I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 15252

Measuring Knowledge

James Heckman
Jin Zhou

APRIL 2022

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 15252

Measuring Knowledge

James Heckman
University of Chicago and IZA

Jin Zhou
University of Chicago

APRIL 2022

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Measuring Knowledge*

Empirical studies in the economics of education, the measurement of skill gaps across demographic groups, and the impacts of interventions on skill formation rely on psychometrically validated test scores that record the proportion of items correctly answered. Test scores are sometimes taken as measures of an invariant scale of human capital that can be compared over time and people. We show that for a prototypical test, invariance is violated. We use an unusually rich data set from an early childhood intervention program that measures knowledge of narrowly defined skills on essentially equivalent subsets of tasks. We examine if conventional, broadly-defined measures of skill are the same across people who are comparable on detailed knowledge measures. We reject the hypothesis of aggregate scale invariance and call into question the uncritical use of test scores in research on education and on skill formation. We compare different measures of skill and ability and reject the hypothesis of valid aggregate measures of skill.

JEL Classification: I210, C81, J71

Keywords: testing child development, psychometrics, measurement of discrimination, human capital, demographic economics

Corresponding author:
James J. Heckman
Center for the Economics of Human Development
and Department of Economics
University of Chicago
1126 East 59th Street
Chicago, IL 60637
USA

E-mail: jjh@uchicago.edu

* Research reported in this publication was supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development of the National Institutes of Health under award number R37HD065072, the Institute for New Economic Thinking, and a grant from an anonymous donor. We thank our partner China Development Research Foundation. This paper was inspired by suggestions of Flavio Cunha made at Rice University in November 2021. Becky Harles made helpful comments. Alejandra Campos and FuyaoWang contributed highly competent and insightful research assistance and commentary.

1 Introduction

A crucial assumption maintained in the literature on skill formation, ethnic skill gaps, and the economics of education is the existence of constant-unit latent skills ("human capital") over ages and inputs, which can be meaningfully compared across time and over people. A corollary but distinct assumption made in empirical work on measuring achievement growth and gaps and value-added measures is the existence of invariant measuring rods for latent skills, which may or may not exist even if there are true latent skill scales.[1] This paper tests for the existence of such invariant measures for prototypical achievement and assessment tests.

An assumption of invariance underlies measures of skill gaps across demographic groups (Cunha et al., 2021), value-added models in education (Konstantopoulos, 2014; Rivkin et al., 2005; Cawley et al., 1999; Hill, 2009; Rockoff, 2004), and studies of skill formation (Agostinelli et al., 2019; Cunha et al., 2010) is the existence of meaningful scales on which to measure the development of children and to compare performance across children at a point in time. Test scores are psychometric creations (see van der Linden, 2016). It has long been noted that any monotonic transformation of a test score is a valid test score and that cardinal comparisons of the type conventionally used to chart student progress over time or comparisons across children are fraught with peril (see, e.g., Cawley et al., 1999; Cunha and Heckman, 2008; Agostinelli et al., 2019; Freyberger, 2021; Cunha et al., 2021). This paper presents tests for the existence of such scales using a unique Chinese data set.

[1] For example, Todd and Wolpin (2007) and others use words spoken by age as measurements of constant-unit skills.

The central assumption in this paper is that mastery of tasks *within a well-defined level* is a true or foundational measure of knowledge. We can chart mastery within levels and compare knowledge and growth across children on a common microscale. Children can either perform a task successfully or not. We use this standard to assess the validity of more aggregative conventional measures of knowledge used in the economics of education and in the study of child development. Our study calls into question the conventional practice that relies on these aggregates as measures of knowledge that can be used to create meaningful comparisons across people or across time.

We use established and widely used measures of skill on narrowly defined tasks developed in the 1960s and 1970s by two influential and leading teams of child development psychologists. These measures of skill are based on the performance of a child. The *weekly* tasks we analyze are well defined, classified into developmental levels, and common across all children. Within narrowly defined levels, we assume, along with the literature, that tasks have the same knowledge content. A child's mastery of these tasks within a level is a precisely defined measure of knowledge. A measure of learning is mastery of progressively more difficult tasks. The question is whether the scales measure growth of the same skills. Mastery can be measured in multiple ways. We explore alternative plausible definitions of mastery and examine agreement among them.

This paper is organized as follows. Section 2 documents the curriculum design for China REACH program, which is our data source. Section 3 presents a model for measuring knowledge. We investigate the stability and comparability of alternative

skill measures over ages in Section 4. Section 5 presents our approach to testing the existence of a constant-unit measuring stick. Section 6 concludes.

