Interactions as Investments: The Microdynamics and Measurement of Early Childhood Learning*

James Heckman and Jin Zhou

Center for the Economics of Human Development
and Department of Economics
University of Chicago
January 3, 2023

*This research was supported in part by NIH grant NICHD R37HD065072, the Institute for New Economic Thinking and an anonymous donor. We thank our partners on the China REACH project: China Development Research Foundation, Beijing Normal University, Child Development Research Group at the University of West Indies, and Dr. Sally Grantham-McGregor. The views expressed in this paper are solely those of the authors and do not necessarily represent those of the funders, partners, or the official views of the National Institutes of Health. This paper was presented as a keynote at the Africa Meeting of the Econometric Society (June 4, 2021), the China Meeting of the Econometric Society (July 1, 2021), the Bonn IZA-Briq Life-Cycle Workshop (August 19, 2021), and at Rice University (November, 2021). We have benefited from the comments of the editor, John List, and four insightful referees, as well as those of our discussant Lance Lochner at a September 2019 Becker Friedman Institute conference in Chicago, as well as those of Flavio Cunha, and Pat Kyllonen. Fuyao Wang, Haihan Tian, and Alejandra Campos contributed highly competent and insightful research assistance and commentary.

Abstract

This paper uses novel experimental data from an early childhood program with high-frequency measurements to investigate the dynamics of skill formation. We show that home-visiting interventions promote child development through quality interactions between home visitors and caregivers. We report non-parametric evidence consistent with dynamic complementarity. We formulate and estimate a dynamic learning model and quantify the sources of early life learning. Using our model, we test the widely held assumption of the existence of constant units of skills that are comparable across levels of skills and ages. We find evidence supporting it for certain skill levels but not for all.

JEL Codes: I3, J1, C5, D2, O12, C9

Keywords: child development, measures of skills, scaffolding, targeting, experiment, IRT, BKT

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

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

1 Introduction

This paper analyzes a low-cost home visiting program in China with unique high-frequency (weekly) data on skill development that is based on a widely-emulated program originally developed in Jamaica that has been shown to be effective in developing child skills (e.g., Grantham-McGregor and Smith, 2016; Gertler et al., 2014, 2022). The study of the effectiveness of home visiting programs isolates a component of successful omnibus programs that include this feature (see Zhou et al., 2022; García and Heckman, 2022).

We investigate the mechanisms producing growth of knowledge on multiple skills in the early years. Using nonparametric methods, we find evidence consistent with a crucial property of learning: dynamic complementarity–acquired skills foster the growth of later skills. We develop and estimate a micro-dynamic model of reinforcement learning to characterize the dynamics of skill formation during early childhood.

The *technology of skill formation* (Cunha and Heckman, 2007) characterizes the growth of child skills at age (stage) $\boldsymbol{a}$: $\boldsymbol{K(a)}$. It is a function of a vector of investments $\boldsymbol{I(a)}$ (including home visits, parenting, interactions with the child, school-based interventions, center care, school stimulation, etc.) and environments $\boldsymbol{G(a)}$ (including neighborhoods, parental education, and public goods):

$$\overbrace{\boldsymbol{K(a+1)}}^{\substack{\text{Skills at}\\ \boldsymbol{a+1}}} = \boldsymbol{f^{(a)}}\Big(\underbrace{\boldsymbol{K(a)}}_{\substack{\text{Skills at}\\ \boldsymbol{a}}}, \overbrace{\boldsymbol{I(a+1)}}^{\text{Investment}}, \underbrace{\boldsymbol{G(a+1)}}_{\substack{\text{Environmental}\\ \text{Variables}}}\Big). \tag{1}$$

For simplicity, we assume that this age- or stage-dependent function is twice continuously differentiable. Key properties of $f^{(a)}$ featured in the literature are self productivity ($\frac{\partial \boldsymbol{K(a+1)}}{\partial \boldsymbol{K(a)}} > 0$), the productivity of investment and beneficial environments ($\frac{\partial \boldsymbol{K(a+1)}}{\partial \boldsymbol{I(a+1)}} > 0$, $\frac{\partial \boldsymbol{K(a+1)}}{\partial \boldsymbol{G(a+1)}} > 0$), and critical and sensitive periods of development ($\boldsymbol{f_2^{(a)}} > \boldsymbol{f_2^{(a')}}$, $a \neq a'$, evaluated at common values, where $\boldsymbol{f_2^{(a)}}$ is $\frac{\partial \boldsymbol{K(a+1)}}{\partial \boldsymbol{I(a+1)}}$). Static complementarity ($\frac{\partial^2 \boldsymbol{K(a+1)}}{\partial \boldsymbol{K(a)} \partial \boldsymbol{I(a+1)}} > 0$) is often found in empirical studies of child development. Investment is more productive the higher the stock of skills; i.e., "skill begets skill." Dynamic complementarity ($\frac{\partial^2 \boldsymbol{K(a+j+1)}}{\partial \boldsymbol{I(a)} \partial \boldsymbol{I}'\boldsymbol{(a+j)}} \geq 0$, for $j > 1$) is a central proposition in the literature. It asserts that investment at earlier life cycle stages makes later investments more productive. It implies that remediation of skill deficits at later stages of the life cycle is more costly (requires more investment) than direct investment at early ages (Heckman and Mosso, 2014).

There are three big questions in this literature. (1) What is $\boldsymbol{I(a)}$ and how to measure it?[1] (2) What are the micro-mechanisms underlying the technology? Child psychologists emphasize that warm and supportive parent/caregiver-child interactions— *"scaffolding"* (Vygotsky, 1978)—are major determinants of child development. (3) How should we measure skills and their growth?

This paper focuses on the mechanisms underlying technology (1) using high frequency (weekly) data on the growth of skills in the treatment group of the China REACH home visiting program. The paper by Zhou, Heckman, Liu, and Lu (2022), examines treatment effects of the program studied at endline and midline and presents

[1]Many different definitions are used. For example, books in the home, time spent in childcare/ playing with the child, parenting styles (e.g., Doepke and Zilibotti, 2019; Kim, 2019; etc.), external interventions at centers or home visits. See, e.g., Cunha and Heckman (2008); Cunha et al. (2010); Del Boca et al. (2014); Agostinelli and Wiswall (2022); Andrew et al. (2020); Doepke and Zilibotti (2019).

and applies methods for correcting for item difficulty. Zhou, Heckman, Liu, Lu, Chang, and Grantham-McGregor (2022) consider the issue of external validity and compliance with the SANS conditions of List (2020). A third paper focuses on issues of measurement (Heckman and Zhou, 2022a).

In the literature, test scores based on passing rates on assessments of cognitive, socioemotional, and other skills are widely used.[2] Such measures have arbitrary scales (e.g., Uzgiris and Hunt, 1975; Cunha and Heckman, 2008; Cunha et al., 2010). Ordinal production functions that compare ranks across people do not suffer from this problem, but at the same time, do not measure levels of attained skill. Value-added measures of school, teacher and student quality assume that constant unit measures are available to make meaningful comparisons.[3]

As a byproduct of our dynamic model, we propose and implement model-based tests of *invariance* of latent skills, a crucial assumption maintained in the value-added and human capital literatures and specifically in previous research on skill formation. It maintains the assumption of the existence of constant-unit latent skills (human capital) over all levels of the same skill. This literature also assumes the existence of constant-unit *measures* of latent skills, which may or may not exist even if constant unit latent skill scales exist.[4] One approach to the problem of defining and

[2]See, e.g., Kautz et al. 2014; OECD 2021.

[3]Cunha et al. (2010, 2021); García and Heckman (2022); Agostinelli and Wiswall (2022); Freyberger (2021).

[4]For example, Todd and Wolpin (2007) and others use words spoken by age as a measurement of the invariant latent skill. The obvious question is whether twice as many words at age 5 is the same amount of knowledge as twice the same at age 8. Are percent changes comparable at different ages? What is the appropriate metric? Are there common scales of knowledge? Is there a single scale to measure the growth of knowledge over time? For all skills? For any particular skill? An assumption of common scales of measurement ignores the finding that multiple skills emerge as a child matures. In addition, many assessments bundle multiple skills (e.g., grades depend on

measuring scale anchors test scores in meaningful outcomes (e.g., earnings, crime). However, objective behavioral anchors at early ages are difficult to find.[5] The recent literature demonstrates the empirical importance of these issues. Freyberger (2021) shows the dramatic consequences of different scalings of skill measures for estimates of the technology of Equation (1).

The current paper addresses the first two questions and the first aspect of the third problem (Existence of invariant latent scales). Heckman and Zhou (2022a) address the second aspect: the issue of existence of invariant *measures* of skills.

In this paper, we first document empirical evidence on the learning process. We examine key mechanisms of home visiting interventions that improve child skill development. We evaluate the impacts on child development of the interactions between home visitors and caregivers and the impact of home visitors' teaching quality. We present evidence consistent with dynamic complementarity without imposing constant-unit invariance assumptions.

We develop and estimate a new stochastic micro-dynamic model of skill formation that formalizes mechanisms proposed in developmental psychology and explains uneven growth of test scores over levels and fadeout of measured skills over age for the same person.[6] We investigate the growth of skills at more granular levels than previous analyses.

We report the following findings. (1) A key mechanism fostering growth of child skills is quality interactions between home visitors and caregivers. (2) We present

cognitive and noncognitive skills) (Borghans et al., 2016). A growing body of evidence challenges the validity of psychometric conventions (see, e.g., Almlund et al., 2011 and Kautz et al., 2014).

[5]For a recent discussion of these problems, see Cawley et al. (1998) and Cunha et al. (2021).

[6]See, e.g., Bronfenbrenner (2005) and Thelen (2005).

evidence consistent with dynamic complementarity using nonparametric methods. (3) Based on this evidence, we develop and estimate a dynamic reinforcement learning model. It unites and extends two highly-influential models of psychometrics: the IRT (Item Response Theory) model and the BKT (Bayesian Knowledge Tracing) model.[7] We add investment and stochastic growth to these frameworks. We find evidence supporting the assumption of invariance of latent skills across levels for certain skills at specific skill levels but not globally. This complements our analysis of invariant *measures* in Heckman and Zhou (2022a).

The paper unfolds in the following way. Section 2 describes the background of the program we analyze and its curriculum.[8] Section 3 presents our evidence on learning patterns. Section 4 discusses the impacts of different interactions on learning. Section 5 presents nonparametric evidence consistent with dynamic complementarity. Section 6 develops a latent Markov process micro-dynamic learning model. Section 7 presents estimates and interpretations. Section 8 concludes.

2 China REACH

The inspiration for the program analyzed is the Jamaican Home Visiting Intervention (Grantham-McGregor and Smith, 2016). It was a randomized home visiting parenting intervention given to a sample of 129 stunted children between 9 and 24 months of age. Substantial positive effects are found for the program through age 34 (i.e., Gertler et al., 2022, 2014). Its success has spawned replications around the world,

[7]See van der Linden (2016).

[8]Zhou et al. (2022) describe it in much greater detail.

e.g., in Bangladesh, China, Colombia, India, Peru (see, e.g., Grantham-McGregor and Smith, 2016).

The program we analyze, *China REACH*, extends and applies the Jamaican protocols. Implemented in 2015 by a large-scale random control trial, it enrolled 1,500 subjects (aged 6 months-42 months) in 111 villages in Huachi county, Gansu province, one of the poorest areas of China. This intervention is not focused on stunted children.

China REACH is a paired-match RCT that minimizes mean square errors of estimates (Bai et al., 2021; Bai, 2022). A non-bipartite Mahalanobis matching method[9] was used to pair villages and randomly select one village within the pair into the treatment group and the other village into the control group. More details of the design of the experiment and balance tests for treatment and control groups can be found in Zhou, Heckman, Liu, and Lu (2022).

