal education. In general, we can see that the education distributions for different household members between the control and treatment groups are very close. The Kolmogorov-Smirnov test shows that the education level distributions are identical between the control and treatment groups.

In Table B.1, we compare household annual income and consumption by categories. Since there are multiple income sources (e.g., wage income, agriculture income, and government subsidy) for the rural households, in the table we lay out main income sources of the control and treatment groups at the baseline. The column of “ p -value” gives the statistics testing whether the mean values are different between the two groups. All p -values are greater than 5%, which means that we cannot reject the null hypothesis that the mean values of the two groups are equal to each other.

Table B.1: Consumption and Income Comparisons (Baseline)

Household Level	Control	Treatment	p -value
Agricultural Income	10284.66	7055.83	0.09
Standard Error	(1655.21)	(1912.39)	
Observations	704	515	
Government Subsidy	2780.42	2321.70	0.06
Standard Error	(220.42)	(244.40)	
Observations	751	567	
Remittance	15632.60	15969.94	0.87
Standard Error	(1233.56)	(2127.09)	
Observations	544	408	
Wage Income (After Tax)	34934.14	31255.06	0.39
Standard Error	(3102.52)	(4276.55)	
Observations	92	64	
Food Consumption	7861.47	9638.88	0.62
Standard Error	(1267.94)	(1956.49)	
Observations	703	513	
Total Consumption	42767.85	41796.35	0.92
Standard Error	(6504.43)	(10021.38)	
Observations	846	629	

p -value is calculated by bootstrapping and clustering at the level of the randomized paired villages.

Figures B.7-B.9 give a summary of the knowledge of pregnancy and the performance in pregnancy. Figure B.7 provides the comparison of pregnancy knowledge between the control and treatment groups. In general, control group individuals have greater knowledge of pregnancy but we cannot reject the null that both groups have equal knowledge of pregnancy. Figure B.8 shows the pregnancy behaviors which would affect child health outcomes. We cannot find significant difference between the control and treatment groups. Almost no mother smokes or drinks during her pregnancy. More than 80% of the mothers had prenatal check experiences, and 60% of the mothers had prenatal checks in the first three months of pregnancy. Also, we cannot find significant differences in the health conditions during pregnancy between the control and treatment groups in Figure B.9.

Table B.2 shows the comparison of Child Development Measures between the control and treatment groups at the baseline. Although we have fewer than 150 children’s Denver test information at the baseline, for these children, we find that there is no significant difference between the two groups for different types of skills. Also, Table B.2 documents that the birth weight and height are very close between the two groups.

Figure B.10 gives the duration of breastfeeding and the time of introducing complementary food to the infants. The distributions are very close for both the control and treatment groups. From Figure B.11, we can find that the changes in parent-child interactions are related to the children’s age. Also, there are no significant differences in both parent-child interactions between the two groups.

From the above comparisons, there are no significant differences in either target outcomes or the non-target variables between the control and treatment groups. In general, the randomization design works well in selecting matched pair villages.

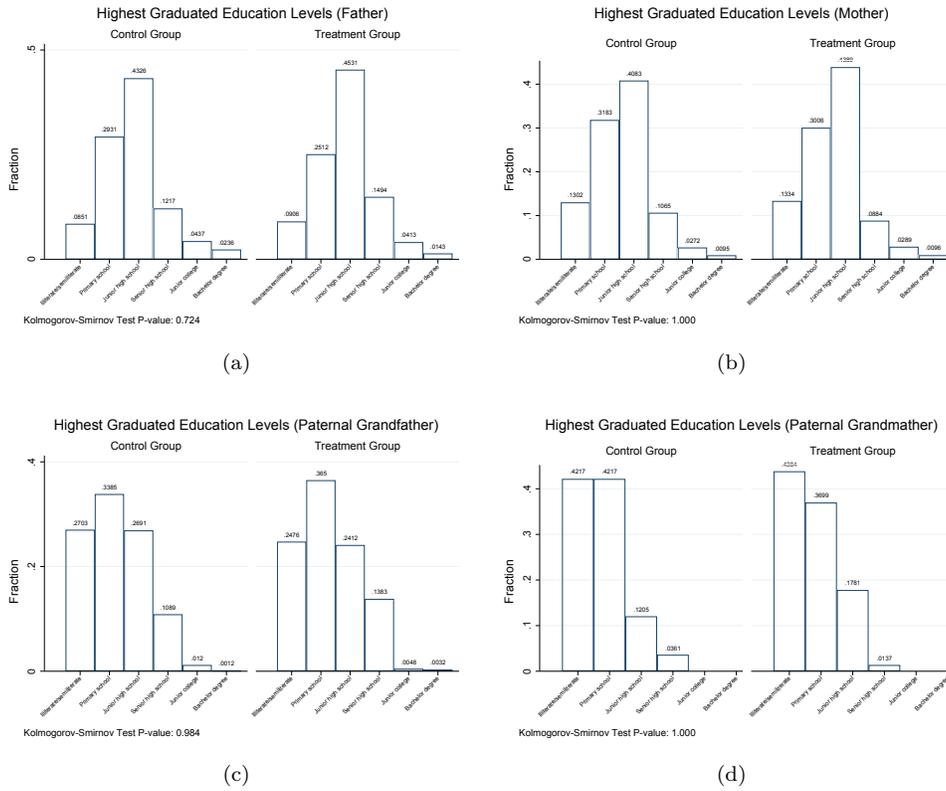


Figure B.6: Family Member Education Levels

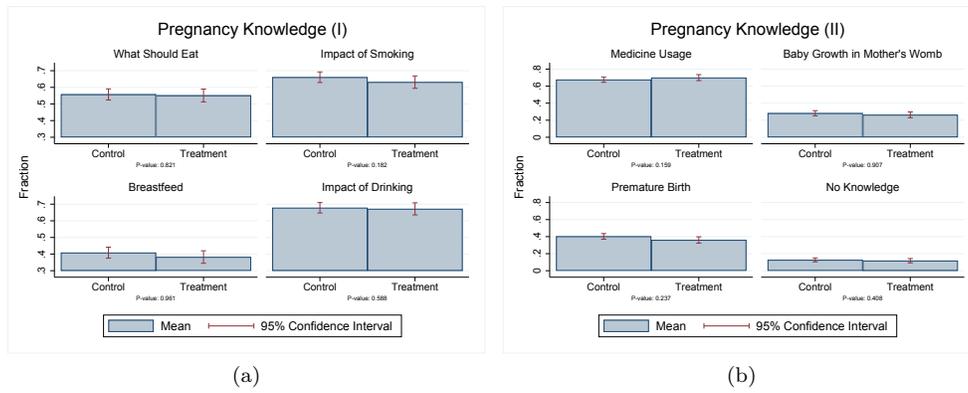


Figure B.7: Pregnancy Knowledge

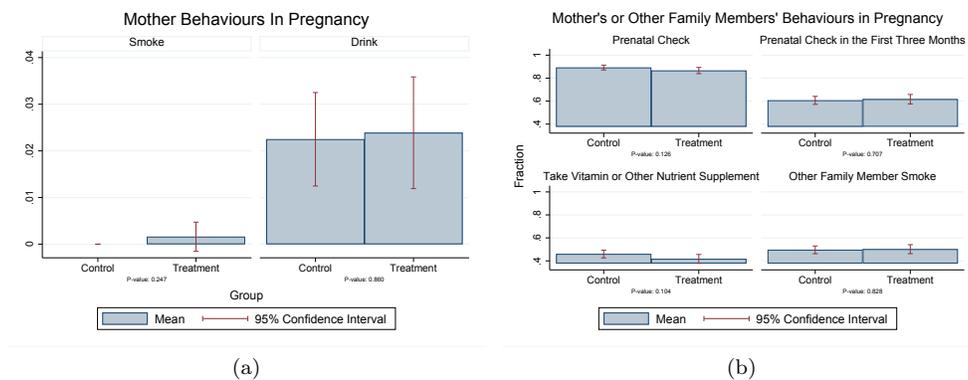


Figure B.8: Pregnancy Behavior

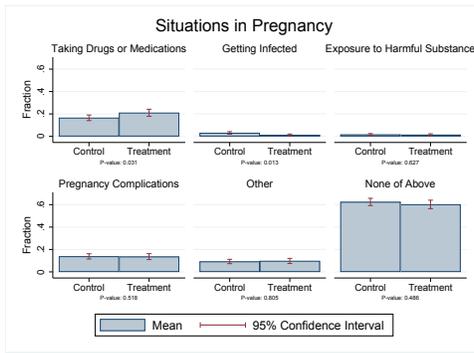
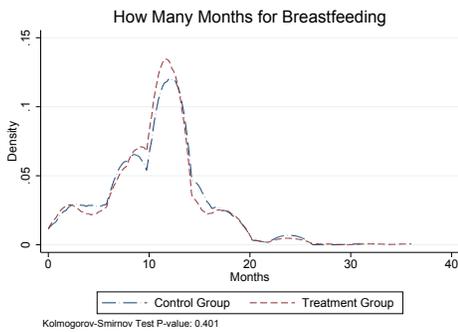
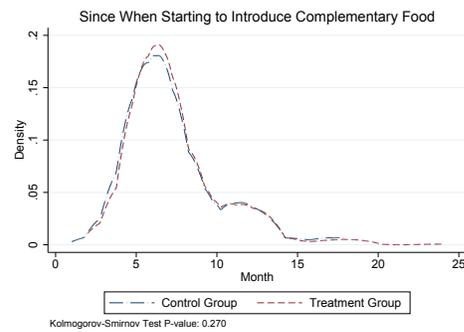


Figure B.9: Situations in Pregnancy

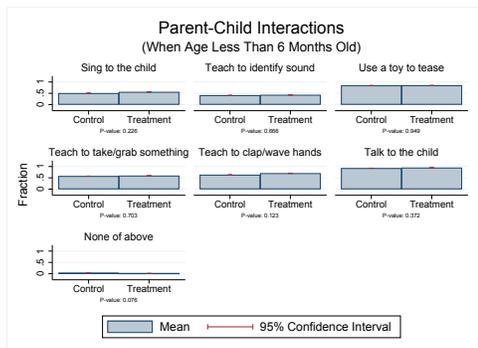


(a)

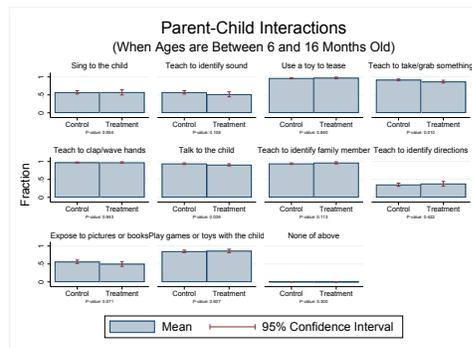


(b)

Figure B.10: Breastfeeding Behavior



(a)



(b)

Figure B.11: Parent-Child Interaction

Table B.2: Child Development Measures (Baseline)

	Control	Treatment	<i>p</i> -value
Denver Test			
Personal/Social	7.27	7.15	0.52
Standard Error	(0.12)	(0.18)	
Fine Motor Adaptive	12.27	12.23	0.81
Standard Error	(0.11)	(0.16)	
Language	11.14	11.08	0.75
Standard Error	(0.13)	(0.18)	
Gross Motor	11.56	11.09	0.02
Standard Error	(0.15)	(0.20)	
Total	42.24	41.56	0.17
Standard Error	(0.37)	(0.49)	
Observations	63	73	
Birth Weight	3.18	3.23	0.03
Standard Error	(.01)	(.02)	
Observations	765	687	
Birth Height	49.83	49.95	0.31
Standard Error	(0.08)	(0.12)	
Observations	675	643	
Monthly age at the baseline	11.39	9.63	0.00
Standard Error	(0.18)	(0.25)	
Observations	852	726	

p-value is calculated by bootstrapping at the level of the randomized paired villages.