2 Our Measures of Skill

The measurement and development of multiple skills in young children has been extensively studied. Uzgiris and Hunt (1975) and Palmer (1971) (UHP) define widely used measures of child development. We draw on these measures, which have been applied to an early childhood program in China.

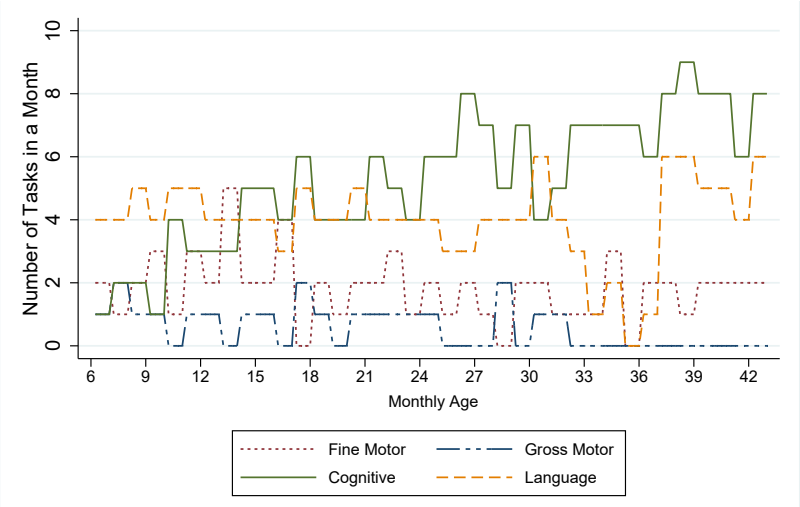
The analyzed China REACH program curriculum is adapted from the Jamaican Reach Up and Learn program, which was designed using UHP as a framework for understanding and supporting child growth and development. The tasks children confront in China REACH cover four domains of skill: fine motor, gross motor, language, and cognitive skills.

China REACH was implemented in 2015 by a large-scale randomized control trial. It enrolled 1,500 aged 9-30 months participants (about 700 participants in the treatment group) in 111 villages in Huachi county, Gansu province, one of the poorest areas of China (Zhou, Heckman, Liu, and Lu, 2021). Trained home visitors visit each treated household weekly and provide one hour of parenting or caregiving guidance. Multiple skills are fostered and tested. The curriculum teaches and encourages caregivers to talk to children through playing games, making toys, singing, reading, and storytelling to stimulate the child's cognitive, language, motor, and socioemotional skill development. We use measurements collected in this intervention.

The intervention follows the UHP script and records child success on it.

Three or four different skill tasks (gross motor, fine motor, language, and cognitive) are taught each week. These profiles describe hierarchies of knowledge. We assume that *knowledge content is the same within levels*. Children's skills are assessed weekly. Figure 1 presents the skill tasks taught and measured at each age. We next discuss the specific skills taught and how they are measured.

Figure 1: Curriculum Task Intensity: The Number of Tasks in a Month in the Curriculum (by Skill Category)



2.1 Cognitive Skills

For cognitive skills, there are thirteen difficulty levels (see Table 1). Figure 2 gives the timing of the measures by age.

Cognitive skills have different dimensions. In the curriculum, the cognitive skills

taught cover spatial skills, knowledge of objects and object functions, order and number, etc. We use knowledge of objects and object functions as an example. Cognitive skill difficulty levels are defined based on the abstract concepts shown in Table 1, such as the child's proficiency in understanding the objects. Seventy-four lessons are sorted into the thirteen ordered difficulty levels.[2] The lessons cover the process of how the child learns to know an object and understand its function.

The cognitive knowledge of objects tasks progresses from a simple understanding of concepts depicted in pictures by acknowledging with vocalizations to using receptive (heard) language to identify certain pictures. Receptive language is a skill developed prior to expressive language whereby children form words to communicate. Children must use expressive language to complete the subsequent lessons, which increase with difficulty as the children must develop more and more language to identify an increasing number of images. To progress, the child must display an increasingly sophisticated understanding of the stories presented, first simply naming actions, then answering questions and talking abstractly about a story. Levels 10, 11, 12, and 13 ask the child to take the information presented and build on it by discussing the uses of the objects depicted and making connections with other images.