The intervention focuses on improving multi-dimensional skill development through a home-visiting model. Trained home visitors who are roughly at the level of education of the mothers of the children studied visit each treated household weekly and provide one hour of caregiving guidance.

Zhou, Heckman, Liu, and Lu (2022) evaluate the treatment effects of the intervention and find that the intervention significantly improves skill development (e.g., language and cognitive, fine motor, and social-emotional skills). To interpret treatment effects, they use item responses on inventories of skill to estimate individual latent skills. They decompose the source of treatment effects and find that enhance-

[9]See Lu et al. (2011).

ment in latent skills explains most of the conventional treatment effects. Zhou et al. (2022) show that the skill profiles for the growth of skills are similar to those of the original Jamaica Home Visiting program, suggesting some generality of our analysis.

2.1 Program Protocols

The program teaches and encourages the mother/grandparent(s) to talk with the child through playing games, making toys, singing, reading, and storytelling to stimulate the child's cognitive, language, motor, and socioemotional skill development.

About three to four different skill tasks (gross motor, fine motor, language, and cognitive) are taught each week. Skills taught are ordered by difficulty levels following profiles developed by Palmer (1971) and Uzgiris and Hunt (1975) and widely applied in the literature on child development. Central to our identification strategy is the assumption that these profiles describe valid hierarchies (levels) of knowledge and that the knowledge content is the same within each level.[10] Child skills are assessed weekly. There are monthly assessments of the quality of home visits recorded by supervisors.

There are 13 difficulty levels for cognitive skills. Table 1 gives the tasks for cognitive skills taught at specific levels and Figure 1 presents the timing of the lessons by age. The tasks start with simply understanding a picture by verbal acknowledgment to using receptive (heard) language to identify pictures. Although the task content progresses by levels, the task content is similar within the same difficulty level. For example, the contents of cognitive skill tasks at level 1 are described in Table 2. All

[10]The difficulty levels are ordered based on the average children's performance (see Palmer, 1971.)

Table 1: Difficulty Level List for Cognitive Skill Tasks

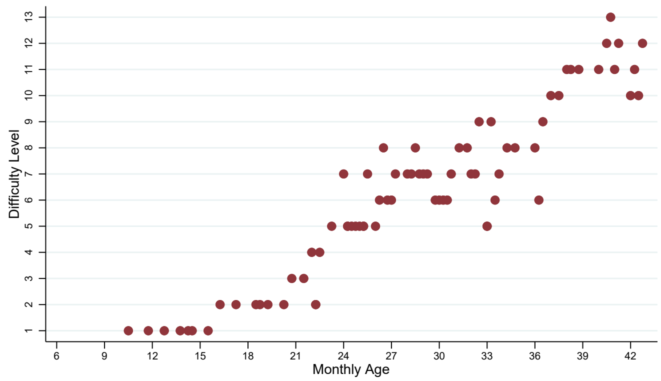
Level 1	Look at the pictures and vocalize
Level 2	Name the objects and ask the baby to point to the pictures accordingly
Level 3	The child can name the objects in one picture, and point to the named picture
Level 4	The child can name the objects in two or more pictures, and point to the named picture
Level 5	The child can point out named pictures, and say names of three or more
Level 6	The child can point out the picture mentioned and correctly name the name of six or more pictures
Level 7	The child can talk about the pictures, answer questions, understand, or name the verbs (eat, play, etc.)
Level 8	The child can follow the storyline, name actions, and answer questions
Level 9	The child can understand stories, talk about the content in the pictures
Level 10	The child can keep up with the development of the story
Level 11	The child can say the name of each graph, discuss the role of each item and then link the graphics in the card together
Level 12	The child can name the things in the picture and link the different pictures together and discuss some of the activities in the pictures
Level 13	The child can name the things in the picture and talk about the function of objects

tasks at that level are virtually identical in task difficulty and relate to the activity of looking at pictures or objects and vocalizing. Appendix A gives comparable information for the other skills.

3 Empirical Evidence on Learning

This section documents the observed learning patterns of the study.

Figure 1: The Timing of Cognitive Skill (Understand Objects) Tasks across Difficulty Levels



Note: Level 1: Look at the pictures and vocalize; Level 13: The child can name the things in the picture and talk about the function of objects.

3.1 High Frequency Data on Learning

Our data on weekly skill growth enable us to move beyond traditional aggregate measures such as the percent of items passed over a diverse range of tasks to examine task by task skill growth and the factors that influence it. To understand the structure of the data analyzed, we introduce some helpful notation.

Let $\mathcal{S}$ be the set of skills taught. Let $\ell(s,a)$ be the level of skill s taught at age a. Mastery of skill s at level ℓ at age a is characterized by:

$$D(s,\ell,a) = \begin{cases} 1 & K(s,\ell,a) \geq \bar{K}(s,\ell) \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where $D(s,\ell,a)$ records mastery (or not) of a skill at a given level at age a. $\bar{K}(s,\ell)$ is the minimum latent skill required to master the task at difficulty level ℓ. This

Table 2: Cognitive Skill Task Content: Look at the Pictures and Vocalize (Level 1)

Difficulty Level	Difficulty Level Aim	Month	Week	Learning Materials	Task Aim and Content
Level 1	Look at the pictures and vocalize	10	2	Picture book A	Look at the pictures and vocalize: baby makes sound when looking at the pictures
Level 1	Look at the pictures and vocalize	11	3	Picture book B	Look at the pictures and vocalize: baby looks at the pictures and vocalize
Level 1	Look at the pictures and vocalize	12	3	Picture book A	Look at the pictures and vocalize: baby makes sound when looking at the pictures
Level 1	Look at the pictures and vocalize	13	3	Picture book B	Look at the pictures and vocalize: baby looks at the pictures and vocalize
Level 1	Look at the pictures and vocalize	14	1	Picture book A	Look at the pictures and vocalize: baby makes sound when looking at the pictures
Level 1	Look at the pictures and vocalize	14	2	Baby doll	Look at the pictures and vocalize: baby makes sound when holding a baby doll
Level 1	Look at the pictures and vocalize	15	2	Picture book B	Look at the pictures and vocalize: The child pronounces while looking at the pictures

characterization is consistent with the classical IRT model (Lord and Novick, 1968). Let $\underline{a}(s,\ell)$ be the first age at which skill s is taught at level ℓ, and let $\bar{a}(s,\ell)$ be the last age at which it is taught at level ℓ. For consecutive lessons in a run, $1+\bar{a}(\ell)-\underline{a}(\ell)$ is the length of run (# of lessons taught on skill s at level ℓ) starting at age $\underline{a}(s,\ell)$. For level ℓ of skill s, we collect the indicators of knowledge in a spell, $\left\{D(s,\ell,a)\right\}_{\underline{a}(s,\ell)}^{\bar{a}(s,\ell)}$.

3.2 Characterizing Learning

In a stationary environment with age-invariant individual heterogeneity and with no learning or growth of knowledge at level ℓ and skill s, the sequences $\{D(s,\ell,a')\}$, $a' \in [\underline{a}(\ell),\bar{a}(\ell)]$ are exchangeable (i.e., they are equally probable for any order within ℓ).[11] With learning, sequences are back-loaded, i.e., for $j>0$, $\Pr(D(s,\ell,a+j) \geq D(s,\ell,a)) \geq \Pr(D(s,\ell,a+j) \leq D(s,\ell,a))$.

Zhou, Heckman, Wang, and Liu (2022) test exchangeability on weekly data and reject that hypothesis, indicating learning. Learning is found even after controlling

[11]See Heckman (1978, 1981).

for maturation and exposure effects that might boost skills in the absence of any intervention (see Appendix B).

Figure 2 characterizes the growth of knowledge in language, cognitive, and fine motor skills.[12] Average passing rates within each difficulty level for language and cognitive tasks increase with age, a pattern consistent with learning. When individuals transition to higher difficulty levels, initial passing rates decline. Subsequent passing rates increase as learning ensues. The dynamic model presented in Section 6 captures this phenomenon. At most levels of fine motor skills, there is—at best—modest learning.[13] Access to detailed weekly data enables us to determine at what stages learning occurs.

3.3 Measures of Learning and Knowledge

Heckman and Zhou (2022a) compare the traditional measure of learning: the proportion of correct answers on a broad range of tasks with two alternative measures of learning and learning speed: time to first mastery and backsliding.

The passing rate on skill s at level ℓ is:

$$p(s, \ell) = \frac{1}{\bar{a}(s, \ell) - \underline{a}(s, \ell) + 1} \sum_{a=\underline{a}(s,\ell)}^{\bar{a}(s,\ell)} D(s, \ell, a). \tag{3}$$

The overall passing rate is:

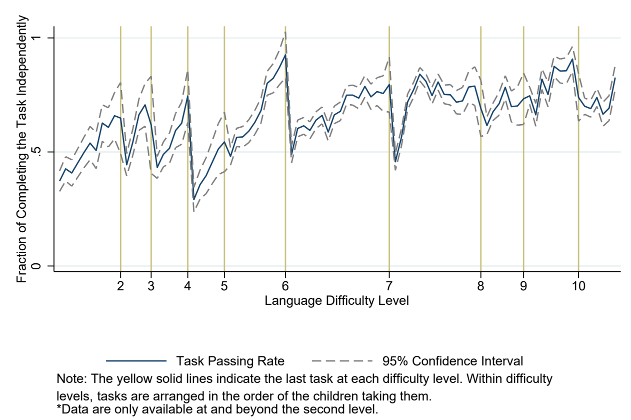
$$p(s) = \sum_{\ell=1}^{L_s} p(s, \ell) w(s, \ell) \tag{4}$$

[12]The program has no measured effect on gross motor skills.

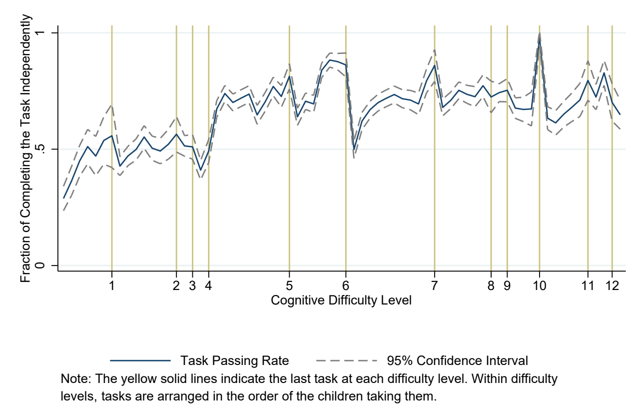
[13]We also measure gross motor skills, but they are very flat with age and are not affected by the intervention, so we do not systematically analyze them in the text.

Figure 2: Average Task Passing Rates by Order and Level

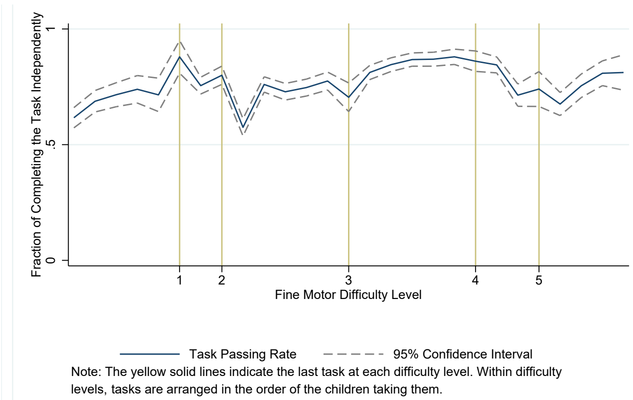
(a) Language



(b) Cognitive



(c) Fine Motor



where L_s is the highest level of skill s and

$$w(s,\ell) = \frac{\bar{a}(s,\ell) - \underline{a}(s,\ell) + 1}{\sum_{\ell=1}^{L_s}(\bar{a}(s,\ell) - \underline{a}(s,\ell)) + 1}. \tag{5}$$

This measure weights passing rates at different difficulty levels by the number of items on it tested.