C Denver II Test

In Figures C.1 and C.2, we present the Denver test implemented during the intervention for both English and Chinese versions. The items are over 99% consistent between the two versions.

DENVER II

DDM, INC. 1-800-419-4729
CATALOG #2115

Examiner:
Date:

Name:
Birthdate:
ID No.:

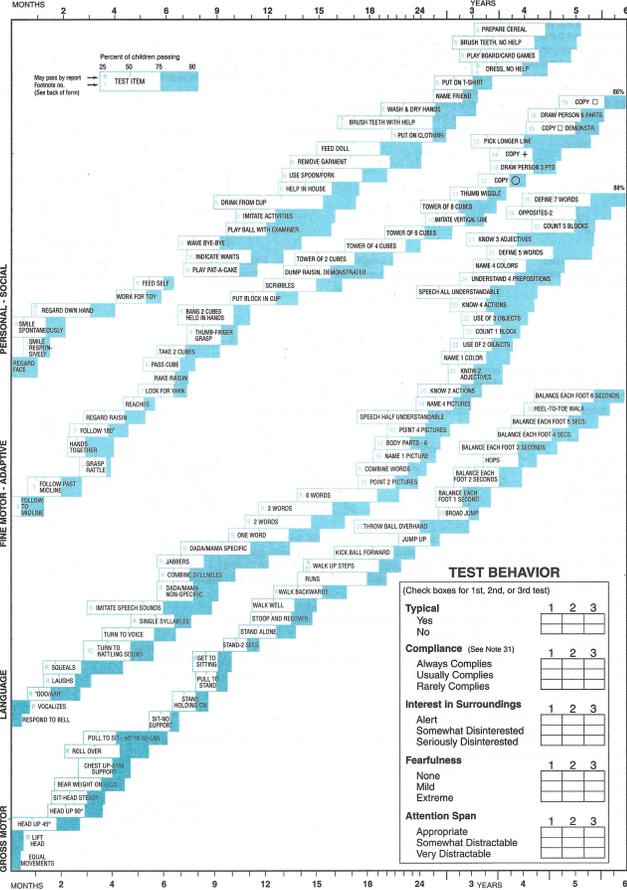


Figure C.1: English Denver Test

上海市小儿发育筛查量表 II

姓名: 检查日期: 20 年 月 日 检查者:
性别: 出生日期: 年 月 日 区号:
民族: 实际年龄: 编号:

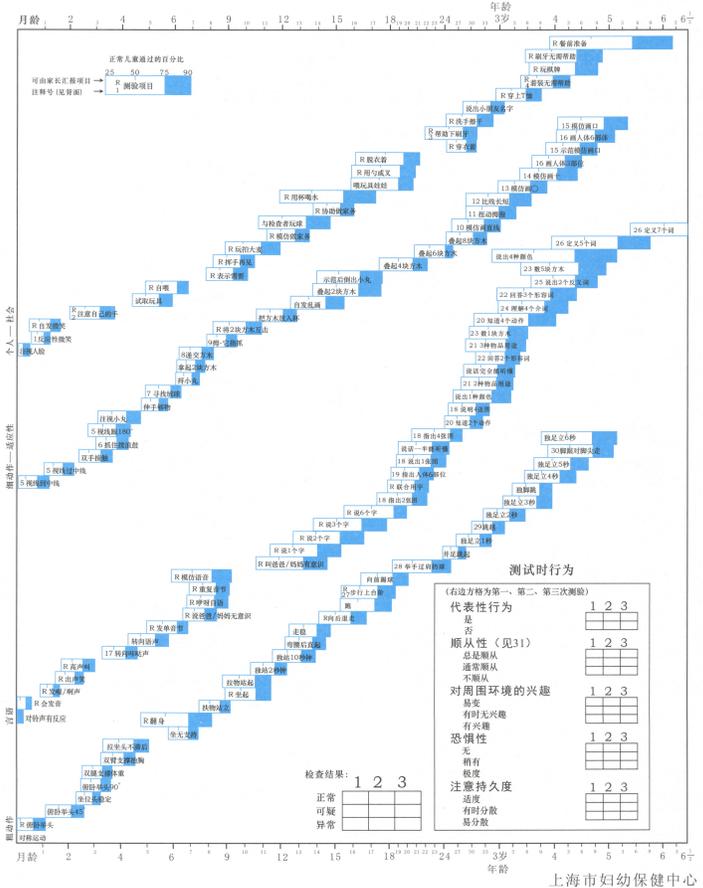


Figure C.2: Chinese Denver Test

D Linear Model Estimates on Raw Scores

Table D.1: Treatment Effects on Total Scores

	(1) All	(2) All	(3) Children ≤ 2 Yrs at Enrollment	(4) All	(5) Children ≤ 2 Yrs at Enrollment
Midline					
Language and Cognitive	0.533*** [0.162, 0.895]	0.569*** [0.161, 0.969]	0.689*** [0.299, 1.090]	0.634*** [0.234, 1.036]	0.754*** [0.347, 1.173]
Fine Motor	0.064 [-0.104, 0.233]	0.064 [-0.075, 0.412]	0.200 [-0.033, 0.444]	0.195 [-0.052, 0.467]	0.228 [-0.014, .488]
Social-Emotional	0.206** [0.044, 0.372]	0.274*** [0.094, 0.452]	0.259*** [0.073, 0.453]	0.285*** [0.115, 0.463]	0.271*** [0.067, 0.477]
Gross Motor	-0.140 [-0.391, 0.110]	-0.121 [-0.398, 0.156]	-0.009 [-0.277, 0.276]	-0.119 [-0.391, 0.148]	-0.031 [-0.295, 0.243]
Endline					
Language and Cognitive	1.031*** [0.599, 1.472]	0.966*** [0.509, 1.427]	1.172*** [0.735, 1.591]	1.041*** [0.601, 1.489]	1.247*** [0.813, 1.687]
Fine Motor	0.224 [-0.006, 0.457]	0.205 [-0.021, 0.424]	0.232** [0.023, 0.454]	0.238* [0.009, 0.472]	0.265** [0.019, 0.530]
Social-Emotional	-0.133 [-0.299, 0.038]	-0.159 [-0.342, 0.022]	-0.093 [-0.260, 0.071]	-0.136 [-0.319, 0.051]	-0.066 [-0.240, 0.109]
Gross Motor	0.085 [-0.244, 0.422]	0.106 [-0.184, 0.405]	0.101 [-0.190, 0.396]	0.112 [-0.174, 0.401]	0.122 [-0.166, 0.405]
Pre-treatment Covariates	No	No	No	Yes	Yes
IPW	No	Yes	Yes	Yes	Yes

1. The 95% confidence intervals in parentheses are constructed by wild bootstrap clustered at the village level.

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

Table D.2: Treatment Effects on Raw Scores

(Female)					
	(1)	(2)	(3)	(4)	(5)
	All	All	Children \leq 2 Yrs at Enrollment	All	Children \leq 2 Yrs at Enrollment
Midline					
Language and Cognitive	0.368	0.365	0.479**	0.388	0.518**
	[-0.089, 0.813]	[-0.105, 0.842]	[0.020, 0.939]	[-0.073, 0.882]	[0.100, 0.996]
Fine Motor	0.135	0.132	0.183	0.156	0.192
	[-0.110, 0.379]	[-0.113, 0.369]	[-0.0337, 0.396]	[-0.103, 0.418]	[-0.030, 0.411]
Social-Emotional	0.352**	0.348**	0.368***	0.382***	0.399***
	[0.098, 0.617]	[0.101, 0.584]	[0.125, 0.609]	[0.126, 0.640]	[0.132, 0.663]
Gross Motor	-0.043	-0.078	-0.034	-0.069	-0.055
	[-0.363, 0.284]	[-0.421, 0.276]	[-0.372, 0.298]	[-0.427, 0.303]	[-0.409, 0.318]
Endline					
Language and Cognitive	0.775*	0.827*	0.859*	0.869*	0.894*
	[0.118, 1.611]	[0.022, 1.618]	[0.007, 1.683]	[0.004, 1.728]	[0.073, 1.749]
Fine Motor	0.350	0.313	0.348	0.347	0.369
	[-0.044, 0.790]	[-0.043, 0.705]	[-0.004, 0.698]	[-0.066, 0.797]	[-0.040, 0.807]
Social-Emotional	-0.147	-0.169	-0.146	-0.164	-0.148
	[-0.339, 0.045]	[-0.363, 0.033]	[-0.363, 0.066]	[-0.333, 0.011]	[-0.325, 0.036]
Gross Motor	0.167	0.209	0.236	0.208	0.223
	[-0.434, 0.724]	[-0.346, 0.779]	[-0.347, 0.804]	[-0.321, 0.728]	[-0.319, 0.739]

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

Table D.3: Treatment Effects on Raw Scores

(Male)					
	(1)	(2)	(3)	(4)	(5)
	All	All	Children \leq 2 Yrs at Enrollment	All	Children \leq 2 Yrs at Enrollment
Midline					
Language and Cognitive	0.716***	0.807***	0.905***	0.853***	0.917***
	[0.204, 1.250]	[0.245, 1.378]	[0.354, 1.467]	[.306, 1.388]	[0.297, 1.520]
Fine Motor	0.046	0.193	0.208	0.211	0.231
	[-0.198, 0.304]	[-0.135, 0.551]	[-0.101, 0.539]	[-0.117, 0.590]	[-0.096, 0.604]
Social-Emotional	0.162	0.240	0.212	0.229	0.188
	[-0.083, 0.411]	[-0.012, 0.499]	[-0.071, 0.485]	[-0.044, 0.500]	[-0.101, 0.480]
Gross Motor	-0.154	-0.086	0.041	-0.126	-0.020
	[-0.437, 0.123]	[-0.447, 0.271]	[-0.277, 0.368]	[-0.426, 0.193]	[-0.306, 0.289]
Endline					
Language and Cognitive	1.198***	0.948**	1.273***	1.037***	1.376***
	[0.548, 1.822]	[0.233, 1.635]	[0.724, 1.837]	[0.375, 1.730]	[0.766, 1.980]
Fine Motor	0.138	0.111	0.128	0.108	0.124
	[-0.118, 0.395]	[-0.151, 0.391]	[-0.132, 0.386]	[-0.146, 0.368]	[-0.136, 0.398]
Social-Emotional	-0.146	-0.194	-0.095	-0.181	-0.089
	[-0.391, 0.115]	[-0.456, 0.065]	[-0.315, 0.144]	[-0.479, 0.121]	[-0.369, 0.216]
Gross Motor	-0.060	-0.067	-0.066	-0.077	-0.059
	[-0.289, 0.171]	[-0.295, 0.168]	[-0.297, 0.156]	[-0.283, 0.134]	[-0.269, 0.161]

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

Figure D.1 shows that the residuals from these regressions are at best weakly correlated across villages.