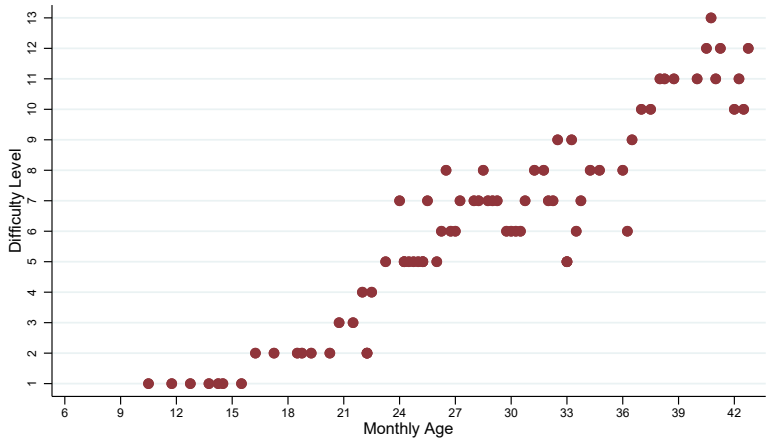
Figure 2 shows the timing of the cognitive skill (knowing objects and understanding object functions) levels in the curriculum. The number of lessons varies across difficulty levels according to the curriculum content itself. Table 2 presents detailed information about the six lessons (and assessments) that are labeled as difficulty level

[2]The difficulty level has ordinal meaning only, not cardinal meaning.

Table 1: Difficulty Level List for the Cognitive Understanding Objects Lessons

Level 1	The child looks at the pictures and vocalizes.
Level 2	Name the objects and ask the child to point to the corresponding pictures.
Level 3	The child can point to one picture and name the objects in it.
Level 4	The child can point to two or more pictures and name the objects in them.
Level 5	The child can point to three or more pictures and name the objects in them.
Level 6	The child can point to six or more pictures and name the objects in them.
Level 7	The child can talk about the pictures, answer questions, and understand or name actions (eat, play, etc.).
Level 8	The child can follow the storyline, answer questions, and name actions.
Level 9	The child can understand stories and talk about the content of 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 graphic, discuss the role of each item, and then link the graphics in the card together.
Level 12	The child can name the objects in the picture, link different pictures together, and discuss some of the activities in the pictures.
Level 13	The child can name the objects in the picture and talk about their functions.

Figure 2: The Timing of Cognitive Skill (Understanding 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.

1 directed to ten-month-old to fifteen-month-old children. In Table 2, all lessons relate to the activity of looking at the pictures or objects and vocalizing, which does not require the child to name or identify the object.

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

Difficulty Level	Month	Week	Learning Materials	Content
1	10	2	Picture book A	The baby makes sounds when looking at the pictures.
1	11	3	Picture book B	The baby looks at the pictures and vocalizes.
1	12	3	Picture book A	The child makes sounds looking at the pictures.
1	13	3	Picture book B	The child makes sounds looking at the pictures.
1	14	1	Picture book A	Mother and child look at the pictures together, and the mother lets the child vocalize and touch the pictures.
1	15	2	Picture book B	Mother and child look at the pictures together, and the mother lets the child vocalize and touch the pictures.

Measures evolve as depicted in Figure 2. As children age and advance across

difficulty levels, they confront more demanding tasks.[3]

2.2 Fine Motor Skills

As another example, consider fine motor drawing lessons, for which there are seven difficulty levels.[4] In general, higher difficulty levels for skills include new content. For example, difficulty level 2 asks the child to mimic circles. The skills at difficulty level 3 include drawing straight lines. We document how the tasks in different difficulty levels are categorized.

Fine motor drawing lessons focus on a child's ability to use writing utensils on progressively more difficult tasks. First, a child is asked to hold utensils to make markings. The child is then asked to copy the markings made by an adult. As the skill levels progress, the child is asked to make markings after only hearing a verbal command from an adult. Finally, the child progresses from abstract shapes to representative drawing (See Table 3.)

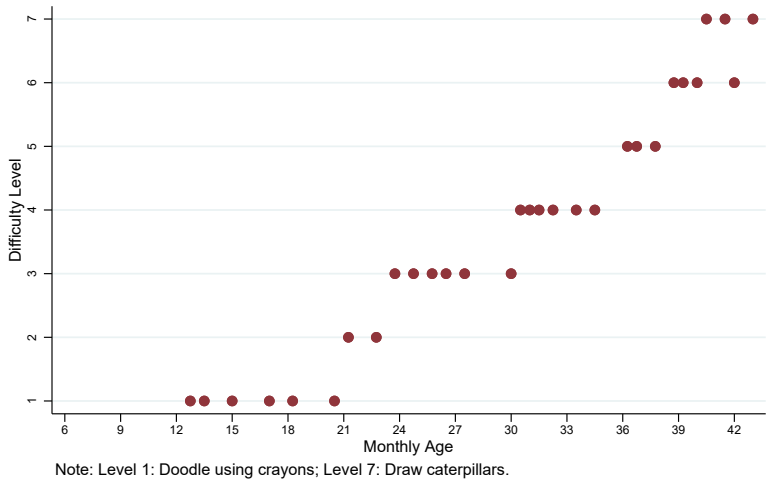
Table 3: Skill Levels for Fine Motor (Drawing) Lessons

Difficulty Level	**Task Content**
Level 1	Doodle using crayons
Level 2	Mimic circles
Level 3	Mimic circles and draw straight lines
Level 4	Draw a circle, vertical line, and horizontal line
Level 5	Draw circles, many lines, and crossed lines
Level 6	Draw a cross (or T), curves, and zigzag curves
Level 7	Draw caterpillars

[3]Occasionally, the protocol reverts to earlier levels of skill to review the child's learning and bolster confidence in their acquired skills.