There are other plausible measures of knowledge and learning. For consecutive learning spells with all participants entering each level at the first lesson, the ***Time to first mastery*** is $d(s,\ell) = \hat{a}(s,\ell) - \underline{a}(s,\ell)$, where for each s and ℓ, $\hat{a}(s,\ell) = \min_a\{D(s,\ell,a) = 1\}_{a=\underline{a}(s,\ell)}^{\bar{a}(s,\ell)}$. This is often used as a measure of intelligence (van der Linden, 2016). Another possible measure is ***Instability*** at level ℓ for skill s as: $\frac{\#\{D(s,\ell,a)=0, a>\hat{a}(s,\ell), a\leq\bar{a}(s,\ell)\}}{\#\{a>\hat{a}(s,\ell), a\leq\bar{a}(s,\ell)\}}\mathbf{1}(\#\{a > \hat{a}(s,\ell), a \leq \bar{a}(s,\ell)\} > 0)$. This captures retention of knowledge.

Heckman and Zhou (2022a) show that in the China REACH data, these measures are correlated in the expected directions. However, the different measures are far from being perfectly correlated, suggesting that they capture different aspects of knowledge.[14] They report that there are two dimensions for each skill and at least five dimensions across all skills. The notion of a single dimension of skill–assumed in standard efficiency unit models in economics and in the psychology of "g" that claims one universal skill predicts performance on all tasks–is grossly inaccurate.

[14]An alternative explanation is substantial measurement error. Our factor analyses of these data show that measurement error ("uniqueness") is a real possibility. See Cunha et al. (2021) for a discussion of measurement errors in measures of achievement.

4 Impacts of Interventions on the Growth of Skills

We now analyze how interventions improve a child's skill development by examining the effects that different interactions have on child learning across difficulty levels. During the intervention, supervisors record assessments of home visitor, caregiver, and child interaction activities at least once per month, making it possible to examine their impacts on skill development. Using these measures, we can evaluate the quality of interaction between home visitors and caregivers and between home visitors and children and their impacts. Trained program supervisors evaluate the quality of home visits in three dimensions: (a) Quality of the home visitor's teaching ability; (b) Interaction quality between the home visitor and the caregiver; and (c) Interaction quality between the home visitor and the child. Appendix C describes the interaction data and the factors that summarize it.

Table 3 reports the impact of program interactions on time to the first mastery of achieving cognitive tasks. It shows a recurrent pattern. The interaction between the home visitor and the caregiver is measured at each skill level. We form an average over all visits. It is the only consistently statistically significant pattern across all difficulty levels.[15] Note that age (maturation) effects are statistically important. Children acquire skills with age and experience. Having a grandmother as the caregiver retards learning speed.

[15] Note that a negative coefficient for a mastery regression means a quicker mastery of the skill.

Table 3: The Effects of Interactions on the Time to Mastery of Cognitive Tasks at Each Level (IV)

	Cognitive Task Difficulty Levels										
	≤2	3	4	5	6	7	8	9	10	11	12
Interaction Quality	-0.923*	-0.212**	0.007	-0.819***	-0.710***	-0.699*	-0.259*	-0.208***	-0.669***	-0.466**	-0.196
Home Visitor and Caregiver	(0.515)	(0.108)	(0.092)	(0.190)	(0.276)	(0.366)	(0.150)	(0.060)	(0.215)	(0.189)	(0.131)
Interaction Quality	-0.082	-0.003	-0.052**	-0.091	-0.042	0.050	-0.015	0.019*	0.004	0.015	-0.049
Home Visitor and Child	(0.130)	(0.028)	(0.021)	(0.077)	(0.068)	(0.100)	(0.055)	(0.010)	(0.042)	(0.050)	(0.045)
Teaching Ability	0.402	0.261**	-0.245**	0.770**	0.600	-0.345	0.231	0.054	0.503***	0.177	-0.205
	(0.548)	(0.101)	(0.123)	(0.342)	(0.370)	(0.434)	(0.204)	(0.060)	(0.167)	(0.274)	(0.164)
Grandmother Rearing[1]	0.032	-0.002	-0.027	0.437*	0.436**	0.006	0.283*	0.100**	0.336**	0.793***	0.117
	(0.255)	(0.062)	(0.071)	(0.258)	(0.176)	(0.225)	(0.155)	(0.045)	(0.162)	(0.209)	(0.088)
Monthly Age	-0.057***	-0.007	0.007	-0.012	-0.018	0.025	0.032*	-0.001	0.002	0.067**	-0.011
	(0.012)	(0.004)	(0.010)	(0.012)	(0.021)	(0.026)	(0.018)	(0.003)	(0.009)	(0.032)	(0.011)
Constant	2.309***	1.033***	0.970***	1.726***	2.463***	2.666*	0.516	1.025***	1.222**	-1.019	1.601***
	(0.762)	(0.259)	(0.311)	(0.540)	(0.687)	(1.541)	(0.498)	(0.095)	(0.532)	(0.884)	(0.453)
Cragg-Donald F	43.494	34.803	22.807	43.648	48.213	96.371	49.372	36.137	54.441	34.974	17.043
Kleibergen-Paap LM	65.949	62.963	43.384	53.898	55.824	89.574	72.079	54.228	90.675	34.408	52.252
Hansen J	1.962	5.604	2.901	0.858	0.779	3.639	3.913	0.754	2.669	1.392	2.451
P-Value(Hansen J)	0.375	0.061	0.234	0.651	0.678	0.162	0.141	0.686	0.263	0.499	0.294

1. % of home visits when grandmother is the primary caregiver.
2. The estimates reported in the table are based on the instrumental variable regression.
3. The variables of teaching ability, interaction quality between home visitor and caregiver (child) are latent factors based on the supervisor recorded measures. See Appendix C.
4. The instrumental variables include mean, max, and min of other village interaction measures through the same home visitor.
5. Time to Mastery is defined as the number of tasks a child takes at the previous difficulty level until the first success (inclusive) at each difficulty level by each skill type.
6. For the first stage, we report Crag-Donald F statistics and Kleinbergen-Paap LM statistics. For overidentification test, we report Hansen J statistic and the p-value of Hansen J statistic.
7. Standard errors are reported in parentheses and clustered at village level.
8.* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

In general, the estimated impacts of interactions between home visitors and caregivers on improving children's skills are positive and statistically significant. Estimated impacts of interactions between home visitors and children are generally not significant, nor is the teaching ability of the visitor. To control for any endogeneity that biases home visitor's interactions with children's latent skills by skill level, we measure interaction outcomes for the same home visitor with children living in different spatially separated villages to construct instruments for the quality of home visitor interaction.[16] When we instrument for home visitor interactions, we find stronger point estimates.

Appendix D reports comparable results for other skills and other measures of knowledge. The interventions have no impact on gross motor skills. In it, we also evaluate the impact of interventions on post-treatment caregiver interactions with children. We measure the frequency of the caregiver playing with the child on the tasks after each home visit.[17] The intervention only promotes the frequency of caregivers play with low-ability children.[18]

5 Nonparametric Tests of a Version of Dynamic Complementarity

Dynamic complementarity arises if early investments affect the productivity of later investments. It governs the extent to which investment at later ages can substitute

[16] The instrumental variables include mean, max, and min of other village interaction measures through the same home visitor. Details are presented in Appendix D.

[17] Specifically, we record the following information: the number of days in a week that the caregiver plays with the child using tasks from the last home visit.

[18] See Table D.17.

for deficient early childhood investment. Heckman and Zhou (2022b) present formal tests of this proposition on our data and find evidence consistent with it. In this section, we present some additional nonparametric evidence that does not require any particular assumption about scales of skills except the maintained assumption of comparability of knowledge *within* skill levels. We use passing rates as our measure of knowledge.

Although children enter the program at different ages, all enrolled children of the same age receive the same lesson. We determine whether late entrants can catch up. This is an aspect of dynamic complementarity: how rapidly do children who enter the program later improve their skills compared to those who entered earlier and had some skill training. Over the age range of 10-25 months, children enter the program more or less randomly with respect to age due to administrative constraints.[19]

None of the children receive training in the program before entry but may acquire skills from home instruction, imitation and maturation. Suppose that a child enters at level $\ell(s)$ for skill s at age $a^+(s,\ell)$. Some may be able to master the task from the outset, but many do not. We compute the probability of mastery for new entrants at entry age $a^+(s,\ell)$ as

$$q(s,\ell,a^+(s,\ell)) = \Pr\left(D(s,\ell,a^+) = 1\right)$$

where age-appropriate lessons are administered at or near a^+. To test this, we use as new entrants children who enroll in the program who have less than one month of exposure to it. $q(s,\ell,a^+(s,\ell))$ is a measure of learning from maturation and exposure

[19]See Figure E.1.

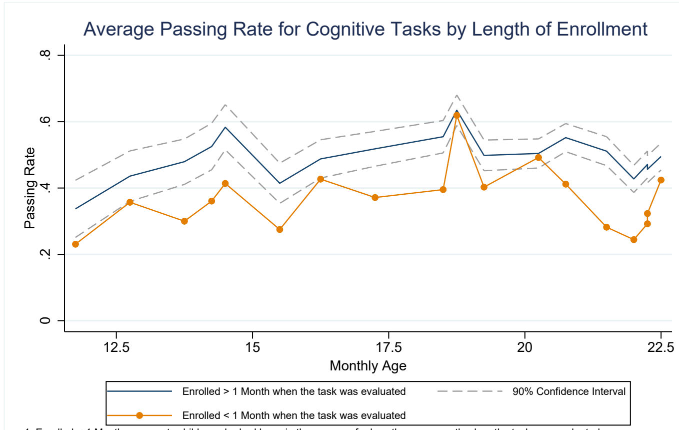
without participating in the program.

We consider performance by age at entry-level. Figure 3 shows the initial passing rate (q) for cognitive tasks by age (length of enrollment). It indicates that the knowledge of those not previously enrolled in the program (i.e., the ones who enrolled less than one month) is less than that of children at the same age who were enrolled in the program longer than one month. For most tasks, the group that is enrolled for longer than one month performs significantly better than new entrants. At the same ages, the endowments for children who just enroll in the program are smaller than those for the children in for longer spells.

Figure 4 compares the passing rates at the designated ages by different enrollment age groups. Since the curriculum is the same for all children of the same age, the ones who enrolled at older ages start at the curriculum with the same age-specific tasks. Therefore, they have shorter exposure of the intervention compared to children enrolled at younger ages. Consistent with dynamic complementarity, they start behind and generally stay behind. The longer the child has been in the program, the higher the passing rate on cognitive tasks. There are few entrants at later ages so we trim noisy data after the age of 30 months old. Comparable patterns appear for other skills.[20]

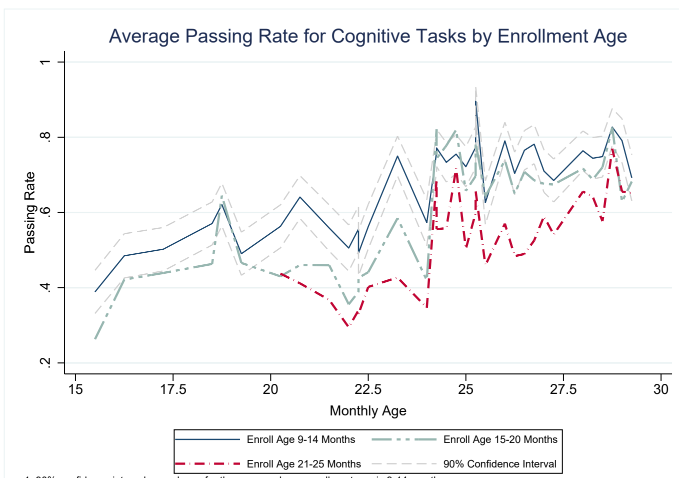
[20]See Appendix E.