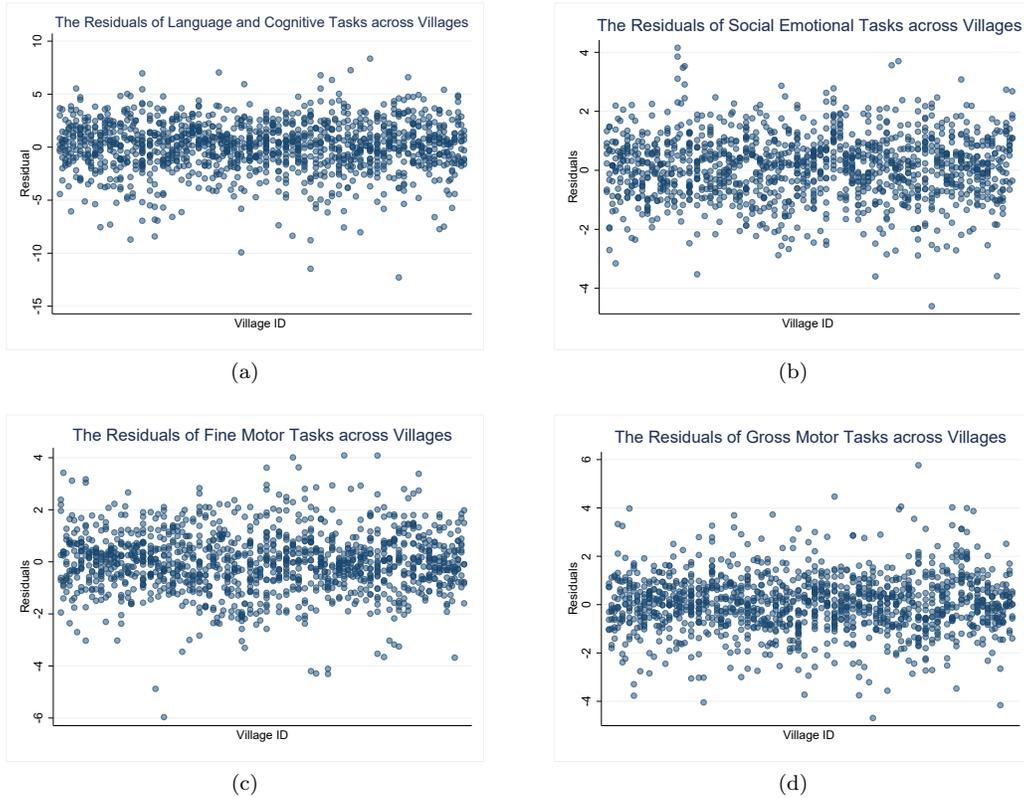


Figure D.1: Test of Residual Independence across Villages

E Wild Bootstrap Procedure

The [Cameron, Gelbach, and Miller \(2008\)](#) procedure:

1. From OLS estimation on the original sample, we obtain the estimates $\hat{\beta}$ and the CRVE $\hat{\Omega}$. Also, based on the null hypothesis $\mathbf{a}'\beta = \mathbf{0}$, we reestimate the model to obtain restricted estimates $\tilde{\beta}$ and residuals $\tilde{\mathbf{u}}$, calculate the cluster robust t statistic t_o
2. Do B iterations of this step. On the b th iteration:
 - (a) Form a sample of V clusters $(\hat{\mathbf{y}}_1^*, \mathbf{X}_1, \dots, \hat{\mathbf{y}}_V^*, \mathbf{X}_V)$ by the following method. For each cluster v , $\mathbf{u}_v^{*b} = w_v^{*b} \tilde{\mathbf{u}}_v$ and the w_v^{*b} are independent realizations of an auxiliary random variable w^* with zero mean and unit variance; then form $\mathbf{y}^{*b} = \mathbf{X}'\tilde{\beta} + \mathbf{u}_v^{*b}$
 - (b) Calculate the bootstrap estimates $\hat{\beta}^{*b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}^{*b}$ and the bootstrap covariance matrix and the bootstrap t -statistic t^{*b}
3. Reject H_0 at the level α if and only if $t_o < t_{[\alpha/2]}^*$ or $t_o > t_{[1-\alpha/2]}^*$

F Data, Attrition, and Nonresponse

This section documents data collection procedures, data attrition problems, and how we address data attrition problems. In January 2015, CDRF collected baseline information in Huachi county; 1,566 children were presented at that time. The RCT design was conducted based on the 1,566 children's survey information and village level administrative data, in which 796 children are in the treated villages and 770 children are in the control villages.

In September 2015, the home visiting intervention started. There was an eight-month gap between the baseline data collection and the first home visit interventions. The local field team made two modifications to the original protocol before they started the first home visit. The first modification is that they included 76 new children in the intervention who were not surveyed in January 2015 but were eligible in September 2015 and, in addition, they excluded most of the children who were older than two years old in September 2015 (about 150 children).² The second modification is that they excluded children

²For these 76 children, the field team collected their baseline information at the midline annual evaluation based on the parents retrospective responses.

with urban hukou (around 90). Therefore, after the two adjustments, in September 2015, the sample size was 1395, including 634 children in the treatment group and 761 children in the control group. In January 2016, 180 children from the younger cohort were added: 89 in the treatment group and 91 in the control group.³ The data we use in our analysis include 1567 observations: 1395 of which were tracked since January 2015, plus 172 children for whom the baseline was January 2016 (8 children are missing from 180 samples). Finally, the sample we use includes 1,567 children, of which 715 are in the treatment group and 852 are in the control group.⁴

Table F.1 summarizes the sample created before the first intervention. The two main modifications targeted only the treatment group, hence most modifications came from the treatment group children. Since there were newly enrolled and also excluded children during this process, we examine the baseline comparison (the final sample with 1567 children) between the control and treated group children in Section B, and find that there is no significant difference between treatment and control groups. When the field group started the home visits, there were 715 children on their name list. Among these 715 children, 705 children participated in the home visits (i.e., compliance rate is above 98%).

Table F.1: Huachi County Data Sample Before the Intervention

(January 2015-January 2016)			
	Total	Treatment	Control
Baseline (January 2015)	1566	796	770
Adjustment in September 2015			
Adjustment 1 (include the children not surveyed in January 2015)	76	76	0
Adjustment 2 (exclude old and urban hukou children)	-247	-238	-9
New Enrollment in January 2016	180	89	91
Missing	-8	-8	0
Final Sample	1567	715	852

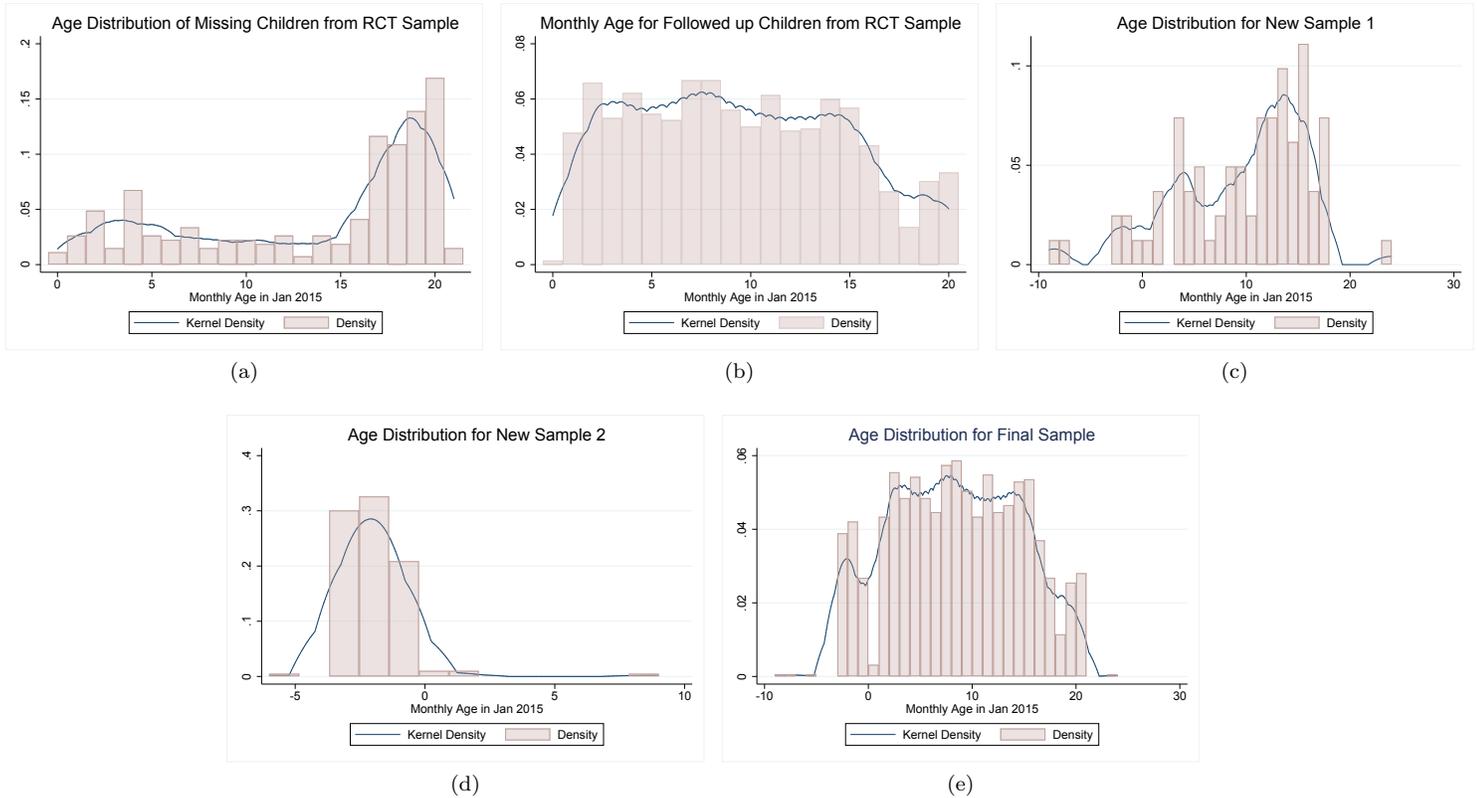


Figure F.1: Age Distribution for the Samples before February 2016

To have robust estimates, in the estimation, we also account for the sample adjustment in September 2015. According to the timeline, there were three stages at which data attrition occurred. The first stage is data attrition before February 2016. Before February 2016, 1,822 children were enrolled in the program, which included 1566 from January 2015, 76 from September 2015, and 180 from January 2016. After January 2016, 245 children were not followed up with. For the home visiting intervention, the final list was based on the 715 children in the treatment group previously discussed, of whom 705

³For these new children, they are a younger cohort from Huachi county and also take the Denver test assessment.