[4]The standard generating the difficulty levels is based on an understanding of the content in the skills.

Figure 3: The Timing of Fine Motor Skill (Drawing) Tasks across Difficulty Levels



Note: Level 1: Doodle using crayons; Level 7: Draw caterpillars.

In addition to tasks of different difficulty levels, the curriculum features multiple lessons and assessments *within* the same difficulty level. The number of lessons within each difficulty level depends on the curriculum. For example, there are six assessments at difficulty level 3 for fine motor drawing skills but only two assessments at difficulty level 2.

Figure 3 gives the timing of each fine motor drawing assessment in the curriculum design. Difficulty level 1 covers the ages from 12 months and 3 weeks to 20 months and 2 weeks. This means that when children are 12 months and 3 weeks old, the home visitor will teach them the first fine motor drawing skill. When they are 20 months and 2 weeks old, the home visitor will teach them the sixth lesson at difficulty level 1. In general, higher difficulty levels appear at later weekly ages. However, there can be some overlap across difficulty levels. When fine motor lessons at difficulty level 7 start, the student still receives lessons at difficulty level 6. Circling back is a strategy

designed to solidify a child's understanding of a concept. Appendix A discusses all the skills we measure.

2.3 Our Key Identifying Assumption

The curriculum we study targets lessons at different skill levels at each weekly age. For each type of skill, task difficulty levels are constructed following UHP. We use mastery of tasks at each level of skill as our fundamental measure of knowledge. Knowledge is acquired in real time. It may be forgotten or retained as children advance through the curriculum, leading to multiple measures of knowledge. Different types of knowledge can be acquired at different levels.

3 A Model for Measuring Knowledge

Our data on weekly skill growth enable us to move beyond the traditional aggregates such as percentage of items passed (as reported in the Denver, Bayley, and most other achievement tests) to examine age-by-age skill growth and the factors that influence it. To understand the structure of our data and alternative ways to measure knowledge and learning, it is helpful to introduce some notation.

Let $\mathcal{S}$ be the set of skills taught. Let $\ell(s, a)$ be the level of skill s taught at age a; $\ell(s, a) \in \{1, \ldots, L_s\}$. L_s is the number of difficulty levels for each skill s. Mastery

of skill s at level ℓ at age a is characterized by a threshold crossing model:

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

where $D(s,\ell,a)$ records mastery (or not) of a skill s at a given level ℓ at age a. $\bar{K}(s,\ell)$ is the minimum latent skill required to master the task at difficulty level ℓ. This characterization is consistent with the classical IRT model in educational psychology (Lord and Novick, 1968; van der Linden, 2016).

Let $\underline{a}(s,\ell)$ be the first age at which skill s is measured at level ℓ, and let $\bar{a}(s,\ell)$ be the last age at which it is measured at level ℓ. For consecutive lessons in a run, $1 + \bar{a}(\ell) - \underline{a}(\ell)$ is the length of the run (# of lessons measured on skill s at level ℓ) starting at age $\underline{a}(s,\ell)$.

For level ℓ of skill s, collect the indicators of knowledge in a spell:

$$\Big\{D(s,\ell,a)\Big\}_{\underline{a}(s,\ell)}^{\bar{a}(s,\ell)}.$$

In a stationary environment with age-invariant heterogeneity with no learning or growth of knowledge at level ℓ, 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).[5]

With learning, sequences are back-loaded. For $j > 0$,

$$\Pr(D(s,\ell,a+j) \geq D(s,\ell,a)) \geq 0.$$

Knowledge acquisition for each skill s at each level ℓ is measured by properties of

[5]See Heckman (1978, 1981).

these arrays and their relationships. Zhou, Heckman, Wang, and Liu (2021) test and reject the hypothesis of no learning for our data. They control for maturation and exposure effects that might boost skills in the absence of any intervention. Even after doing so, they reject exchangeability and find evidence of knowledge growth throughout the program.

Figure 4 characterizes the growth of knowledge of language, cognitive, and fine motor skills. Average passing rates within each difficulty level for language and cognitive tasks increase with age, a pattern consistent with learning. When individuals transition to a higher difficulty level, initial passing rates decline. Subsequent passing rates increase as learning ensues.