Figure 3: Cognitive Tasks Performance Comparison by Length of Enrollment



1. Enrolled < 1 Month represents children who had been in the program for less than one month when the task was evaluated.
2. Enrolled > 1 Month represents children who had been in the program for more than one month when the task was evaluated, who continued to stay in the program for 2 years.
3. 90% confidence intervals are shown for both groups.
4. Tasks with fewer than 10 observations in either group are omitted.

Figure 4: Average Passing Rate for Cognitive Tasks by Enrollment Age



1. 90% confidence intervals are shown for the groups whose enrollment age is 9-14 months

Table 4: Cognitive Skills: Time to First Mastery by Enrollment Age

	Cognitive Difficulty Level										
	2	3	4	5	6	7	8	9	10	11	12
	Enroll (10-15) vs. (16-20)										
Mean (Age 10-15)	2.019	1.127	1.203	1.544	1.479	1.765	1.314	1.198	1.235	1.407	1.133
Mean (Age 16-20)	2.174	1.209	1.215	1.667	1.608	2.023	1.447	1.078	1.355	1.611	1.178
p-value (Age 10-15 v.s. 16-20)	0.336	**0.071**	0.830	0.368	0.216	**0.050**	0.119	**0.005**	0.309	0.190	0.628
step down p-value	0.834	0.447	0.850	0.834	0.755	0.363	0.588	**0.056**	0.834	0.737	0.850
N	384	268	221	442	427	431	416	249	192	167	103
	Enroll (10-15) vs. (21-25)										
Mean (Age 10-15)	2.019	1.127	1.203	1.544	1.479	1.765	1.314	1.198	1.235	1.407	1.133
Mean (Age 21-25)	1.385	1.194	1.156	2.118	2.161	2.424	1.877	1.100	1.483	1.576	1.275
p-value (Age 10-15 v.s. 21-25)	**0.002**	0.325	0.488	**0.001**	**0.000**	**0.000**	**0.000**	**0.033**	**0.047**	0.268	0.172
step down p-value	**0.012**	0.604	0.604	**0.008**	**0.000**	**0.001**	**0.000**	0.154	0.169	0.604	0.475
N	264	189	173	395	389	387	373	236	211	191	132
	Enroll (16-20) vs. (21-25)										
Mean (Age 16-20)	2.174	1.209	1.215	1.667	1.608	2.023	1.447	1.078	1.355	1.611	1.178
Mean (Age 21-25)	1.385	1.194	1.156	2.118	2.161	2.424	1.877	1.100	1.483	1.576	1.275
p-value (Age 16-20 v.s. 21-25)	**0.001**	0.851	0.413	**0.009**	**0.000**	**0.017**	**0.000**	0.518	0.225	0.795	0.209
step down p-value	**0.013**	0.956	0.873	**0.068**	**0.001**	0.106	**0.003**	0.892	0.743	0.956	0.743
N	224	141	138	385	386	384	369	293	267	240	175

1. Group (10–15) represents children whose monthly ages are between 10 and 15 at enrollment.
2. Group (16–20) represents children whose monthly ages are between 16 and 20 at enrollment.
3. Group (21–25) represents children whose monthly ages are between 21 and 25 at enrollment.
4. Time to first mastery is defined as the number of tasks a child takes until the first success (inclusive) at each difficulty level during the intervention by skill type.
5. Step down p values are constructed by multiple hypotheses between the earlier enrolled group and later enrolled group based on Romano and Wolf (2005a,b).
6. Step down p-values are conducted by 5000 times of bootstrap.

Table 4 documents the test in detail. We ask whether those who start later catch up in terms of time to first mastery. In the vast majority of cases, they do not. In Heckman and Zhou (2022b), we show that dynamic complementarity does not operate uniformly across ability groups: normal and low-ability children display stronger dynamic complementarity effects, but high-ability children do not.

This analysis has to be qualified. If there are critical and sensitive early periods in the intervals missed by late entrants, our evidence is also consistent with that phenomena, as well as with dynamic complementarity.

To this point, we have presented empirical evidence that learning exists, the intervention boosts skill development mainly through the interactions between home visitors and caregivers, and our data are consistent with dynamic complementarity. We next develop a dynamic model to formalize these findings.

6 Mechanisms Generating Child Learning

To motivate our approach to estimating the weekly dynamics of skill formation, we consider a simple model for one level of skill before presenting our general model. The more general model is the simple model applied to each skill at each level.

The program fosters skill at ages $a \in [0, \ldots, \bar{A}]$. Lessons are the same for all participants at age a. Define $K(a)$ as the level of "knowledge" at age a with the initial value $K(0)$. Lessons with identical skill content are taught and examined using a series of tasks. A person exhibits *mastery* of a skill at level $\bar{K}$ if $K(a) \geq \bar{K}$. Let $D(a) = 1$ if a person at age a masters the skill, so $D(a) = \mathbf{1}\big(K(a) \geq \bar{K}\big)$.

Mastery is measured at each age.

Consider a deterministic model of skill formation. Skill evolves via

$$K(a) = K(a-1) + \delta(a)\eta K(a-1) + V(Q(a)), \tag{6}$$

where η is an ability to learn parameter that is individual specific and assumed positive ($\eta > 0$), and $\delta(a)$ is the "lesson" at age a for everyone enrolled. $V(Q(a))$ captures variables $Q(a)$, such as family background and investments received at home, as well as maturation effects through ages that affect the evolution of skills. And, $V(Q(a))$ also operates independently of the level of $K(a-1)$. We assume skill invariance *within* each designated skill level.[21] Skills are additive in the metric that quantifies K.

In this framework, *Self-Productivity* is $\dfrac{\partial K(a)}{\partial K(a-1)} = 1 + \delta(a)\eta$. *Investment Productivity* is $\dfrac{\partial K(a)}{\partial \delta(a)} = \eta K(a-1)$. *Static Complementarity* between skills and investment at age $a-1$ defined as: $\dfrac{\partial^2 K(a)}{\partial K(a-1)\partial \delta(a)} = \eta > 0$. *Dynamic complementarity* arises from investment at age a on the productivity of future investments is defined as $\frac{\partial^2 \boldsymbol{K(a+j+1)}}{\partial \boldsymbol{\delta(a)}\partial \boldsymbol{\delta'(a+j)}}$.[22]

Adding Shocks

A multiplicative version of the model turns out to fit the data on skill growth

[21]The tasks within each difficulty level are essentially the same.

[22]Heckman and Zhou (2022b) show that dynamic complementarity can be affected by (a) complementarity between skills and investment in period $a+j$, (b) self-productivity (e.g., the marginal productivity of investment), and (c) the transmission of period a investment to latent skills in period $a+j+1$.

very well. [23] Adding i.i.d. idiosyncratic shocks in growth rates ($\varepsilon(a)$) on a log scale, skill acquisition is characterized by:

$$\ln K(a) - \ln K(a-1) \doteq \delta(a)\eta + V(Q(a)) + \varepsilon(a). \tag{7}$$

Accounting for initial conditions, Equation (7) becomes:

$$\ln K(a) \doteq \eta \sum_{j=1}^{a} \delta(j) + \sum_{j=1}^{a} V(Q(j)) + \sum_{j=1}^{a} \varepsilon(j) + \ln K(0) \tag{8}$$

where $\varepsilon(a)$ is i.i.d. across all a with $E(\varepsilon(a)) = 0$. The model exhibits dynamic complementarity, self-productivity and investment productivity. It introduces random walk growth in skill levels following Rutherford (1955).

Adding stochastic shocks to learning growth accounts for fadeout or acceleration off deterministic growth paths. The entire literature on fadeout of test scores (see, e.g., Duncan et al., 2022) assumes deterministic growth profiles. We allow for stochastic growth and fadeout of measured skill within a lifetime.

Define $U(a) = \sum_{j=1}^{a} \varepsilon(j)$, a random walk, $\Delta(a) = \sum_{j=1}^{a} \delta(j)$ is cumulative lessons, and $\Lambda(a) = \sum_{j=1}^{a} V(Q(j))$. In this notation, the probability of mastery of the skill at age a is $\Pr(D(a) = 1) = \Pr(\ln K(0) + U(a) + \Lambda(a) + \eta\Delta(a) > \ln \bar{K})$, where we assume $\eta \perp\!\!\!\perp \varepsilon(j)$ for all j so shocks are from the same distribution and independent of ability level. Conditioning on η, assumed to be independent of $U(a)$ and $K(0)$, we obtain

[23] Heckman and Zhou (2022c) compare the empirical performance of multiplicative and additive models. In many aspects, the qualitative results from each are very similar but quantitative results are somewhat better for the multiplicative model as characterized by model specification tests.

$$\Pr(D(a) = 1 \mid \eta, \Delta(a), \Lambda(a), K(0)) = \int_{\ln \bar{K} - \eta\Delta(a) - \Lambda(a) - \ln K(0)}^{\infty} dF(U(a)). \tag{9}$$

The General Model

Using the notation introduced in Section 3, Equation (2), the general model has the same structure as the simple model applied to skills at each level where $\mathcal{S}$ is the set of skills taught, $\ell(s,a)$ is the level of skill s taught at age a, and $\ell(s,a) \in \{1, \ldots, L_s\}$, where L_s is the number of levels of difficulty for each skill s.

Shocks at level ℓ for age a—$\varepsilon_\ell(s,a)$—are assumed to be independent across a. Their distributions may vary with ℓ and a. When estimating the model, we assume that they are i.i.d. within ℓ. $\eta(s)$ may vary by age a[24] and $\delta(a)$ captures the content of the curriculum. Thresholds (passing standards) $\bar{K}(s,\ell)$ may also change across levels, as may $V_\ell(Q(a))$.

This model is a contribution to mathematical psychology. It unites and extends two fundamental psychometric models: the Item Response Theory (IRT) model (Lord and Novick, 1968) and the Bayesian Knowledge Tracing (BKT) model (Corbett and Anderson, 1994). The essential feature of the IRT model is captured by the threshold crossing feature (2). The BKT model is captured by the dynamics of the model (6). Unlike the BKT model, knowledge $\boldsymbol{K(a)}$ in our model is affected by education and investment, which is captured by $\boldsymbol{I(a)}$, so that we depart from its mechanical growth trajectory feature to account for investment that affects

[24] In the estimation, η includes the interaction measures and the measure of grandmother appearance. Therefore, η changes as lessons change.

learning.[25]

By allowing for level-specific shocks, we account for the possibility that different difficulty levels within an assessment may have different variances and thresholds. This is indeed what we find in our estimates. We can explain "fadeout of measured skills within a lifetime" by allowing for level-specific differences in difficulty of an assessment and level-specific responses of test takers. We next define two notions of skill invariance and show how we test for it within our model.

6.1 Testing Skill Invariance

As previously noted, there are two different interpretations of invariance of skill. The first interpretation, which is examined in this paper, is that there exist latent skills that generate model outcomes, and they are comparable across ages and levels of skill. This assumption underlies all human capital models since Ben-Porath (1967). The second interpretation is that there are invariant *measures* of skill comparable across skill levels (Agostinelli and Wiswall, 2022). The existence of scale-invariant measures in this sense requires invariant latent skills in the first sense.[26]

Skill invariance in the first sense assumes a common scale within and across all difficulty levels ℓ for each skill type s, although scales may vary across s. Heckman and Zhou (2022a) conduct nonparametric tests of skill invariance of measures (con-

[25] Deonovic et al. (2018) compare the IRT and BKT models and criticize them for not including investment as a determinant of learning.