⁴Here, the reason why the final sample is 1,567 and not 1,566 is that we find two children share the same ID in the Denver test. We also find both children's information in the weekly home visit records. Therefore, we include the additional child.

had at least one home visit (close to 99% of children in the treatment group had been treated). By July 2017, the average number of home visits was 74.

Since both the midline and endline child development assessments were conducted in a short time window (e.g., two weeks), and the annual review assessments were conducted in the town center hospital, data attrition appeared at both rounds of assessments. Table F.1 shows the decomposition of the 1567 children in the followed-up sample. At the midline, there were 1301 children who attended the Denver test (i.e., 636 in the treatment group, and 633 in the control group); for family survey information, information on more than 1,430 children was recorded. At the endline, there were 1,073 children who attended the Denver test examination (i.e., 529 in the treatment group and 544 in the control group). The family survey was conducted for 1,189 children (i.e., 569 in the treatment group and 620 in the control group).⁵

We use the inverse probability weighting method to address missing data problems (Tsiatis, 2006). The IPW estimator solves the following minimization problem:

$$\min_{\beta} \sum_{i=1}^N \left(\frac{s_i}{\hat{p}(z_i)} \right) (y_i - x_i' \beta)^2$$

where s_i is the indicator if we can use observation i , $P(z_i)$ is the propensity score of observation i being observed. z_i is a vector of baseline variables which are always observed for everyone. Here, we mainly focus on the estimation of propensity scores at different stages. We present the distribution of the propensity score of the samples with missing data in Figure F.2 and compare them with the distribution in the observed outcome samples. In the figures, we show the distribution of the probabilities of being missing at three stages for the observed samples and missing data samples. Our model performs very well in predicting missingness: for observed outcomes samples, the probability of missing is low. For most of them the probability is less than 0.2. For the samples with missing outcome data, the propensity scores are close to being uniformly distributed. For both rounds, the estimated propensity scores are far away from 1 which means we do not need to trim the data. We thus avoid the inconsistency due to data trimming (Maasoumi and Wang, 2019).

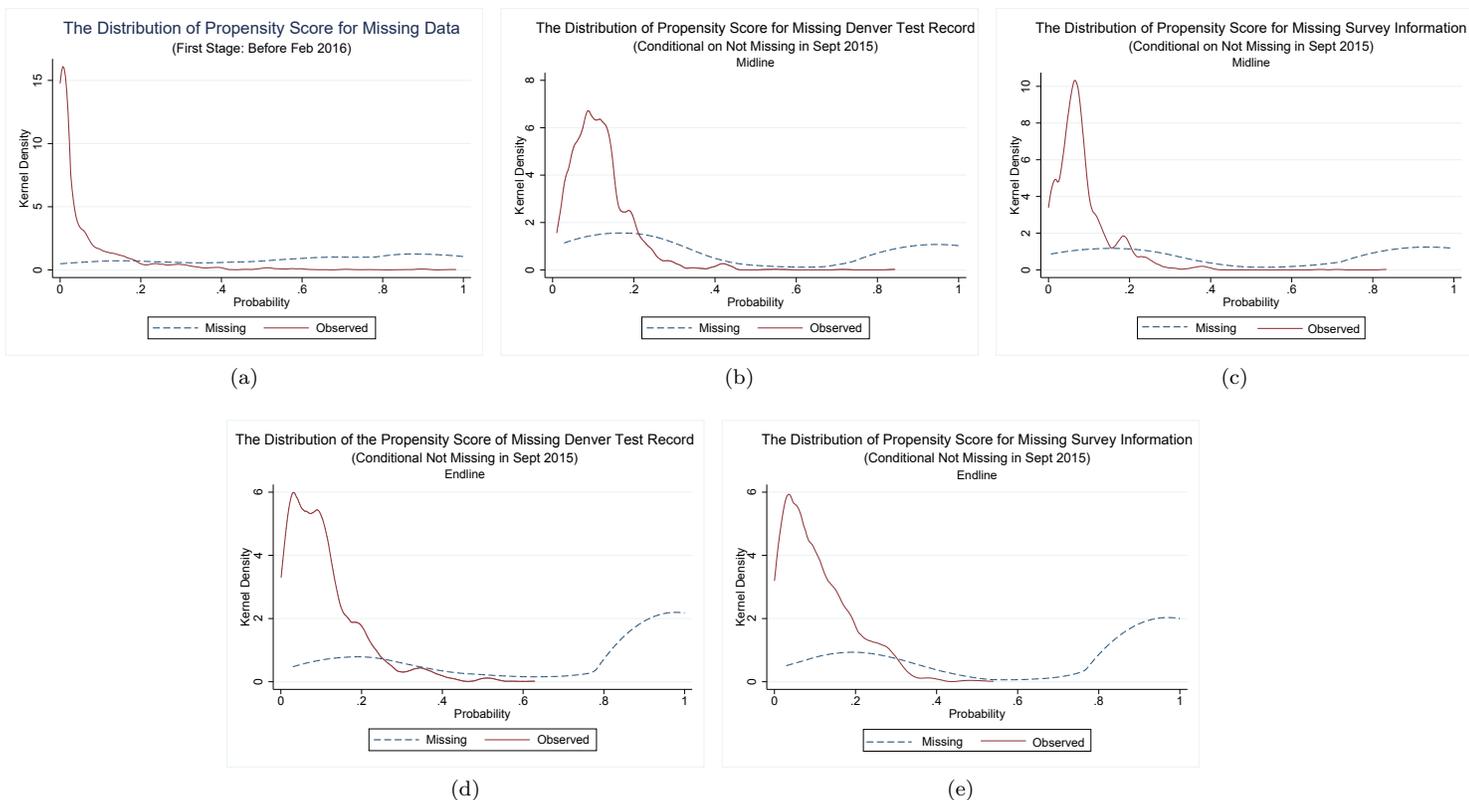


Figure F.2: Missing Data Propensity Score Distributions

⁵The data missing in the two year annual evaluations can be treated as independent events (e.g., there is no evidence to show being missing from the July 2016 evaluation is correlated with being missing from July 2017 evaluation).

Table F.2: Propensity Score for Missing Data

	First Stage	Second Stage		Third Stage	
	Miss Before January 2016	Miss Denver At Midline	Miss Survey At Midline	Miss Denver At Endline	Miss Survey At Endline
Older Than 24 Months (In September 2015)	1.5674*** (0.2163)	0.0568 (0.1491)	0.0654 (0.1548)	0.2484+ (0.1361)	0.1042 (0.1300)
Monthly Age (In September 2015)	0.0341* (0.0160)	0.0135 (0.0087)	0.0151+ (0.0092)	-0.0084 (0.0076)	0.0012 (0.0072)
The population of the Village	0.0011 (0.0008)	-0.0004 (0.0003)	-0.0003 (0.0003)	0.0003 (0.0003)	0.0005+ (0.0003)
Number of Households in the Village	-0.0001 (0.0010)	-0.0000 (0.0004)	0.0002 (0.0005)	0.0001 (0.0004)	0.0004 (0.0004)
Size of Working Population	-0.0042** (0.0016)	0.0004 (0.0006)	0.0006 (0.0006)	-0.0010+ (0.0005)	-0.0008 (0.0005)
Poor Village or Not	1.3732** (0.5066)	0.1078 (0.2058)	0.2037 (0.2129)	-0.0161 (0.1833)	0.0833 (0.1702)
Number of Persons Receiving Social Welfare	-0.0051** (0.0017)	0.0006 (0.0008)	-0.0005 (0.0009)	0.0025*** (0.0007)	0.0017** (0.0007)
Mean Years of Schooling among Villagers	-0.5857* (0.2714)	0.1179 (0.1230)	0.1485 (0.1218)	-0.1169 (0.1083)	-0.1114 (0.1020)
Fraction of Interviewed Children Who Are Left-Behind	1.1623 (2.6825)	-0.4473 (0.9716)	0.7090 (0.9408)	-1.0534 (0.8324)	-1.3794+ (0.7728)
HOME - Sum of Warmth/Responsiveness Items	0.0253 (0.0410)	0.0247 (0.0271)	0.0153 (0.0284)	0.0049 (0.0239)	-0.0063 (0.0226)
HOME - Sum of Verbal Skills Items	0.1647+ (0.0908)	0.0661 (0.0587)	0.0992 (0.0626)	-0.0273 (0.0498)	-0.0643 (0.0478)
HOME - Sum of Harshness/Discipline Items	-0.0201 (0.1206)	0.0399 (0.0863)	0.0448 (0.0882)	-0.2120* (0.0876)	-0.1247+ (0.0778)
HOME - Sum of Stimulation/Teaching Items	0.0046 (0.0263)	-0.0182 (0.0181)	-0.0206 (0.0189)	-0.0179 (0.0161)	-0.0060 (0.0154)
HOME - Dum of Outings Items	0.1590** (0.0614)	0.0305 (0.0430)	0.0323 (0.0448)	-0.0356 (0.0383)	-0.0311 (0.0362)
Fraction of Children Taking Nutrition Package	-0.3803* (0.1533)	-0.2581* (0.1011)	-0.1977+ (0.1063)	-0.0393 (0.0887)	-0.1036 (0.0842)
Fraction of Children Taking Nutrition Package without Interruption	-0.5557* (0.2332)	-0.0612 (0.1340)	-0.0765 (0.1401)	0.0259 (0.1144)	0.0342 (0.1091)
Number of Eligible Kids Living at Home in Interviewed Households in This Village	0.8225*** (0.1450)	-0.0958* (0.0416)	-0.0681 (0.0433)	-0.0574 (0.0353)	-0.0614+ (0.0337)
Fraction of Parents Willing to Participate in This Village	0.3958 (0.3759)	0.0042 (0.2008)	-0.0963 (0.2029)	-0.1754 (0.1720)	-0.0485 (0.1707)
Number of Eligible Kids in Households That Would Be Willing to Participate	-0.7809*** (0.1464)	0.0992* (0.0454)	0.0723 (0.0471)	0.0592 (0.0381)	0.0508 (0.0364)
Fraction of Interviewed Households Planning to Migrate with the Child	0.1123 (0.2381)	0.1415 (0.1565)	0.1828 (0.1613)	0.1495 (0.1415)	0.0961 (0.1374)
Distance between Home Visitor's Home and the Village	0.2559*** (0.0666)	0.0126 (0.0105)	0.0071 (0.0057)	-0.0013 (0.0063)	-0.0046 (0.0064)
Whether Living in Chengguan Village	0.6431** (0.1975)	-0.6649* (0.2681)	-0.9191** (0.3190)	0.1176 (0.1898)	0.1962 (0.1809)
Family Migrate out of County	2.5438*** (0.3410)	1.6588* (0.6442)	1.7525** (0.6755)		
Refuse Home Visit	2.8099*** (0.3545)	-5.0985 (167.8907)	-5.1383 (166.4867)		
Refuse Home Visit in September 2015		6.3453 (167.8891)	6.3040 (166.4851)		
Constant	-4.1707 (2.5367)	-2.2239* (1.0230)	-2.9191** (0.9568)	0.7335 (0.9001)	0.9051 (0.8452)
Observations	1823	1576	1576	1576	1576

Standard errors in parentheses

+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

G Measures of Interactions and Extracted Factors

G.1 Measures of Interactions

China REACH collects weekly records of child performance on lessons that measure child development during the weekly home visit intervention. The supervisors record home visitor, parent, and child interaction activities at least once per month, making it possible for us to examine their impacts. These measures are only recorded for the treatment group. We exploit variation within the treatment group, which, as we document below, is substantial.