3.1 Measures of Knowledge and Knowledge Acquisition

The traditional measure of knowledge of a skill is the proportion of correct answers over all levels of difficulty. A more refined measure within an assessment is defined within a skill and difficulty level (s, ℓ). The passing rate on skill s at level ℓ is:

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

The overall passing rate is:

$$p(s) = \frac{\sum_{\ell=1}^{L_s} \Big\{1 + \bar{a}(s,\ell) - \underline{a}(s,\ell)\Big\} p(s,\ell)}{\sum_{\ell=1}^{L_s} \Big\{1 + \bar{a}(s,\ell) - \underline{a}(s,\ell)\Big\}}, \tag{2}$$

which weights all items across all difficulty levels equally and puts more weight on

difficulty levels with more items. This measure is an aggregate measure that does not recognize the sampling of (s, ℓ) items, the retention of knowledge, or the speed of acquisition.

We define other plausible measures of knowledge and knowledge acquisition, which we also measure. For consecutive learning spells with all participants entering each level at the first lesson, we define ***time to first mastery*** as $d(s,\ell) = \hat{a}(s,\ell) - \underline{a}(s,\ell)$, where for each s and ℓ, $\hat{a}(s,\ell) = \min_a\{D(s,\ell,a) = 1\}_{a=\underline{a}(s,\ell)}^{\bar{a}(s,\ell)}$. We define ***time to full mastery*** as $\tilde{a}(s,\ell) = \min_a[D(s,\ell,a) = 1, \forall a \geq \tilde{a}(s,\ell)]$.[6] Time to full mastery is $\tilde{a}(s,\ell) - \underline{a}(s,\ell)$. Some would call speed of mastery an ability and not a pure measure of knowledge. Other measures of learning are possible, such as time to mastery of two items in a row after $\hat{a}(s,\ell)$, etc. ***Backsliding*** at level ℓ for skill s is:

$$\frac{\#\{D(s,\ell,a) = 0, a > \hat{a}(s,\ell), a \leq \bar{a}(s,\ell)\}}{\#\{a > \hat{a}(s,\ell), a \leq \bar{a}(s,\ell)\}} \mathbf{1}(\#\{a > \hat{a}(s,\ell), a \leq \bar{a}(s,\ell)\} > 0).$$

3.2 Correlations with Conventional Test Scores

It is instructive to examine the correlation between the measures just defined and traditional achievement scores. We use Denver tests as traditional scores (Appelbaum, 1978). They are very closely related to Bayley scores used to measure child development (Rubio-Codina and Grantham-McGregor, 2020; Rubio-Codina et al., 2016). Tables 4a–4d present the correlations between the Denver scores at midline

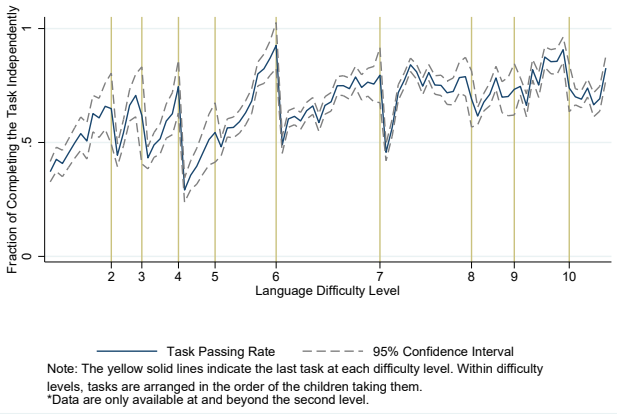
[6] We define time to first mastery using the number of tasks a child takes until the first success (inclusive) at each difficulty level by skill type. Similarly, time to full mastery is the number of tasks a child takes to succeed and not fail afterwards at each difficulty level during the intervention by skill type.

and endline for combined language-cognitive, fine motor, gross motor, and socioemotional skills, as well as average passing rates, the common measure of "knowledge," cumulated up to the date at which the Denver test is administered. The Denver tests were administered twice during the intervention: the midline was administered about nine months into the intervention, and the endline was administered about twenty-one months into the intervention. We consider other measures of knowledge below.

Most of the measures are significantly correlated with the children's Denver test scores in the expected directions. The Denver score is positively correlated with the average passing rate across tasks during the intervention. Notice, however, the strong correlations between Denver tasks tailored to a particular skill and the components of knowledge from all skills. This might suggest a one-dimensional model of skill. However, we test and reject that model. Heckman and Zhou (2021) summarize estimates of the dimensionality of these measures. There are two dimensions for each measure and at least five dimensions across all measures of knowledge. Knowledge is not one-dimensional, and the existence of Galton's "g" as a solo measure of ability is called into question.