[26] Agostinelli and Wiswall define invariant measures in the following way. Let $K_i(a)$ be child i's human capital in the first sense. Let $Z_i(m, a)$ be child i's score on a measure of $K_i(a)$ at age a. Invariant measures is defined as: $E(Z_i(m, a) \mid K_i(a) = \tau) = E(Z_{i'}(m, a + t) \mid K_{i'}(a + t) = \tau)$; $i \neq i'$, all t. This is a property of a *measure* of a latent skill, assumed to be skill invariant in the first sense.

stant units across levels) using passing rates within narrowly defined levels of tests as measures of knowledge, assuming skill invariance within the same levels. Cognitive and language aggregate Denver test scores are not the same for each group with identical knowledge as measured by passing rates within the same levels. These findings are robust across almost all the levels of knowledge. This paper develops and applies a model-based test of skill invariance in the first sense and rejects this assumption as well for most skills and most levels.

Under skill invariance in the first sense, index $K(s, \ell, a)$ cumulates *across* levels, so the measures of knowledge growth are well-defined. This requires, among other things, that in the absence of depreciation (or appreciation) associated with transitions across levels,

$$\underbrace{K(s, \ell, \underline{a}(s, \ell))}_{\text{Initial condition at level } \ell} = \underbrace{K(s, \ell - 1, \bar{a}(s, \ell - 1))}_{\text{Terminal condition at level } \ell-1} .$$

This is a property of latent variables. If all components of the technology of skill formation (Equation (1)) shift across levels, the hypothesis of skill-invariant scales lacks testability because the scale is not directly observed, and technology parameters can be redefined to impose invariance. Some parameters must be invariant across levels to conduct this test, although they need not necessarily be the same parameters across all levels. We test for skill invariance in the first sense, maintaining the assumption of invariance *within the same levels*. Our proof of model identification in Appendix F makes this point precise. Note that the assumed lack of depreciation (or appreciation) is only a property at boundaries. There can exist either (or both) in interior segments.

If scales change across levels, but human capital scales are somehow connected, it follows that

$$K(s, \ell, \underline{a}(s, \ell)) = \Gamma_\ell(K(s, \ell - 1, \bar{a}(s, \ell - 1))),$$

where Γ_ℓ is a general function. If there is total depreciation of skills in transitions from $\ell - 1$, Γ_ℓ is the zero function. Skill invariance in the first sense at $\ell - 1$ sets $\Gamma_\ell = I$, the identity function. In this paper, we only consider affine transformations for $\Gamma_\ell(\cdot)$:

$$\Gamma_\ell(K(s, \ell, \underline{a}(s, \ell))) = \gamma_{0,\ell} + \gamma_{1,\ell}(K(s, \ell - 1, \bar{a}(s, \ell - 1))). \tag{10}$$

Setting $\gamma_{0,\ell} = 0$ and $\gamma_{1,\ell} = 1$ captures the notion of skill invariance. More general transformations are admissible but we use the affine transformation as a first order linear approximation of the general function.

We now present the intuition for how we can test for skill invariance in the first sense. We do not have direct measures of latent skills. Instead, we have strings of binary task performances for children enrolled in the program, from which we can infer their skills up to scale as in the standard binary threshold crossing model (see, e.g., Matzkin, 1992).

6.2 Model Identification

In order to avoid notational complexity, we use a simplified notation for a single skill to motivate essential ideas underlying model identification. A formal proof is presented in Appendix F. We use means and covariances because we assume normal errors in estimation. However, drawing on Heckman and Vytlacil (2007) and Matzkin (1992, 2007), we show in Appendix F that we can nonparametrically identify the joint distributions of unobserved variables up to normalizations under conditions stated in those papers.

Define the latent index $K(1,a)$ for skill at level 1 at age a. This corresponds to $K(s,1,a)$ for a particular skill s, which is kept implicit. We simplify Equation (8) to read:

$$\ln K(1,a) = \underbrace{\eta \sum_{j=1}^{a} \delta_1(j)}_{\text{learning}} + \underbrace{V_1(a)}_{\substack{\text{maturation and}\\ \text{exposure effects}}} + \underbrace{U_1(a)}_{\text{shocks}} + \ln K(0), \tag{11}$$

where $K(1,a)$ is the latent index (skill) of a binary outcome model at difficulty level 1 at weekly age a, and $K(0)$ is the initial condition. $\ln K(0) = \mu_0(Z) + \Upsilon$, where Z are background variables, $E(\Upsilon) = 0$, $\Upsilon \perp\!\!\!\perp \eta$, and $Z \perp\!\!\!\perp \Upsilon$. $U_1(a) = \sum_{j=1}^{a} \varepsilon_1(j)$, where $\varepsilon_1(j)$ is a task-specific shock at difficulty level 1 at weekly age j, which is assumed to be i.i.d. with variance $\sigma^2_{\varepsilon(1)}$. We assume that $\varepsilon_1(j) \perp\!\!\!\perp (\eta, \Upsilon)$ for all j. We parameterize $\delta_1(a)\eta(X) = \bar{\beta}_1(X) + \omega$, where X are covariates including ability and interactions. $X \perp\!\!\!\perp [\omega, \varepsilon_1(j)]$ for all j. ω is an individual-specific random shock, with $E(\omega) = 0$, and $\omega \perp\!\!\!\perp (\Upsilon, \varepsilon_1(j))$ for all j. It captures heterogeneity in learning ability. To simplify the analysis, we assume that $\omega_\ell = \omega$ for $\ell \in \{1, \dots, L\}$. We can relax this assumption and still achieve identification. However, we have to take a position on the dependence across ω_j.[27] We assume that the learning component $\delta_1(a)$ is constant within each level but can differ across levels. $V_1(a)$ is shorthand for $\sum_{j=1}^{a} V_1(Q(j))$.

[27] One attractive alternative assumption that secures identification is $\omega_j = \rho\omega_{j-1} + \tau_j$, where τ_j is mean zero, i.i.d over j.

Equation (11) can be rewritten in the notation for the general case allowing for heterogeneity in $\ln K(0)$:

$$\ln K(1,a) = \mu_1 + \mu_0(Z) + V_1(a) + \bar{\beta}_1(X)a + \underbrace{\left\{ a\omega + \sum_{j=1}^{a} \varepsilon_1(j) + \Upsilon \right\}}_{\Psi_1(a)} \tag{12}$$

where $\text{Var}(\Psi_1(a)) = a^2\sigma_\omega^2 + a\sigma_{\varepsilon(1)}^2 + \sigma_\Upsilon^2 := \sigma^2(1,a)$, where $\sigma^2(1,1) = \sigma_\omega^2 + \sigma_{\varepsilon(1)}^2 + \sigma_\Upsilon^2$.

Under conditions given in Matzkin (1992, 2007), with sufficient variation in the regressors in period j, $\underline{a}(1) \leq j \leq \bar{a}(1)$, we can identify

$$\frac{\mu_1^*}{\sigma(1,j)}, \quad \frac{\mu_0(Z)}{\sigma(1,j)}, \quad \frac{\bar{\beta}_1(X)}{\sigma(1,j)}, \quad \frac{V_1(a)}{\sigma(1,j)},$$

where $\mu_1^* = \mu_1 - \bar{K}(1)$ and μ_1 collects any other model intercepts. If any slope coefficient is common across j and j', we can identify the ratio of $\frac{\sigma(1,j)}{\sigma(1,j')}$. Under this condition, with one normalization (e.g., $\sigma(1,j) = 1$), we can identify μ_1^*, $\mu_0(Z)$, $\bar{\beta}_1(X)$, $V_1(a)$ up to scale. Since we can identify the ratio of $\frac{\sigma(1,j)}{\sigma(1,j')}$, $\sigma(1,a)$, $\sigma(1,a')$ are identified up to a normalization (e.g., $a, a' \neq j$) (see Heckman, 1981 and Heckman and Vytlacil, 2007).

Using the definition of $\sigma^2(1,a) := a^2\sigma_\omega^2 + a\sigma_{\varepsilon(1)}^2 + \sigma_\Upsilon^2$, we have the following equations:

$$\begin{aligned}
\sigma^2(1,a) &= a^2\sigma_\omega^2 + a\sigma_{\varepsilon(1)}^2 + \sigma_\Upsilon^2 \\
\sigma^2(1,a') &= (a')^2\sigma_\omega^2 + a'\sigma_{\varepsilon(1)}^2 + \sigma_\Upsilon^2 \\
\sigma^2(1,j) &= j^2\sigma_\omega^2 + j\sigma_{\varepsilon(1)}^2 + \sigma_\Upsilon^2.
\end{aligned}$$

In these equations, the left-hand sides are identified up to the scale (i.e., $\sigma^2(1,j)=1$). On the right-hand sides, there are three unknown terms σ^2_ω, $\sigma^2_{\varepsilon(1)}$, and σ^2_Υ. When $a \geq 3$ (i.e., three different tasks at level one), we can identify all three terms: σ^2_ω, $\sigma^2_{\varepsilon(1)}$, and σ^2_Υ with sufficient variation in a and j.

Adopting a similar notation for levels $\ell > 1$, if we assume skill invariant measures connecting level 1 with level 2 (i.e., $\gamma_{0,2}=0$, and $\gamma_{1,2}=1$), we can connect latent skill $\ln K(1,\bar{a}(1))$ (the index of the last age $\bar{a}(1)$ of the last task at level 1) to the initial skill at level 2, $\ln K(2,\underline{a}(2))$: $\ln K(1,\bar{a}(1)) = \ln K(2,\underline{a}(2))$. The latent skill at level 2 at age a can be written as:

$$\begin{aligned}\ln K(2,a) &= \mu_2 + V_2(a) + \bar{\beta}_2(X)(a-\bar{a}(1)) + \sum_{j=\underline{a}(2)}^{a} \varepsilon_2(j) + \ln K(1,\bar{a}(1)) \\ &= \mu_1 + \mu_2 + \mu_0(Z) + V_1(\bar{a}(1)) + V_2(a) + \bar{\beta}_2(X)(a-\bar{a}(1)) + \bar{\beta}_1(X)\bar{a}(1) \\ &+ \underbrace{\left\{ \sum_{j=\underline{a}(2)}^{a} \varepsilon_2(j) + (a-\bar{a}(1))\omega + \sum_{j=1}^{\bar{a}(1)} \varepsilon_1(j) + \bar{a}(1)\omega + \Upsilon \right\}}_{\Psi_2(a)}. \end{aligned} \tag{13}$$

Given the initial normalization at level 1 (i.e., $\sigma(1,j)=1$) and identification of the parameters in the first level (up to scale), we can identify $V_2(a)$ and $\bar{\beta}_2(X)$ up to scale $\sigma(2,a)$, where

$$\begin{aligned}\Psi_2(a) &= \sum_{j=\underline{a}(2)}^{a} \varepsilon_2(j) + (a-\bar{a}(1))\omega + \sum_{j=1}^{\bar{a}(1)} \varepsilon_1(j) + \bar{a}(1)\omega + \Upsilon \\ \sigma^2(2,a) &:= \mathrm{Var}\Psi_2(a) \\ \mathrm{Var}\Psi_2(a) &= \sigma^2_\Upsilon + a^2\sigma^2_\omega + (a-\underline{a}(2))\sigma^2_{\varepsilon(2)} + \bar{a}(1)\sigma^2_{\varepsilon(1)}. \end{aligned}$$

Notice that we have identified σ^2_ω, $\sigma^2_{\varepsilon(1)}$, and σ^2_Υ, and the only term not identified in $\mathrm{Var}\Psi_2(a)$ is $\sigma^2_{\varepsilon(2)}$. We now discuss how to identify this term. Consider the covariance term $\mathrm{Cov}\left(\frac{\Psi_2(a)}{\sigma(2,a)}, \frac{\Psi_2(a')}{\sigma(2,a')}\right)$

$$
\begin{aligned}
\mathrm{Cov}\left(\frac{\Psi_2(a)}{\sigma(2,a)}, \frac{\Psi_2(a')}{\sigma(2,a')}\right) &= \frac{\sigma^2_\Upsilon + aa'\sigma^2_\omega + (\bar{a}(1) - \underline{a}(1))\sigma^2_{\varepsilon(1)} + \min((a - \underline{a}(2)), (a' - \underline{a}(2)))\sigma^2_{\varepsilon(2)}}{\sigma(2,a)\sigma(2,a')} \\
&= \frac{\sigma^2_\Upsilon + aa'\sigma^2_\omega + (\bar{a}(1) - \underline{a}(1))\sigma^2_{\varepsilon(1)} + \min((a - \underline{a}(2)), (a' - \underline{a}(2)))\sigma^2_{\varepsilon(2)}}{\sqrt{\sigma^2_\Upsilon + a^2\sigma^2_\omega + (a - \bar{a}(1))\sigma^2_{\varepsilon(2)} + \bar{a}(1)\sigma^2_{\varepsilon(1)}}\sqrt{\sigma^2_\Upsilon + (a')^2\sigma^2_\omega + (a' - \bar{a}(1))\sigma^2_{\varepsilon(2)} + \bar{a}(1)\sigma^2_{\varepsilon(1)}}}
\end{aligned}
$$

In the equation just written, we observe the left-hand side value. On the right-hand side, the only unknown term is the variance of shocks at level 2 (i.e., $\sigma^2_{\varepsilon(2)}$). Therefore, we can identify the value of $\sigma^2_{\varepsilon(2)}$. After identifying $\sigma^2_{\varepsilon(2)}$, we can identify the scale of variance term $\sigma^2(2,a)$. Then, we can identify $V_2(a)$ and $\bar{\beta}_2(X)$ up to $\sigma(2,a)$.