We have detailed measures to evaluate the interaction quality between home visitors and caregivers and the visited children per visit. These observation-based measures were recorded by the program supervisors who randomly visited each household at least once per month at randomly selected times. During the home visit, the program supervisor evaluated the home visit’s quality in three dimensions: the quality of the home visitor’s teaching ability, the interaction quality between home visitor and caregiver; and the interaction quality between the home visitor and the child. We use the following measure as the teaching ability: “1. Does the home visitor bring the curriculum to the household? 2. Does the home visitor properly use the curriculum? 3. Has the home visitor prepared the home visit in advance? 4. Has the home visitor chosen the teaching materials and tasks which are suitable to the child’s age? 5. Does the home visitor focus on language development? 6. During the home visit, what is the home visitor’s attitude?” The measures we use for the interaction quality are listed in Table G.1.

Table G.1: Observed Interactions

Between Home Visitor and Caregiver	
Has the home visitor explained the task content and lesson target to the caregiver?	
Has the home visitor shown the lessons and given examples to the caregiver?	
Does the home visitor ask the caregiver to play the lessons with the child alone?	
Does the caregiver ask the home visitor about lessons in the next week?	
Has the home visitor listened to the caregiver?	
Has the home visitor answered the caregiver’s questions?	
Has the home visitor asked for the caregiver’s opinions?	
Does the home visitor encourage and help the caregiver?	
Is the relationship between the home visitor and caregiver friendly, understandable, and cooperated?	
Has the home visitor discuss with a caregiver or other persons about the content not related to the home visiting?	
Between Home Visitor and Child	
Has the home visitor shown the lessons and given examples to the child?	
Has the home visitor explained the lesson to the child?	
Does the home visitor listen to the child and respond to the child’s voice or action?	
Does the home visitor praise the child when the child tries to master one task?	
Does the home visitor use language to communicate with the child when the child is completing the lessons?	
Does the home visitor give the child enough time to explore the materials and finish the lessons?	
Is the relationship between the home visitor, and the child friendly, understandable, and cooperative?	

Note: The interaction quality measures are recorded by the supervisor of the program at least once per month.

Table G.2: Measure Variances

Measure		Teaching Ability	Home Visitor-Caregiver Interaction	Home Visitor-Child Interaction
Midline	Standard Deviation (σ)	0.410	0.696	1.032
	Coefficient of Variation ($\frac{\sigma}{M}$)	0.324	0.260	0.189
Endline	Standard Deviation (σ)	0.400	0.701	1.002
	Coefficient of Variation ($\frac{\sigma}{M}$)	0.355	0.335	0.209

G.2 Factor Model of Interaction

As documented above, we have detailed measures on home visitor’s teaching ability, the interaction measures between home visitor and caregiver (child). To summarize them, we estimate the latent factors of home visitor’s teaching ability, and the interaction quality factors between home visitor and caregiver (child). We use a separate notation in this section.

Denote $M_{ia}^{j,l}$ as the measure j at household i at the child’s age a and the γ_{ia}^l is the latent factors l represents different factors (i.e., teaching ability, the interaction quality between home visitor and caregiver, and the interaction quality between home visitor and child).

$$M_{ia}^{j,l} = X'_{ia} \beta + \alpha^j \gamma_{ia}^l + \epsilon_{ia}^{j,l} \tag{1}$$

We have estimated the factor model by MLE assuming normal errors. We estimate the latent factor l based on the empirical Bayes method: the empirical conditional posterior distribution of the latent factor is given as

$$g(\gamma^l | M^l, X; \beta, \alpha) = \frac{\mu(M^l | X, \gamma^l; \beta, \alpha, \phi(\gamma^l))}{\int \mu(M^l | X, \gamma^l; \beta, \alpha, \phi(\gamma^l)) d\gamma^l} \tag{2}$$

Table G.4: Treatment Effects on Standardized Social-Emotional Scores

Variable	(1)	(2)	(3)	(4)	(5)
Denver Tasks	All	All	Children \leq 2 Yrs at Enrollment	All	Children \leq 2 Yrs at Enrollment
	Midline				
Treatment	0.509*	0.637**	0.339	0.617**	0.293
	[0.093,0.935]	[0.166,1.093]	[-0.080,0.750]	[0.203,1.031]	[-0.116,0.700]
Treatment \times Interaction Between Home Visitor and Caregiver	0.968**	0.971*	1.192**	0.906*	1.181**
	[0.278,1.722]	[0.153,1.815]	[0.406,1.962]	[0.093,1.709]	[0.386,1.949]
Treatment \times Interaction Between Home Visitor and Child	0.664**	0.568**	0.554*	0.459*	0.442
	[0.179,1.176]	[0.135,1.050]	[0.086,1.055]	[0.009,0.944]	[-0.050,0.977]
Treatment \times Teaching Ability	-0.930	-0.785	-0.845	-0.538	-0.636
	[-2.581,0.650]	[-2.263,0.687]	[-2.261,0.749]	[-2.030,1.047]	[-2.132,0.990]
	Endline				
Treatment	-0.270	-0.327	-0.211	-0.305	-0.180
	[-0.675,0.148]	[-0.789,0.109]	[-0.622,0.195]	[-0.716,0.138]	[-0.567,0.226]
Treatment \times Interaction Between Home Visitor and Caregiver	-0.470	-0.391	-0.698	-0.522	-0.761
	[-1.355,0.574]	[-1.357,0.687]	[-1.505,0.134]	[-1.465,0.585]	[-1.493,0.032]
Treatment \times Interaction Between Home Visitor and Child	-0.181	-0.123	-0.168	-0.157	-0.166
	[-0.542,0.163]	[-0.450,0.226]	[-0.472,0.170]	[-0.451,0.162]	[-0.491,0.177]
Treatment \times Teaching Ability	1.778*	1.400	1.643*	1.588*	1.675*
	[0.437,3.095]	[-0.108,2.847]	[0.231,3.027]	[0.150,2.973]	[0.146,3.024]
Pre-Treatment Covariates	No	No	No	Yes	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. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Therefore, the latent factor estimates is given as $\hat{\gamma}^l = \int \gamma g(\gamma|M^l, X; \beta, \alpha)d\gamma^l$.

Table G.3: Treatment Effects on Standardized Language and Cognitive Scores

Variable	(1)	(2)	(3)	(4)	(5)
Denver Tasks	All	All	Children \leq 2 Yrs at Enrollment	All	Children \leq 2 Yrs at Enrollment
	Midline				
Treatment	0.384*	0.451*	0.468*	0.498*	0.497*
	[0.008,0.746]	[0.033,0.862]	[0.036,0.911]	[0.101,0.887]	[0.054,0.944]
Treatment \times Interaction Between Home Visitor and Caregiver	0.379	0.509	0.451	0.559	0.495
	[-0.220,1.009]	[-0.117,1.199]	[-0.151,1.098]	[-0.033,1.154]	[-0.079,1.097]
Treatment \times Interaction Between Home Visitor and Child	0.223	0.337	0.265	0.291	0.235
	[-0.200,0.634]	[-0.048,0.707]	[-0.119,0.644]	[-0.065,0.643]	[-0.127,0.577]
Treatment \times Teaching Ability	1.028	0.585	0.622	0.631	0.649
	[-0.532,2.483]	[-0.976,2.06]	[-0.981,2.133]	[-0.804,2.028]	[-0.839,2.063]
	Endline				
Treatment	1.000***	0.924***	0.996***	1.018***	1.062***
	[0.590,1.405]	[0.498,1.356]	[0.613,1.375]	[0.580,1.458]	[0.666,1.451]
Treatment \times Interaction Between Home Visitor and Caregiver	0.959	1.21*	1.051*	1.343**	1.231**
	[-0.056,2.051]	[0.254,2.282]	[0.252,1.857]	[0.357,2.416]	[0.386,2.098]
Treatment \times Interaction Between Home Visitor and Child	0.255	0.223	0.203	0.178	0.176
	[-0.283,0.696]	[-0.273,0.622]	[-0.332,0.632]	[-0.357,0.626]	[-0.385,0.654]
Treatment \times Teaching Ability	-0.512	-0.803	-0.871	-0.764	-0.916
	[-2.725,1.63]	[-2.891,1.274]	[-3.146,1.297]	[-2.885,1.401]	[-3.176,1.482]
Pre-Treatment Covariates	No	No	No	Yes	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. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table G.5: Treatment Effects on Standardized Fine Motor Scores

Variable	(1)	(2)	(3)	(4)	(5)
Denver Tasks	All	All	Children \leq 2 Yrs at Enrollment	All	Children \leq 2 Yrs at Enrollment
	Midline				
Treatment	0.251	0.334	0.397	0.349	0.416
	[-0.246,0.762]	[-0.203,0.887]	[-0.116,0.903]	[-0.192,0.908]	[-0.081,0.932]
Treatment \times Interaction Between Home Visitor and Caregiver	0.392	0.574	0.593	0.594	0.618
	[-0.357,1.209]	[-0.253,1.548]	[-0.200,1.465]	[-0.333,1.553]	[-0.252,1.506]
Treatment \times Interaction Between Home Visitor and Child	-0.054	-0.188	-0.234	-0.205	-0.233
	[-0.523,0.383]	[-0.781,0.385]	[-0.916,0.35]	[-0.760,0.344]	[-0.838,0.323]
Treatment \times Teaching Ability	0.547	0.626	0.608	0.612	0.522
	[-1.295,2.479]	[-1.142,2.358]	[-1.195,2.456]	[-1.334,2.374]	[-1.363,2.408]
	Endline				
Treatment	0.661*	0.665*	0.665*	.733**	.700*
	[0.122,1.176]	[0.113,1.219]	[0.118,1.225]	[0.190,1.276]	[0.144,1.245]
Treatment \times Interaction Between Home Visitor and Caregiver	0.348	0.435	0.314	0.764	0.695
	[-0.533,1.277]	[-0.460,1.419]	[-0.495,1.256]	[-0.111,1.727]	[-0.140,1.615]
Treatment \times Interaction Between Home Visitor and Child	-0.247	-0.205	-0.236	-0.256	-0.291
	[-0.758,0.195]	[-0.654,0.235]	[-0.719,0.251]	[-0.791,0.212]	[-0.851,0.207]
Treatment \times Teaching Ability	-0.604	-0.826	-0.797	-0.961	-0.971
	[-2.398,1.066]	[-2.581,0.806]	[-2.648,0.91]	[-2.699,0.696]	[-2.784,0.799]
Pre-Treatment Covariates	No	No	No	Yes	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. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table G.6: Treatment Effects on Standardized Gross Motor Scores