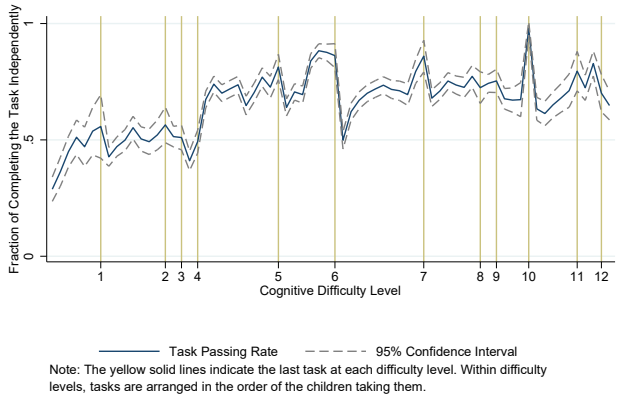
Figure 4: Average Task Passing Rate by Order and Level

(a) Language*



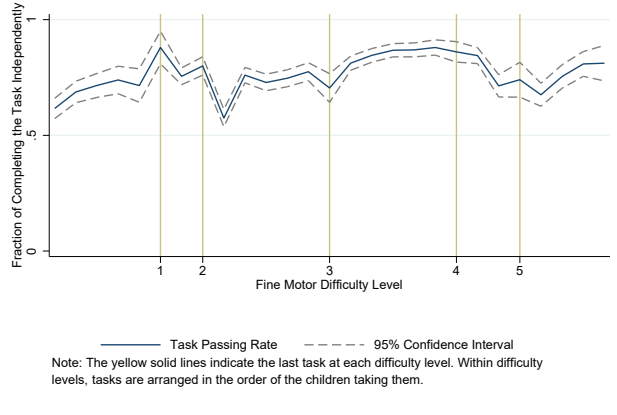
Note: The yellow solid lines indicate the last task at each difficulty level. Within difficulty levels, tasks are arranged in the order of the children taking them.
*Data are only available at and beyond the second level.

(b) Cognitive



Note: The yellow solid lines indicate the last task at each difficulty level. Within difficulty levels, tasks are arranged in the order of the children taking them.

(c) Fine Motor



Note: The yellow solid lines indicate the last task at each difficulty level. Within difficulty levels, tasks are arranged in the order of the children taking them.

Table 4a: Correlation between Average Passing Rate (Up to Midline/Endline Measurement Age) and Denver Scores

		Average Passing Rate			
		Language	Cognitive	Fine Motor	Gross Motor
Denver Score (Midline)	Language and Cognitive	0.039**	0.078***	0.061**	0.043**
	Fine Motor	0.040**	0.076***	0.057**	0.086***
	Gross Motor	0.027	0.080***	0.054*	0.011
	Socioemotional	0.100***	0.118***	0.068**	0.068***
Denver Score (Endline)	Language and Cognitive	0.078***	0.098***	0.099***	0.058***
	Fine Motor	0.011	0.042***	0.042**	0.017
	Gross Motor	0.075***	0.088***	0.064***	0.055***
	Socioemotional	0.005	0.024*	0.044**	-0.001

1. Average passing rate is the passing rate for the intervention tasks at each difficulty level by each skill type.
2. For the Denver score (midline) rows, the measures of average passing rate are evaluated from the time of enrollment up to Denver midline measurement age and similarly for the Denver score (endline) rows.
3. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4b: Correlation between Time to First Mastery (Up to Midline/Endline Measurement Age) and Denver Scores

		Time to First Mastery			
		Language	Cognitive	Fine Motor	Gross Motor
Denver Score (Midline)	Language and Cognitive	-0.044**	-0.064***	-0.081***	-0.048**
	Fine Motor	-0.044**	-0.043**	-0.054*	-0.049**
	Gross Motor	-0.030	-0.078***	-0.034	-0.008
	Socioemotional	-0.071***	-0.073***	-0.060**	0.000
Denver Score (Endline)	Language and Cognitive	-0.076***	-0.069***	-0.052**	0.019
	Fine Motor	-0.024	-0.027*	-0.017	-0.002
	Gross Motor	-0.071***	-0.071***	-0.012	-0.027
	Socioemotional	-0.020	-0.023	0.029	0.003

1. 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 each skill type.
2. For the Denver score (midline) rows, the measures of time to mastery are evaluated from the time of enrollment up to Denver midline measurement age and similarly for the Denver score (endline) rows.
3. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4c: Correlation between Instability (Up to Midline/Endline Measurement Age) and Denver Scores