From the previous discussion, we can identify, for all $\ell \geq 2$, the variance $\sigma(\ell, a)$ without imposing additional normalization at levels ℓ ($\ell \geq 2$). The only normalization we need is on the scale of variance term $\sigma(1,j) = 1$ at level 1.

Under conditions established in Matzkin (2007) and Heckman and Vytlacil (2007), we can nonparametrically identify the distributions of $\varepsilon_1(a)$ and $\varepsilon_2(a')$ for each a and a' in the appropriate intervals and the technologies at each level subject to the initial normalization. Details concerning nonparametric identification are discussed in Appendix F. We do not develop this point further because we adopt parametric models in making our estimates. The conditions just developed extend in a straightforward way to higher levels, $\ell > 2$. All higher-level parameters are identified up to the initial normalization at level 1.

6.2.1 Testing the Skill Invariance Assumption

Under skill invariance characterized by Equation (10) with $\gamma_{0,\ell} = 0$ and $\gamma_{1,\ell} = 1$, we obtain tight restrictions on the coefficients across levels. Relaxing scale invariance adds two new parameters $(\gamma_{0,2}, \gamma_{1,2})$ to Equation (13):

$$\ln K(2,a) = \gamma_{0,2} + \mu_2 + V_2(a) + \bar{\beta}_2(X)(a - \bar{a}(1)) + \sum_{j=\underline{a}(2)}^{a} \varepsilon_2(j) + \gamma_{1,2} \ln K(1, \bar{a}(1)).$$

Notice that scale invariance imposes a proportionality restriction across functions common to $\ln K(2,a)$ and $\ln K(1,a)$. Going across levels,

$$\mathrm{Cov}\left(\frac{\Psi_2(a)}{\sigma(2,a)}, \frac{\Psi_1(a')}{\sigma(1,a')}\right) = \gamma_{1,2}\left\{aa'\sigma_\omega^2 + (a' - \underline{a}(1))\sigma_{\varepsilon(1)}^2 + \sigma_\Upsilon^2\right\}\frac{1}{\sigma(2,a)\sigma(1,a')},$$
$$a > \bar{a}(1);\ \underline{a}(1) \leq a' < \bar{a}(1).$$

From the previous analysis, the term in braces is identified up to the previously stated normalization at the first level. Thus $\gamma_{1,2}$ is identified, and we can test if $\gamma_{1,2} = 1$. Testing $\gamma_{0,2} = 0$ requires stronger assumptions. We need model intercepts to be invariant, which is difficult to maintain given that $\bar{K}(2)$ is absorbed in any estimated intercept, and we expect that the difficulty levels are increasing in ℓ. As before, we can estimate $\ln \bar{K}(2)$ up to scale net of intercepts, and we can identify the scale.

7 Estimation Results

We use the method of simulated moments to estimate the model for each specific skill s. We adjust for clustering in our sample using the paired cluster bootstrap. Details are provided in Appendix G. The moments used in forming the estimates are presented in Table H.1. The model passes goodness of fit tests (see Appendix H).[28] Appendix H also plots model predictions vs. data for each skill, with and without skill invariance[29]. In general, imposing the skill invariance assumption produces worse fits, a point developed further below. We report estimates in the text that do not impose skill invariance. Estimates imposing skill invariance are presented in Appendix I.

7.1 Estimates

We report our empirical results by skill level.

7.1.1 Language Skills

Figure 5a displays estimates of the minimum skill level required at each level. This is defined relative to $\bar{K}(1)$, assuming no shift in model intercepts for each skill across levels apart from that due to skill accumulation. As expected, the skill level required to pass tasks monotonically increases across difficulty levels. We do not impose any restriction on the order of the $\bar{K}(\ell)$. The estimates show that, on average, the

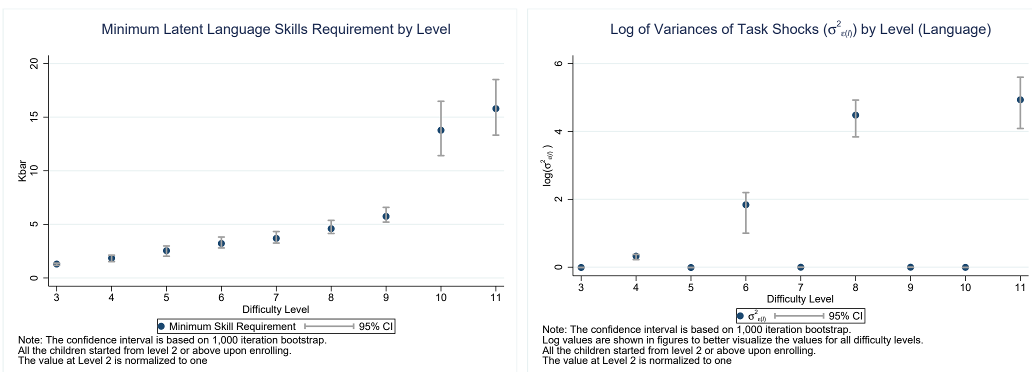
[28]When we separate estimates by gender, we find no differences in the structural parameters. The initial conditions favor girls and that explains their better scores on the tests. See Zhou et al. (2022).

[29]See Figures H.1, H.7, and H.13 for language, cognition, and fine motor skills, respectively.

difficulty levels in the curriculum are consistent with child task performance. The variances of shocks at each level display different patterns, reflecting differentials in ability. Figure 5b presents estimates of the variances. The variances at levels 6, 8, and 11 are larger than the variances at other levels. We plot the task passing rates at these three levels in Figure 6, and we find that the large variances are associated with a larger range of passing rates. Passing rates do not monotonically increase by task order within the same level (see Figure 6). Level-specific shocks can intrude to alter the monotonicity delivered by the deterministic model and to capture the lack of fit of the model to the data.[30]

Figure 5: Language Skill

(a) $\bar{K}(\ell)$ (b) $\sigma^2_{\varepsilon(\ell)}$

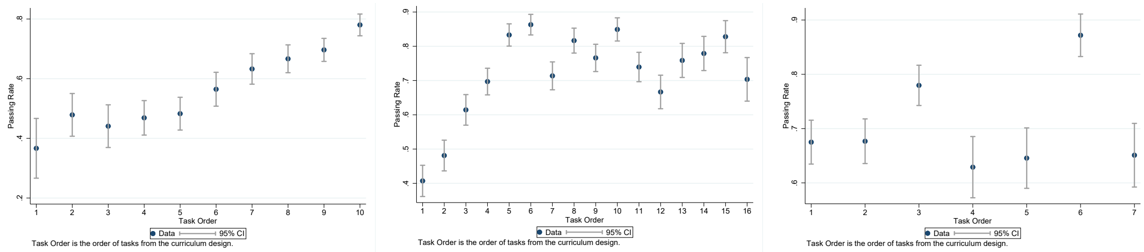


Note that “fadeout” as measured by passing rates appears within level 11 and across levels 6-11 as a consequence of patterns of item difficulties and variances. This occurs despite the stochastically monotonic increase in knowledge.

[30]See Figure H.1b.

Figure 6: Average Passing Rate of Language Tasks: $p(s, \ell)$ (Raw Data)

(a) Level 6 (b) Level 8 (c) Level 11

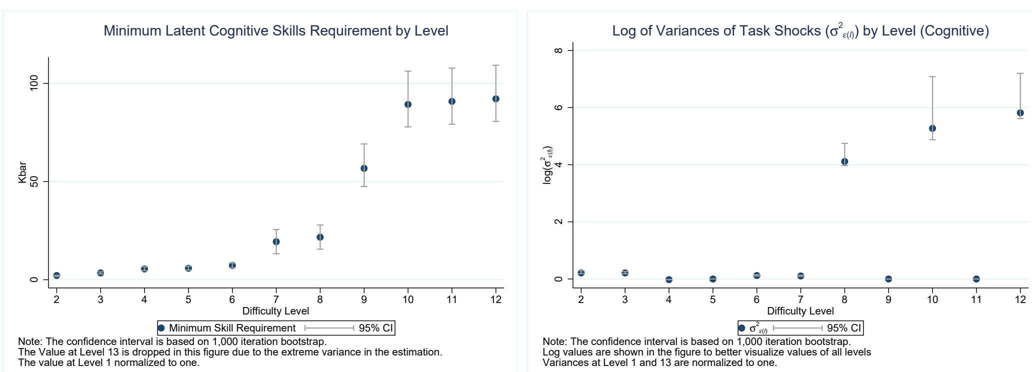


7.1.2 Cognitive Skills

The pattern for the estimated parameters for cognitive skills is similar to that for language skills. For certain difficulty levels, passing rates are not monotone within levels, thus explaining "fadeout" even when, on average, skill levels are increasing.

Figure 7: Cognitive Skill

(a) $\bar{K}(\ell)$ (b) $\sigma^2_{\varepsilon(\ell)}$



7.1.3 Fine Motor Skills

A similar pattern arises for fine motor skills.