Variable	(1)	(2)	(3)	(4)	(5)
Denver Tasks	All	All	Children \leq 2 Yrs at Enrollment	All	Children \leq 2 Yrs at Enrollment
	Midline				
Treatment	-0.018	-0.041	-0.004	-0.083	-0.075
	[-0.581,0.552]	[-0.612,0.524]	[-0.636,0.649]	[-0.571,0.438]	[-0.626,0.46]
Treatment \times Interaction Between Home Visitor and Caregiver	0.469	0.629	0.528	0.518	0.454
	[-0.439,1.277]	[-0.327,1.615]	[-0.553,1.539]	[-0.395,1.432]	[-0.551,1.421]
Treatment \times Interaction Between Home Visitor and Child	0.214	0.145	0.114	0.194	0.164
	[-0.205,0.666]	[-0.343,0.627]	[-0.350,0.600]	[-0.321,0.705]	[-0.367,0.701]
Treatment \times Teaching Ability	-0.518	-0.520	-0.422	-0.559	-0.508
	[-2.604,1.470]	[-2.606,1.489]	[-2.363,1.492]	[-2.228,1.254]	[-2.359,1.262]
	Endline				
Treatment	0.030	0.110	0.128	0.141	0.160
	[-0.582,0.638]	[-0.475,0.708]	[-0.456,0.734]	[-0.424,0.716]	[-0.374,0.722]
Treatment \times Interaction Between Home Visitor and Caregiver	0.105	0.103	0.141	0.291	0.296
	[-1.357,1.597]	[-1.205,1.463]	[-1.066,1.337]	[-1.092,1.657]	[-0.980,1.504]
Treatment \times Interaction Between Home Visitor and Child	0.408	0.279	0.267	0.266	0.249
	[-0.447,1.316]	[-0.574,1.213]	[-0.613,1.296]	[-0.661,1.317]	[-0.700,1.283]
Treatment \times Teaching Ability	0.593	0.694	0.716	0.582	0.651
	[-1.884,3.083]	[-1.801,3.239]	[-2.029,3.32]	[-2.048,3.288]	[-2.071,3.204]
Pre-Treatment Covariates	No	No	No	Yes	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. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

H Robustness Check for Factor Normalization

Factor models require normalizations as long as we seek to separate factors from factor loadings. We use the items in the category “Baseline” to normalize the first four loadings for the four skills studied in this paper. We use a self-explanatory simplified notation. The items listed assign $\alpha^{jk} = 1$ to the items listed and zero otherwise, while the remaining factor loadings are freely specified. In Table H.2, we show means of the latent factor loadings under different normalizations and also test whether they are different from our original normalization estimates. We find that the results are quite stable across different normalization choices if we choose the normalized items in the medium difficulty level range.

Table H.1: The List of Normalized Task Items

	Original	(1)	(2)	(3)	(4)	(5)
Social-Emotional	Wash and Dry Hands	Wash and Dry Hands	Wash and Dry Hands	Wash and Dry Hands	Drink From Cup	Drink From Cup
Fine Motor	Imitate Vertical Line	Thumb Wiggle	Imitate Vertical Line	Imitate Vertical Line	Imitate Vertical Line	Thumb Wiggle
Language and Cognitive	Combine Words	Combine Words	Name Body Parts 6	Name Body Parts 6	Name Body Parts 6	Name Body Parts 6
Gross Motor	Broad Jump	Broad Jump	Broad Jump	Balance Each Foot 1 Second	Balance Each Foot 1 Second	Balance Each Foot 1 Second

Table H.2: Skill Loading Mean Comparison under Different Normalizations

	Control						Treatment						
	Original	(1)	Different Normalizations				Original	(1)	Different Normalizations				
			(2)	(3)	(4)	(5)			(2)	(3)	(4)	(5)	
	Language and Cognitive Skill Loadings						Language and Cognitive Skill Loadings						
Mean	0.442	0.453	0.442	0.462	0.439	0.454	Mean	0.656	0.697	0.697	0.725	0.721	0.709
S.D.	(0.371)	(0.320)	(0.346)	(0.354)	(0.380)	(0.381)	S.D.	(0.478)	(0.447)	(0.457)	(0.447)	(0.479)	(0.462)
<i>p</i> -value		0.335	0.969	0.125	0.876	0.599	<i>p</i> -value		0.008	0.001	0.000	0.037	0.076
	Gross Motor Skill Loadings						Gross Motor Skill Loadings						
Mean	0.712	0.710	0.716	0.741	0.737	0.768	Mean	0.668	0.639	0.690	0.690	0.681	0.692
S.D.	(0.422)	(0.425)	(0.388)	(0.373)	(0.419)	(0.451)	S.D.	(0.453)	(0.405)	(0.433)	(0.397)	(0.418)	(0.416)
<i>p</i> -value		0.897	0.690	0.088	0.345	0.069	<i>p</i> -value		0.092	0.013	0.120	0.529	0.299
	Fine Motor Skill Loadings						Fine Motor Skill Loadings						
Mean	0.437	0.431	0.465	0.429	0.451	0.470	Mean	0.539	0.543	0.565	0.556	0.539	0.541
S.D.	(0.269)	(0.248)	(0.263)	(0.249)	(0.300)	(0.292)	S.D.	(0.240)	(0.228)	(0.237)	(0.219)	(0.240)	(0.252)
<i>p</i> -value		0.348	0.005	0.606	0.438	0.082	<i>p</i> -value		0.672	0.097	0.390	0.989	0.944
	Social-Emotional Skill Loadings						Social-Emotional Skill Loadings						
Mean	0.259	0.245	0.238	0.227	0.220	0.214	Mean	0.223	0.177	0.190	0.188	0.168	0.171
S.D.	(0.275)	(0.279)	(0.270)	(0.278)	(0.314)	(0.338)	S.D.	(0.260)	(0.227)	(0.227)	(0.202)	(0.276)	(0.290)
<i>p</i> -value		0.119	0.005	0.059	0.138	0.118	<i>p</i> -value		0.000	0.001	0.020	0.037	0.046

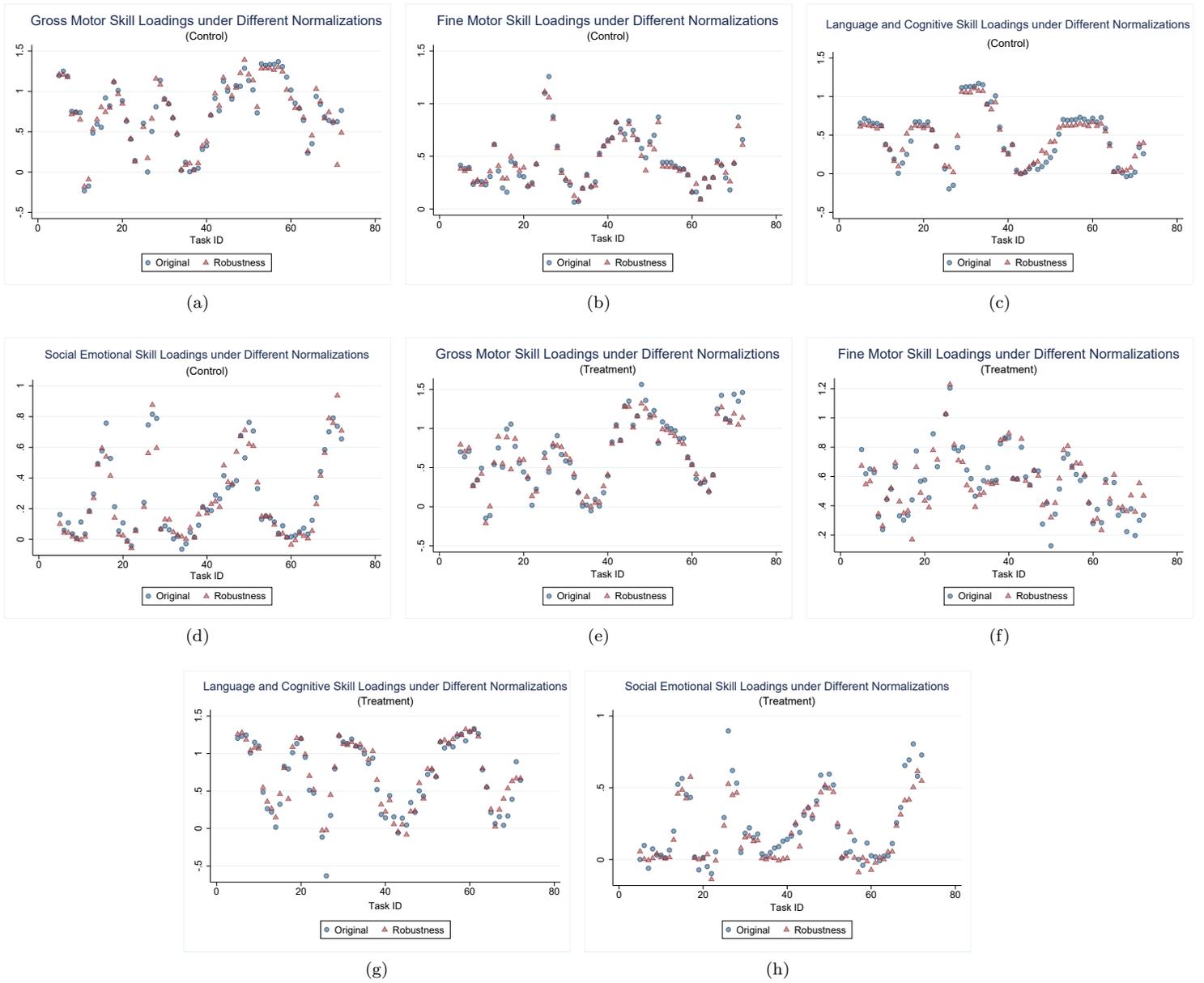


Figure H.1: The Comparison of Latent Skill Loadings under Different Normalizations

Table H.3: The Test of Equality in the Loadings between Treatment and Control ($\alpha^{j,1} = \alpha^{j,0}$)