		Instability			
		Language	Cognitive	Fine Motor	Gross Motor
Denver Score (Midline)	Language and Cognitive	-0.049**	-0.110***	-0.101***	-0.063**
	Fine Motor	-0.032	-0.058**	-0.058*	-0.103***
	Gross Motor	-0.023	-0.033	-0.101***	-0.032
	Socioemotional	-0.022	-0.094***	-0.050	-0.038
Denver Score (Endline)	Language and Cognitive	-0.070***	-0.063***	-0.043*	-0.078***
	Fine Motor	-0.026	-0.040**	-0.021	-0.031
	Gross Motor	-0.061***	-0.074***	-0.048**	-0.061**
	Socioemotional	0.003	-0.019	-0.041*	-0.032

1. Instability is defined as the proportion of fails after the first success at each difficulty level by each skill type.
2. For the Denver score (midline) rows, the measures of instability are evaluated from the time of enrollment up to Denver midline measurement age and similarly for the Denver score (endline) rows.
3. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4d: Correlation between Time to Full Mastery (Up to Midline/Endline Measurement Age) and Denver Scores

		Time to Full Mastery			
		Language	Cognitive	Fine Motor	Gross Motor
Denver Score (Midline)	Language and Cognitive	-0.062***	-0.076***	-0.126***	-0.015
	Fine Motor	-0.040**	-0.034	-0.033	-0.035
	Gross Motor	-0.010	-0.025	-0.085**	0.031
	Socioemotional	-0.022	-0.029	-0.028	0.008
Denver Score (Endline)	Language and Cognitive	-0.049***	-0.046**	-0.082***	-0.078**
	Fine Motor	-0.022	-0.036**	-0.070**	-0.050
	Gross Motor	-0.030	-0.024	-0.020	-0.066**
	Socioemotional	-0.028	-0.001	-0.027	-0.044

1. Time to full mastery is defined as the number of tasks a child takes to succeed and not fail afterwards at each difficulty level during the intervention by each skill type.
2. For the Denver score (midline) rows, the measures of time to full mastery are evaluated from the time of enrollment up to Denver midline measurement age and similarly for the Denver score (endline) rows.
3. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Denver scores are negatively correlated with the time the child takes to achieve first success and negatively correlated with the proportion of fails after the first suc-

cess. Compared to fine and gross motor scores, the language and cognitive scores have more statistically significant correlations with these measures. The program significantly improves measured language and cognitive skills. The correlations between the Denver scores (endline and midline) and our other measures of knowledge are generally much weaker, suggesting that Denver scores do not capture other dimensions of knowledge as well as conventional passing measures.

3.2.1 Correlations with Measures at the Time the Denver Test Is Taken

In addition to correlating knowledge measured over intervals, it is useful to measure knowledge at the time the Denver tests are taken. Tables 5a–5d report such correlations.

Table 5a: Correlation between Average Passing Rate (At Midline/Endline Measurement Age) and Denver Scores

		Average Passing Rate			
		Language	Cognitive	Fine Motor	Gross Motor
Denver Score (Midline)	Language and Cognitive	0.101**	0.074*	0.100	0.050
	Fine Motor	0.149***	0.069	0.170***	0.097*
	Gross Motor	0.147***	0.062	0.142**	0.012
	Socioemotional	0.128***	0.043	0.066	0.012
Denver Score (Endline)	Language and Cognitive	0.004	0.127*	0.058	-0.076
	Fine Motor	-0.249**	-0.066	-0.086	0.308
	Gross Motor	-0.085	0.198***	0.057	0.118
	Socioemotional	-0.216*	0.129**	0.115	0.078

1. Average passing rate is the passing rate for the intervention tasks at each difficulty level by each skill type.
2. For the Denver score (midline) rows, the measures of average passing rate are the difficulty levels evaluated at Denver midline measurement age, and similarly for the Denver score (endline) rows.
3. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 5b: Correlation between Time to First Mastery (At Midline/Endline Measurement Age) and Denver Scores

		Time to First Mastery			
		Language	Cognitive	Fine Motor	Gross Motor
Denver Score (Midline)	Language and Cognitive	-0.056	0.072	-0.045	-0.046
	Fine Motor	-0.052	0.012	0.006	0.018
	Gross Motor	-0.085*	0.013	-0.069	0.045
	Socioemotional	-0.039	-0.032	0.017	-0.013
Denver Score (Endline)	Language and Cognitive	0.091	-0.114	-0.004	0.076
	Fine Motor	-0.026	-0.010	0.038	-0.308
	Gross Motor	-0.049	-0.207***	0.047	-0.118
	Socioemotional	0.187	-0.250***	0.034	-0.078

1. 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 each skill type.