Figure 8: Average Passing Rate of Cognitive Tasks: $p(s, \ell)$ (Raw Data)

(a) Level 8 (b) Level 10 (c) Level 12

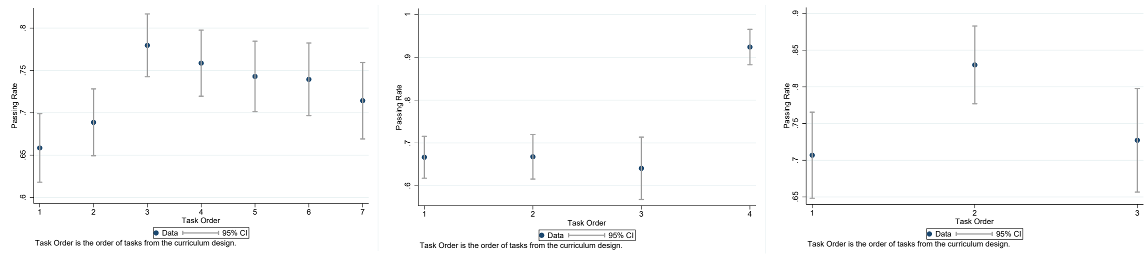


Figure 9: Fine Motor Skill

(a) $\bar{K}(\ell)$ (b) $\sigma^2_{\varepsilon(\ell)}$

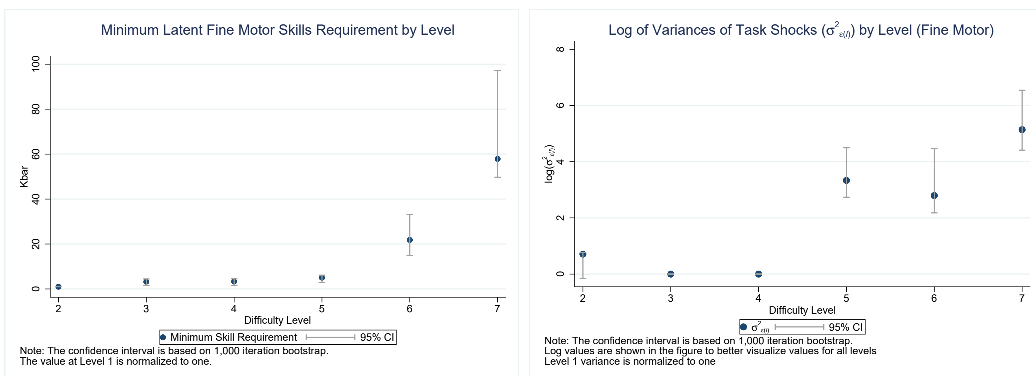
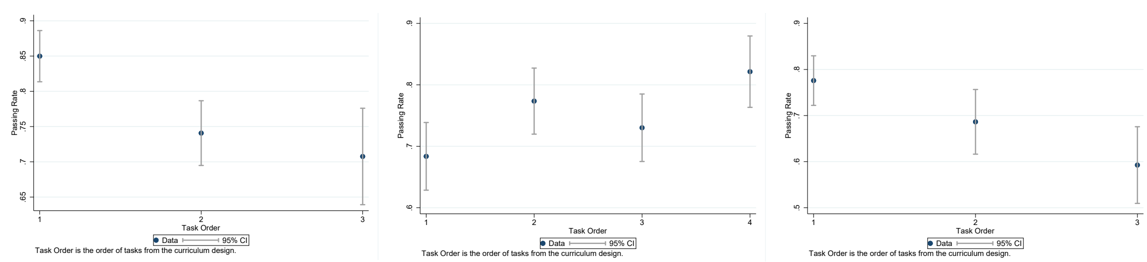


Figure 10: Average Passing Rate of Fine Motor Tasks: $p(s, \ell)$ (Raw Data)

(a) Level 5 (b) Level 6 (c) Level 7



7.2 Learning Components and Task Performance

In this section, we examine how the learning component in our structural model $\delta_\ell E(\eta)$ explains child task performance. The δ_ℓ term captures the curriculum content at each difficulty level, which is common across all children. The $\eta(X)$ term includes interaction quality measures between home visitors and caregivers/children, home visitors' teaching quality, and grandmother rearing during the intervention.

The intervention interaction variables (entered as X in $\beta_\ell(X)$) are significant determinants of child learning for each task. This finding is consistent with the results reported in Section 4. The interaction between the home visitor and the caregiver is the only consistently positive interaction that promotes skills (see Appendix I).[31] The grandmother, as the main caregiver, often has significantly negative effects on learning.[32]

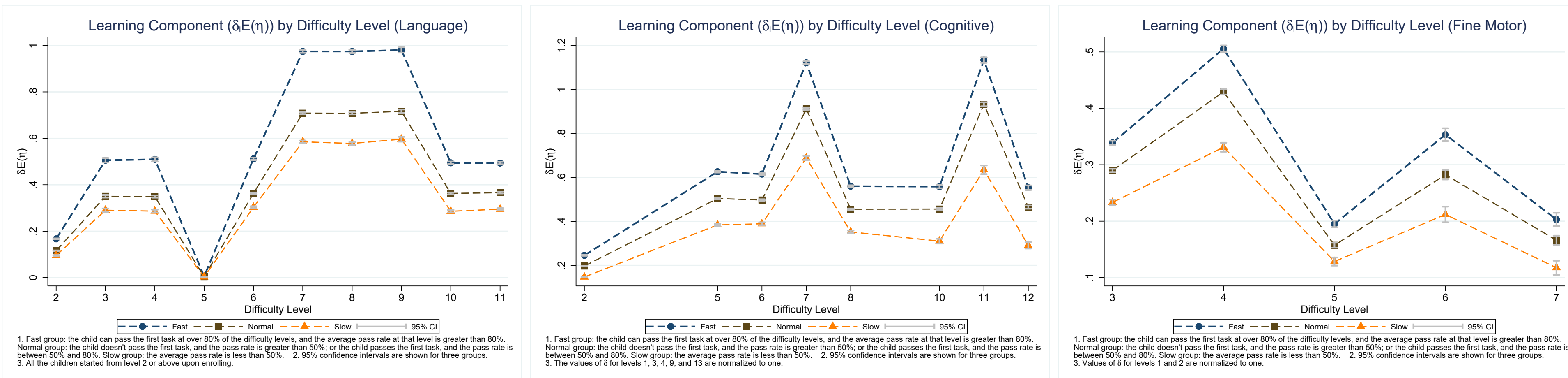
Rapid learning (high-ability) children have significantly higher values of the learning component during the intervention for all skills. This finding is consistent across all difficulty levels for all skills (see Figure 11). We also find that higher caregiver education levels are significantly associated with better language skills when children are first enrolled in the program (see Table I.1). There is substantial learning for children exposed to more educated mothers.

[31] All the estimation results are presented in Appendix I.

[32] Grandmothers' education is low on average (3 years).

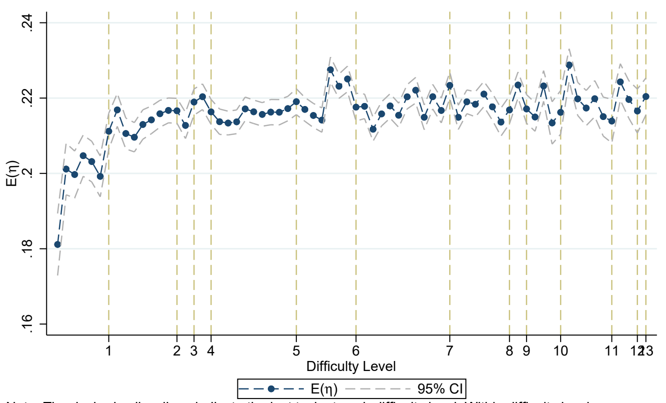
Figure 11: Estimates of $\delta_\ell E(\eta)$ by Ability Group

(a) Language (b) Cognitive (c) Fine Motor



* Intervals are of the form $(j-1, j)$. The parameter for the interval is indexed by the upper value, j.

Figure 12a: Learning Component $E(\eta(X))$ of Cognitive Tasks by Level



Note: The dashed yellow lines indicate the last task at each difficulty level. Within difficulty levels, tasks are arranged in the order of the curriculum design.

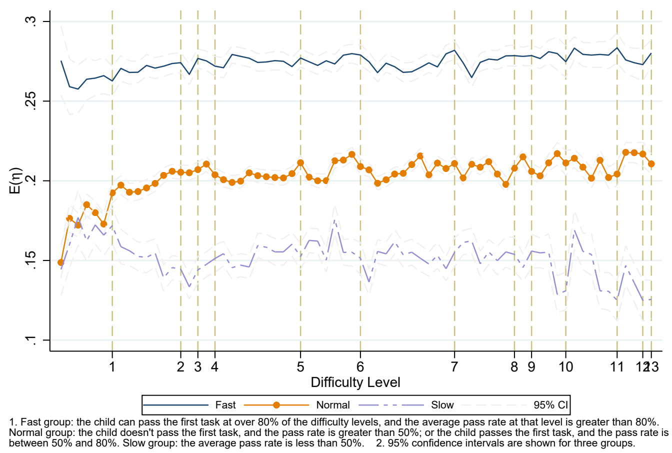
We now focus on how the $\eta(X)$ term affects child performance on tasks. Figure 12a shows the mean of $\eta(X)$ for each cognitive task. We identify it using β_ℓ and normalize $\delta(1) = 1$. There is an increasing pattern of $E(\eta)$ within difficulty levels. In Figure 12b, we break down the estimated $E(\eta)$ values by ability group.[33] Children in the normal ability group contribute the most growth in learning. Children in the fast group master the task quickly, usually on the first try. Thus, they have little subsequent learning growth when they are instructed on the same task multiple times. For children in the normal group, performance improves as they learn the task multiple times. This pattern is consistent with our estimates showing that the learning component $E(\eta)$ increases within a difficulty level, especially strongly for children in the normal group. This finding is also found for other skills.[34] For fine

[33]Fast group: the child passes the first task at over 80% of the difficulty levels, and the average passing rate at that level is greater than 80%. Normal group: the child does not pass the first task, and the passing rate is greater than 50%; or the child passes the first task, and the passing rate is between 50% and 80%. Slow group: the average passing is less than 50%.

[34]See Figures J.1-J.4 in Appendix J.

motor tasks, there is a similar pattern for tasks greater than 4, although learning is not substantial at any level.

Figure 12b: Learning Component $E(\eta(X))$ of Cognitive Tasks by Level and Ability Group



Appendix Tables J.1-J.3 compare each interaction component by family education background, child ability category, and age of enrollment. As expected, the interaction quality between the home visitor and caregiver contributes the most to the learning component η. The interaction quality between the home visitor and the caregiver is higher for the household with higher family education levels. Also, the interaction quality measures are significantly different by ability groups and age of enrollment.

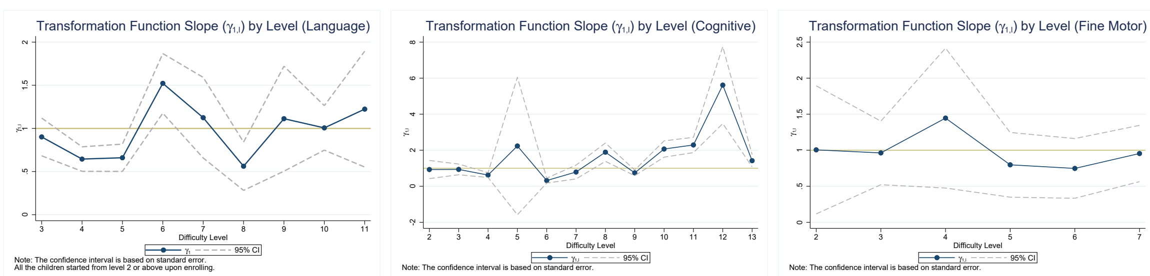
7.3 Testing Skill Invariance in the First Sense

Under our parameterization of skill invariance in the first sense, $\gamma_{1,\ell} = 1$. Note that $\gamma_{1,\ell} = 1$ implies the validity of a constant-unit latent skill between ℓ and $\ell - 1$.

Figure 13a shows that estimates of $\gamma_{1,\ell}$ for each skill level for models estimated without imposing the restriction $\gamma_{1,\ell} = 1$. Table 5 shows the χ^2 test results for each level and skill. Our estimates partially support skill invariance. For language and cognitive skills, at some levels, skill invariance cannot be rejected. For example, we cannot reject skill invariance for language skills between levels 8-11 (i.e., 8-9, 9-10, and 10-11).[35] However, it is decisively rejected at levels 4-6. Table 6 lists the task content for difficulty levels 8-11; it shows that the task content is very similar across these different levels. However, the null hypothesis of skill invariance across all levels is rejected. The evidence for skill invariance across levels 8-9, 9-10, and 10-11 makes sense given the similarity of the tasks at those levels.

Figure 13: Tests of the Null Hypothesis of Skill Invariance in the First Sense

(a) Language Skill (b) Cognitive Skill (c) Fine Motor Skill



[35] $\gamma_{1,\ell} = 1$ implies a uniform scale for latent skill variables between level ℓ and level $\ell - 1$. For example, the coefficient at level 8 for language skills (i.e., 0.562) presents the scale between level 7 and level 8.