	Social-Emotional Skills	Fine Motor Skills	Language and Cognitive Skills	Gross Motor Skills
Step Down p -values				
Social-Emotional Tasks				
Play Ball with Examiner	0.147	0.042	0.028	0.032
Help in House	0.193	0.049	0.024	0.131
Drink From Cup	0.054	0.038	0.021	0.059
Feed Doll	0.497	0.021	0.031	0.019
Use Spoon/fork	0.199	0.195	0.023	0.390
Remove Garment	0.121	0.239	0.026	0.036
Put on Clothing	0.152	0.018	0.066	0.039
Brush Teeth with Help	0.027	0.087	0.153	0.045
Name Friend	0.047	0.230	0.361	0.312
Put on T-shirt	0.519	0.437	0.390	0.031
Dress No Help	0.365	0.290	0.201	0.362
Play Board/Card Games	0.238	0.119	0.182	0.383
Brush Teeth no Help	0.476	0.436	0.511	0.433
Language and Cognitive Tasks				
DaDa/MaMa Specific	0.336	0.026	0.098	0.030
One Word	0.162	0.030	0.270	0.176
Two Words	0.051	0.025	0.447	0.167
3 Words	0.051	0.015	0.130	0.041
6 Words	0.033	0.013	0.109	0.174
Point 2 Pictures	0.108	0.012	0.124	0.284
Body Parts 6	0.076	0.034	0.081	0.045
Name 1 Picture	0.431	0.035	0.109	0.073
Speech Half Understandable	0.061	0.018	0.181	0.061
Point 4 Pictures	0.454	0.112	0.510	0.091
Know 2 Actions	0.048	0.067	0.046	0.035
Name 4 Pictures	0.072	0.161	0.057	0.049
Name 1 Color	0.167	0.150	0.450	0.037
Use of 2 Objects	0.102	0.107	0.164	0.043
Speech all Understandable	0.059	0.387	0.416	0.060
Know 2 Adjectives	0.045	0.393	0.332	0.059
Use of 3 Objects	0.206	0.101	0.602	0.018
Count 1 Block	0.072	0.458	0.469	0.041
Know 4 Actions	0.153	0.463	0.510	0.063
Understand 4 Prepositions	0.084	0.216	0.253	0.135
Know 3 Adjectives	0.198	0.187	0.091	0.109
Opposites 2	0.318	0.211	0.060	0.303
Count 5 Blocks	0.298	0.253	0.407	0.038
Name 4 Colors	0.078	0.134	0.051	0.200

Table H.4: The Test of Equality in the Loadings between Treatment and Control ($\alpha^{j,1} = \alpha^{j,0}$)

	Social-Emotional Skills	Fine Motor Skills	Language and Cognitive Skills	Gross Motor Skills
Step Down p -values				
Fine Motor Tasks				
Scribbles	0.162	0.033	0.034	0.037
Tower of 2 Cubes	0.058	0.060	0.031	0.429
Dump Raisin Demonstrated	0.131	0.032	0.020	0.090
Tower of 4 Cubes	0.428	0.050	0.032	0.073
Tower of 6 Cubes	0.146	0.031	0.248	0.010
Tower of 8 Cubes	0.274	0.016	0.219	0.035
Thumb Wiggle	0.173	0.613	0.668	0.039
Pick Longer Line	0.080	0.284	0.218	0.103
Copy Circle	0.039	0.371	0.481	0.043
Copy +	0.165	0.514	0.026	0.147
Draw Person 3 Parts	0.120	0.310	0.162	0.331
Gross Motor Tasks				
Get to Sitting	0.344	0.090	0.048	0.101
Pull to Stand	0.564	0.085	0.052	0.088
Stand 2 Seconds	0.607	0.087	0.040	0.060
Stand 10 Seconds	0.555	0.066	0.036	0.047
Stoop and Recover	0.180	0.052	0.025	0.236
Walk Well	0.093	0.044	0.023	0.051
Walk Backwards	0.075	0.087	0.024	0.421
Runs	0.370	0.048	0.019	0.445
Walk up Steps	0.383	0.036	0.018	0.052
Kick Ball Forward	0.128	0.028	0.020	0.254
Throw Ball Overhand	0.322	0.019	0.099	0.084
Jump up	0.239	0.016	0.095	0.087
Balance Each Foot 1 Second	0.203	0.013	0.239	0.396
Balance Each Foot 2 Second	0.241	0.342	0.249	0.375
Balance Each Foot 3 Second	0.045	0.158	0.031	0.044
Hops	0.152	0.521	0.073	0.161
Balance Each Foot 4 Second	0.674	0.133	0.030	0.017
Balance Each Foot 5 Second	0.395	0.509	0.159	0.014
Heel-to-toe Walk	0.688	0.358	0.076	0.347
Balance Each Foot 6 Seconds	0.477	0.101	0.059	0.351

I Consistency and Asymptotic Normality of Individual Factors (Wang, 2020) and a Factor Estimation Procedure from Chen, Fernández-Val, and Weidner (2021)

We use the analysis of Wang (2020), who proves the consistency and asymptotic normality of estimators of $\theta_i^{j(k)}$ under conditions we satisfy. We actually know some of the factors he estimates, so our model is a special case of his. We apply an estimation procedure for factors proposed by Chen, Fernández-Val, and Weidner (2021).⁶ We use a simplified self-explanatory notation in this appendix to facilitate exposition.

- (1) For each iteration k , given the set of parameters $\{\beta^{j(k)}, \theta_i^{j(k)}, \alpha^{j(k)}, \delta^{j(k)}\}$, define $\mu_{ij}^j = \mathbf{X}^{j'} \beta^{(k)} + \delta^{j(k)} + (\theta_i^{j(k)})' \alpha^{j(k)}$.
- (2) **E-step:** Calculate

$$\begin{aligned} \hat{Y}_i^{j(k)} &= E(Y_{ij}^{j*} | Y_i^j, \mathbf{X}_i^j, \beta^{j(k)}, \alpha^{j(k)}, \delta^{j(k)}, \theta_i^{j(k)}) \\ &= \mu_i^{j(k)} + (Y_i^j - \Phi^{j(k)}(\mu_i^{j(k)})) \phi^{j(k)}(\mu_i) / \{\Phi(\mu_i^{j(k)}) (1 - \Phi(\mu_i^{j(k)}))\}. \end{aligned}$$

- (3) **M-step** conditional maximization steps:

- Update β^j : $\beta^{j(k+1)} = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' (\hat{Y}^{j(k)} - \delta^{j(k)} - (\theta^{j(k)})' \alpha^{j(k)})$.
- Update δ^j : $\delta^{j(k+1)} = \sum_i (\hat{Y}_i^{j(k)} - \mathbf{X}_i^{j'} \beta^{j(k+1)} - (\theta_i^{j(k)})' \alpha^{j(k)}) / N_{I_j}$.
- Update $\alpha_{j,m}^{j(k+1)}$, where m indicates the m th latent factor, since for each latent factor, we have one item m^* with loading $\alpha_{m^*, m^*} = 1$ and $\alpha_{m^*, j \neq m^*} = 0$

$$\alpha_{j,m}^{(k+1)} = \frac{\sum_i (\hat{Y}_i^{j(k)} - \mathbf{X}_i^{j'} \beta^{j(k+1)} - \delta^{j(k+1)}) (\hat{Y}_i^{m^*(k)} - (\mathbf{X}_i^{m^*})' \beta^{j(k+1)} - \delta^{m^*(k+1)})}{\sum_i (\hat{Y}_i^{m^*(k)} - (\mathbf{X}_i^{m^*})' \beta^{j(k+1)} - \delta^{m^*(k+1)})^2}.$$

- Update $\theta_i^{j(k+1)}$, in this step, we use the closed form solution to update the individual level latent factors, which is more robust than the method proposed in Chen, Fernández-Val, and Weidner (2021).

$$\theta_i^{j(k+1)} = (\hat{Y}_i^{j(k)} - \mathbf{X}_i^{j'} \beta^{j(k+1)} - \delta^{j(k+1)})' \alpha^{j(k+1)} (\alpha^{j(k+1)} \alpha^{j(k+1)})^{-1}.$$

- (4) Iterate until convergence.

J Fit of Estimated Models to Sample Data

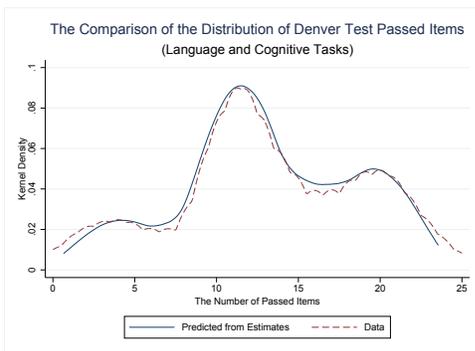


Figure J.1: Model Fit for Language and Cognitive Tasks

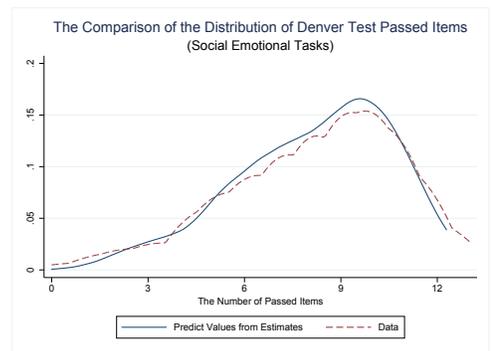


Figure J.2: Model Fit for Social-Emotional Tasks

⁶Wang's analysis assumes no \mathbf{X}_i^j and identifies and develops a consistent estimator of θ_i^j , as well as of factor loadings α^j . It is trivial to apply his analysis when components of θ_i^j are known, i.e., the \mathbf{X}_i^j .

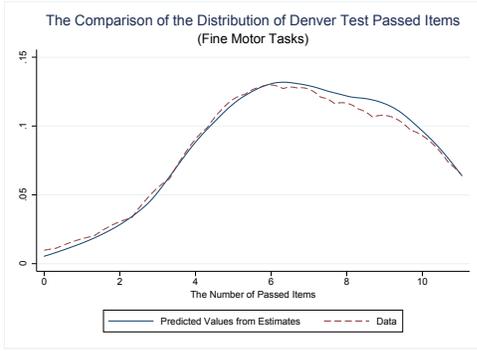


Figure J.3: Model Fit for Fine Motor Tasks

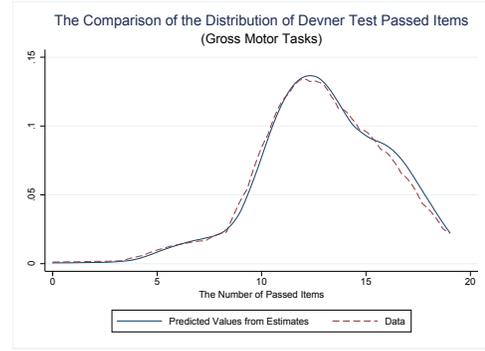


Figure J.4: Model Fit for Gross Motor Tasks

K Stochastic Dominance

We test for stochastic dominance of the estimated skill curves. Figure K.1a, the generalized Lorenz curve, shows the average cumulative values at each cumulative proportion observation. At each cumulative proportion, the treated children have higher language skills. Similarly, Figure K.1b gives the maximum language and cognitive skill values at each percentile. It is clear that the treated group has larger language skill values at each percentile. Figures K.2–K.4 show the same measures for social-emotional, fine motor and gross motor skills, respectively. We can find similar patterns for social-emotional and fine motor skills but not for gross motor skill.