Table 5: Skill Invariance Hypothesis Tests by Levels

	Language			Cognitive			Fine Motor		
	Slope($\gamma_{1,\ell}$)	$\chi^2(\cdot)$	p-value	Slope($\gamma_{1,\ell}$)	$\chi^2(\cdot)$	p-value	Slope($\gamma_{1,\ell}$)	$\chi^2(\cdot)$	p-value
Level 2				0.929	0.012	0.914	1.005	0.000	0.992
Level 3	0.901	0.546	0.460	0.936	0.010	0.922	0.963	0.022	0.883
Level 4	0.645	20.193	0.000	0.621	0.142	0.707	1.446	0.774	0.379
Level 5	0.66	9.382	0.002	2.235	3.899	0.048	0.798	0.720	0.396
Level 6	1.522	5.063	0.024	0.317	17.482	0.000	0.748	1.277	0.258
Level 7	1.125	0.182	0.670	0.791	0.362	0.547	0.955	0.034	0.853
Level 8	0.562	8.195	0.004	1.893	4.237	0.040			
Level 9	1.113	0.113	0.737	0.744	3.432	0.064			
Level 10	1.006	0.001	0.970	2.068	12.211	0.000			
Level 11	1.223	0.375	0.540	2.292	10.927	0.001			
Level 12				5.614	14.351	0.000			
Level 13				1.420	4.333	0.037			
Total		44.051	0.000		71.398	0.000		2.827	0.830

1. For each level we test the null hypothesis that $\gamma_{1,\ell}$=1 .
2. The column of p-value reports the probability of not rejecting the null hypothesis.
3. The row "Total" tests whether the scale invariance assumption is valid across all the levels.
4. Our data for language tasks starts from level 2.

Table 6: Difficulty Level List for Language (Learn words) Tasks

Level 8	The child points to the pictures which are being named, names one or more pictures, and mimics the sound of the objects.
Level 9	The child points to the pictures which are being named, names two or more pictures, makes the sound of the objects.
Level 10	The child points at 7 or more than 7 pictures and talks about them.
Level 11	Teach the child some simple descriptive words and the child names objects at home, and tells the usage of those objects.

We also test for skill invariance in the first sense for cognitive and fine motor skill tasks. Similarly, we reject the null of skill invariance across all the levels of the cognitive skill tasks. However, we find evidence in support of skill invariance for fine motor skill tasks, which mainly test drawing skills.

In sum, our estimates do not support skill invariance in the first sense across all levels for both language and cognitive skills, but the assumption cannot be rejected for some levels and some skills. For example, we cannot reject skill invariance between levels 8, 9, and 10 for tasks testing language skills. Skill invariance appears to be a valid description of fine motor skills at all levels. Our findings call into question standard practice that relies on skill invariant measures for analyzing skill growth and value-added.

Our evidence on skill invariance in the first sense is based on a parametric normal specification. This limits the generality of our findings. As previously noted, it is possible to estimate a nonparametric version of the model. That is a task left for the future.

8 Conclusion

This paper uses novel experimental data on a widely-emulated home visiting program implemented in rural China. We study its mechanisms for improving child skills, as documented in Zhou, Heckman, Wang, and Liu (2022). We investigate the impacts of different types of interactions on child achievement measures: interactions between home visitors and caregivers, interactions between home visitors and children, the quality of the teacher, and the frequency of the caregiver playing with the child after the class. High-quality interactions between the home visitor and the caregiver significantly improve child skill development in multiple dimensions, but the other features of the program are not generally effective. We find evidence consistent with dynamic complementarity using methods that do not rely on arbitrary measures of skills.

We develop and estimate a dynamic learning model to rationalize our evidence on program impacts. The model captures patterns of learning in our data and explains how skills evolve at weekly levels. We measure the growth in knowledge across difficulty levels. Our model explains the frequently noted phenomenon of "fadeout of measured skills within a lifetime" as a consequence of the stochastic nature of learning and the variation in performance across skill assessments. We introduce learning through investment and stochastic shocks into the standard IRT and BKT models of psychometrics.

High-ability children start strong and their knowledge generally does not improve within levels because they generally master tasks at the first attempt. Normal-ability children learn more but they have more to learn. Low-ability children also learn, but very slowly. Parental play accelerates their learning, but not that of children of other ability levels. Going forward, in designing the program, adaptive lessons that accelerate high-ability children will promote greater learning for high-ability children.

We formally test whether skill invariance in the sense of the existence of a constant units scale holds across skill levels. We find evidence supporting such skill invariance for certain skills at certain difficulty levels, but we reject the assumption as a global characterization, except for fine motor skills. This finding calls into question standard practice that assumes the existence of invariant measures for analyzing child development and the value added of teachers and schools. This evidence is in accord with the findings of Heckman and Zhou (2022a) showing the nonexistence of invariant measures of skills across levels.

There is clearly room for improvement in our research. Allowing for the cross-productivity of different skills in shaping the growth of skills, following Cunha et al. (2010) is an obvious and important extension left for the future. So is a semiparametric implementation of the model.

References

Agostinelli, F. and M. Wiswall (2022). Estimating the technology of children's skill formation. NBER Working Paper No. 22442, Accepted at *Journal of Political Economy*.

Almlund, M., A. L. Duckworth, J. J. Heckman, and T. Kautz (2011). Personality psychology and economics. In E. A. Hanushek, S. Machin, and L. Wößmann (Eds.), *Handbook of the Economics of Education*, Volume 4, Chapter 1, pp. 1–181. Amsterdam: Elsevier B. V.

Andrew, A., O. Attanasio, B. Augsburg, M. Day, S. Grantham-McGregor, C. Meghir, F. Mehrin, S. Pahwa, and M. Rubio-Codina (2020). Effects of a scalable home-visiting intervention on child development in slums of urban India: Evidence from a randomised controlled trial. *Journal of Child Psychology and Psychiatry 61*(6), 644–652.

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

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

Ben-Porath, Y. (1967, August). The production of human capital and the life cycle of earnings. *Journal of Political Economy 75*(4, Part 1), 352–365.

Borghans, L., B. H. H. Golsteyn, J. J. Heckman, and J. E. Humphries (2016). What do grades and achievement tests measure? Submitted to *Proceedings of the National Academy of Sciences*.

Bronfenbrenner, U. (Ed.) (2005). *Making Human Beings Human: Bioecological Perspectives on Human Development*. SAGE Publications.

Cawley, J., J. J. Heckman, and E. J. Vytlacil (1998). Cognitive ability and the rising return to education. NBER Working Paper 6388, National Bureau of Economic Research.

Corbett, A. T. and J. R. Anderson (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction 4*(4), 253–278.

Cunha, F. and J. J. Heckman (2007, May). The technology of skill formation. *American Economic Review 97*(2), 31–47.

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.

Cunha, F., E. Nielsen, and B. Williams (2021). The econometrics of early childhood human capital and investments. *Annual Review of Economics 13*(1), 487–513.

Del Boca, D., C. J. Flinn, and M. Wiswall (2014). Household choices and child development. *Review of Economic Studies 81*(1), 137–185.

Deonovic, B., M. Yudelson, M. Bolsinova, M. Attali, and G. Maris (2018). Learning meets assessment. *Behaviormetrika 45*(2), 457–474.

Doepke, M. and F. Zilibotti (2019). *Love, Money, and Parenting: How Economics Explains the Way we Raise our Kids.* Princeton University Press.

Duncan, G., A. Kalil, M. Mogstad, and M. Rege (2022). Investing in early childhood development in preschool and at home. In L. Woessmann, E. Hanushek, and S. J. Machin (Eds.), *Handbook of the Economics of Education*, Volume 6. North-Holland. Forthcoming.

Freyberger, J. (2021). Normalizations and misspecification in skill formation models.

García, J. L. and J. J. Heckman (2022). Parenting promotes social mobility within and across generations. Under Review, *Annual Reviews*.

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, April). Measuring knowledge and learning. Working Paper 29990, National Bureau of Economic Research.

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

Heckman, J. J. (1978, September). Simple statistical models for discrete panel data developed and applied to test the hypothesis of true state dependence against the hypothesis of spurious state dependence. *Annales de l'inséé 30/31*, 227–269.

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. and S. Mosso (2014, August). The economics of human development and social mobility. *Annual Review of Economics 6*(1), 689–733.

Heckman, J. J. and E. J. Vytlacil (2007). Econometric evaluation of social programs, part I: Causal models, structural models and econometric policy evaluation. In J. J. Heckman and E. E. Leamer (Eds.), *Handbook of Econometrics*, Volume 6B, Chapter 70, pp. 4779–4874. Amsterdam: Elsevier B. V.

Heckman, J. J. and J. Zhou (2022c). Comparing multiplicative and additive specifications for models of reinforcement learning. Under Preparation, University of Chicago.

Kautz, T., J. J. Heckman, R. Diris, B. ter Weel, and L. Borghans (2014). Fostering and measuring skills: Improving cognitive and non-cognitive skills to promote lifetime success. Technical report, Organisation for Economic Co-operation and Development, Paris. Available at https://www.oecd.org/edu/ceri/Fostering-and-Measuring-Skills-Improving-Cognitive-and-Non-Cognitive-Skills-to-Promote-Lifetime-Success.pdf.

Kim, J. H. (2019). Parenting skill and child development. Unpublished.

List, J. A. (2020). Non est disputandum de generalizability? A glimpse into the external validity trial. NBER Working Paper 27535.

Lord, F. M. and M. R. Novick (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley Publishing Company.

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

Matzkin, R. L. (1992, March). Nonparametric and distribution-free estimation of the binary threshold crossing and the binary choice models. *Econometrica 60*(2), 239–270.

Matzkin, R. L. (2007). Nonparametric identification. In J. J. Heckman and E. E. Leamer (Eds.), *Handbook of Econometrics*, Volume 6B. Amsterdam: Elsevier.

OECD (2021). *Beyond Academic Learning: First Results from the Survey of Social and Emotional Skills*. OECD Publishing.

Palmer, F. H. (1971). *Concept training curriculum for children ages two to five.* Stony Brook, NY: State University of New York at Stony Brook.

Romano, J. P. and M. Wolf (2005a, March). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association 100*(469), 94–108.

Romano, J. P. and M. Wolf (2005b). Stepwise multiple testing as formalized data snooping. *Econometrica 73*(4), 1237–1282.

Rutherford, R. S. G. (1955, July). Income distributions: A new model. *Econometrica 23*(3), 277–294.

Thelen, E. (2005, 04). Dynamic systems theory and the complexity of change. *Psychoanalytic Dialogues 15*, 255–283.

Todd, P. E. and K. I. Wolpin (2007, Winter). The production of cognitive achievement in children: Home, school, and racial test score gaps. *Journal of Human Capital 1*(1), 91–136.

Uzgiris, I. C. and J. M. Hunt (1975). *Assessment in Infancy: Ordinal Scales of Psychological Development*. Urbana, Illinois: University of Illinois Press.

van der Linden, W. J. (2016). *Handbook of Item Response Theory: Volume 1: Models*. Boca Raton, FL: CRC Press.

Vygotsky, L. S. (1978). *Mind in Society: The Development of Higher Psychological Processes*. Cambridge, MA: Harvard University Press.

Zhou, J., J. Heckman, B. Liu, and M. Lu (2022). The impacts of a prototypical home visiting program on child skills. Working Paper 27356, National Bureau of Economic Research.

Zhou, J., J. Heckman, F. Wang, and B. Liu (2022). Early childhood learning patterns for a home visiting program in rural China. *Journal of Community Psychology*.

Zhou, J., J. J. Heckman, B. Liu, M. Lu, S. M. Chang, and S. Grantham-McGregor (2022). Comparing China REACH and the Jamaica home visiting program. Revised and Resubmitted, *Pediatrics*.