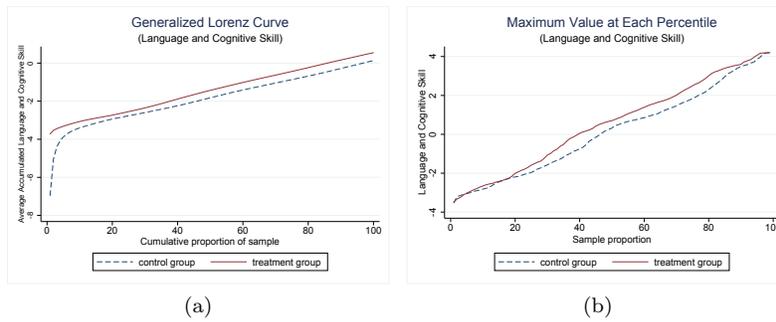


Figure K.1: Language and Cognitive Skills Stochastic Dominance Curves

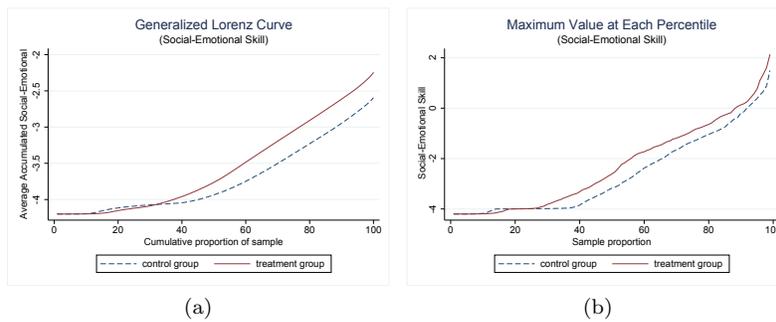


Figure K.2: Social-Emotional Skills Stochastic Dominance Curves

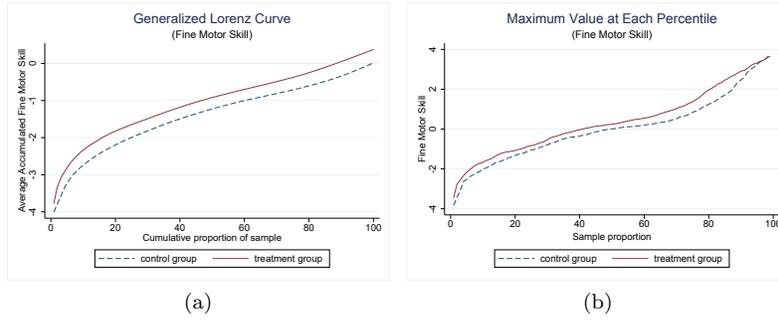


Figure K.3: Fine Motor Skills Stochastic Dominance Curves

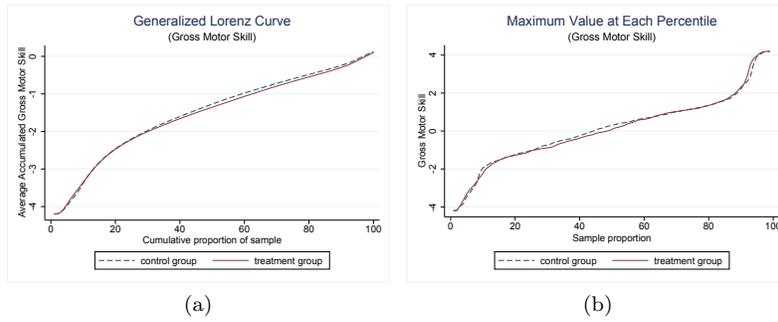
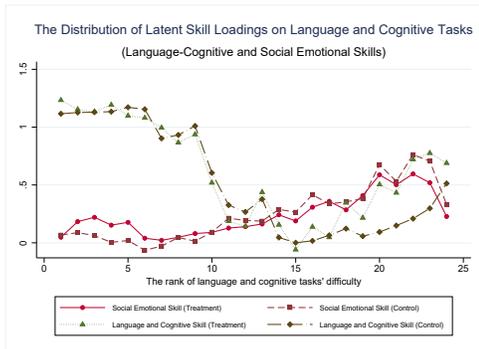
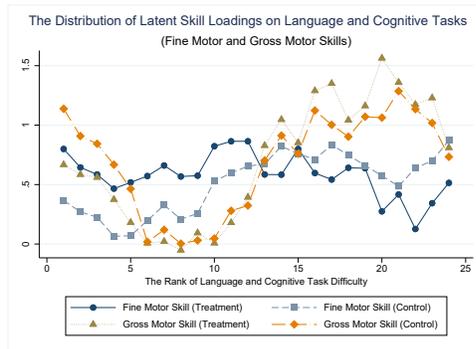


Figure K.4: Gross Motor Skills Stochastic Dominance Curves

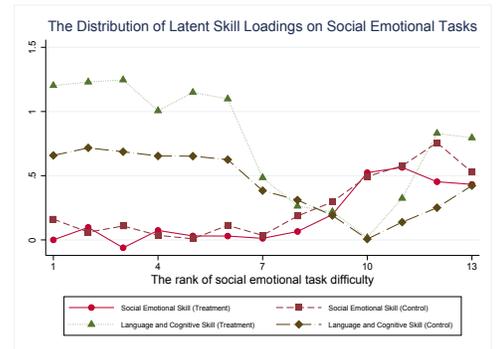
We plot the point estimates of loadings for each task in Figure K.5. We find that the scale of language and cognitive skill loadings on language and cognitive tasks is much higher than the loadings of social-emotional skill on language and cognitive tasks. This finding also holds with other types of tasks. This means that improving language and cognitive skill could boost the child's overall Denver test performance given the same level social-emotional and language skills.



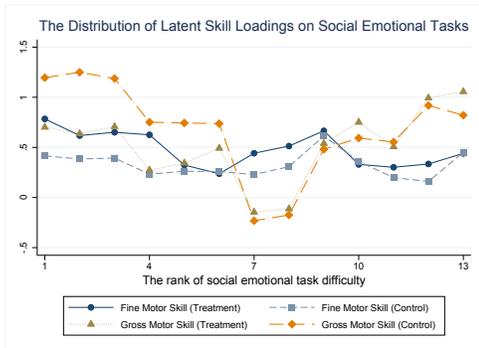
(a)



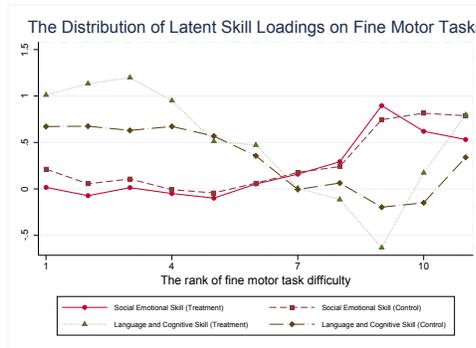
(b)



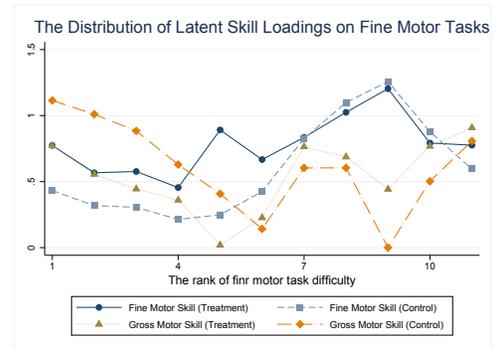
(c)



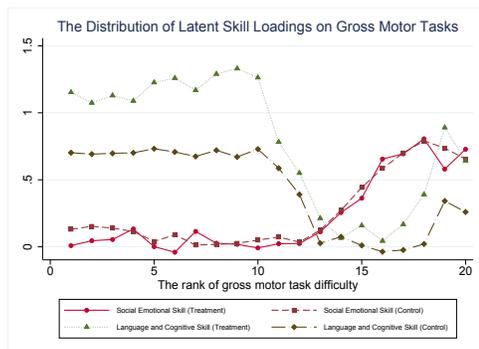
(d)



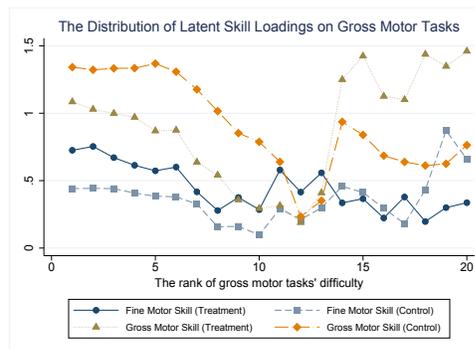
(e)



(f)



(g)



(h)

Figure K.5: The Distribution of Latent Skill Loadings

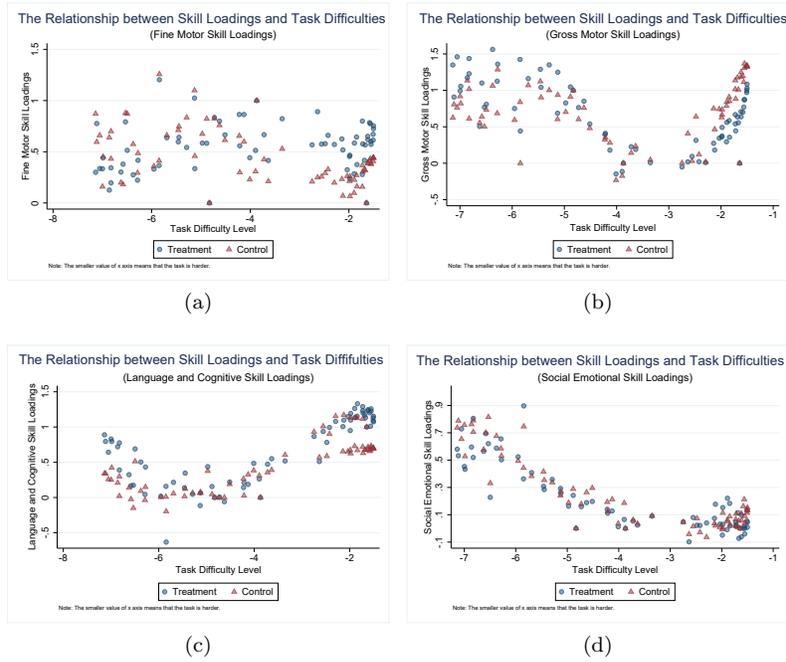
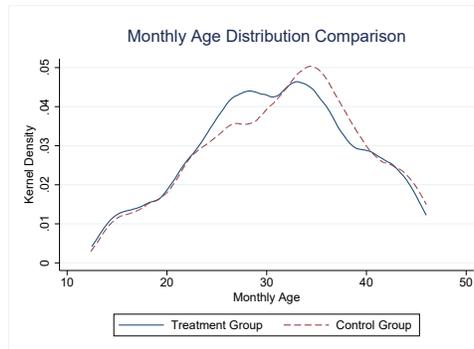


Figure K.6: The Relationship between Latent Skill Loadings and Task Difficulties

L The Monthly Age Distribution Comparison

Figure L.1: The Monthly Age Distribution Comparison



Note: The p -value of the Kruskal-Wallis test is 0.18.

References

- 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.
- CHEN, M., I. FERNÁNDEZ-VAL, AND M. WEIDNER (2021): “Nonlinear Factor Models for Network and Panel Data,” *Journal of Econometrics*, 220(2), 296–324.
- MAASOUMI, E., AND L. WANG (2019): “The Gender Gap between Earnings Distributions,” *Journal of Political Economy*, 127(5), 2438–2504.
- TSIATIS, A. (2006): *Semiparametric Theory and Missing Data*. New York: Springer.
- WANG, F. (2020): “Maximum likelihood estimation and inference for high dimensional generalized factor models with application to factor-augmented regressions,” *Journal of Econometrics*.