2. For the Denver score (midline) rows, the measures of time to mastery are the difficulty levels evaluated at Denver midline measurement age, and similarly for the Denver score (endline) rows.

3. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 5c: Correlation between Instability (At Midline/Endline Measurement Age) and Denver Scores

		Instability			
		Language	Cognitive	Fine Motor	Gross Motor
Denver Score (Midline)	Language and Cognitive	-0.148***	-0.074	-0.044	0.044
	Fine Motor	-0.049	-0.056	-0.091	-0.025
	Gross Motor	-0.004	-0.004	-0.019	0.048
	Socioemotional	-0.061	-0.026	0.012	0.129*
Denver Score (Endline)	Language and Cognitive	-0.294*	-0.025	0.064	.
	Fine Motor	0.069	0.086	0.026	.
	Gross Motor	-0.078	-0.183*	0.029	.
	Socioemotional	-0.038	-0.128	-0.086	.

1. Instability is defined as the proportion of fails after the first success at each difficulty level by each skill type.

2. For the Denver score (midline) rows, the measures of instability are the difficulty levels evaluated at Denver midline measurement age and similarly for the Denver score (endline) rows.

3. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 5d: Correlation between Time to Full Mastery (At Midline/Endline Measurement Age) and Denver Endline Scores

		Time to Full Mastery			
		Language	Cognitive	Fine Motor	Gross Motor
Denver Score (Midline)	Language and Cognitive	-0.072	0.093*	0.001	0.150
	Fine Motor	0.045	-0.037	-0.051	0.062
	Gross Motor	0.012	0.015	-0.064	0.095
	Socioemotional	0.010	-0.029	0.013	0.006
Denver Score (Endline)	Language and Cognitive	0.118	0.027	-0.271**	.
	Fine Motor	-0.038	-0.008	-0.040	.
	Gross Motor	0.217	-0.027	-0.069	.
	Socioemotional	-0.174	-0.146	-0.167	.

1. Time to full mastery is defined as the number of tasks a child takes to succeed and not fail afterwards at each difficulty level during the intervention by each skill type.
2. For the Denver score (midline) rows, the measures of time to full mastery are the difficulty levels evaluated at Denver midline measurement age and similarly for the Denver score (endline) rows.
3. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

The contemporaneous measures of knowledge are much more weakly correlated with the Denver scores. Cumulative measures are more predictive.

While all the correlations are in the expected direction, the different measures are far from perfectly correlated, suggesting that they capture different aspects of knowledge.[7] Table 6 shows the correlations between different measures of knowledge. Time to first mastery is strongly negatively correlated with passing rates but much more weakly correlated with knowledge retention. Instability (backsliding) is at best weakly correlated with speed (time to mastery). The different measures of knowledge capture aspects of learning.

[7]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 error in such measures.

Table 6: Correlations between Different Measures of Knowledge

Correlation Variables	Language	Cognitive	Fine Motor	Gross Motor
Time to First Mastery vs. Avg. Passing Rate	-0.641***	-0.677***	-0.688***	-0.607***
Time to First Mastery vs. Instability	0.181***	0.208***	0.175***	-0.035
Avg. Passing Rate vs. Instability	-0.810***	-0.831***	-0.857***	-0.932***
Time to Full Mastery vs. Avg. Passing Rate	0.137***	0.193***	0.022	0.181***
Time to Full Mastery vs. Instability	0.170***	0.209***	0.253***	0.589***
Time to Full Mastery vs. Time to First Mastery	0.237***	0.155***	0.049*	-0.518***

Notes: 1. Average passing rate is the passing rate for the intervention tasks at each difficulty level by each skill type. 2. For intervention tasks, instability is defined as the proportion of fails after the first success at each difficulty level by each skill type. 3. Time to first mastery is defined as the number of tasks a child takes until the first success (inclusive) at each difficulty level. 4. Time to full mastery is defined as the number of tasks a child takes to succeed and not fail afterwards at each difficulty level during the intervention by each skill type.
5. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

4 Stability of Mastery of Skills over Time

Using our data and measures, we can define ability groups and determine the stability of membership in the ability categories. Ability categories are defined by the speed of mastering the task (time to the first correct answer). It is conventional to measure ability by the speed of learning, while learning is defined by eventual mastery of tasks. We examine how distinct these measures actually are.

Table 7 defines the categories. There is strong persistence of passing rates across difficulty levels. The fast group is defined by two conditions: (1) members pass the first task for more than 80% of difficulty levels, and (2) members pass all skill-specific tasks at an average rate of more than 80%. Figures 5a–5d show that passing rates are persistent. Figures 6a–6d and 7a–7d show similar persistence for other measures of knowledge. The full mastery measure is quite noisy. Ability predicts the proportion

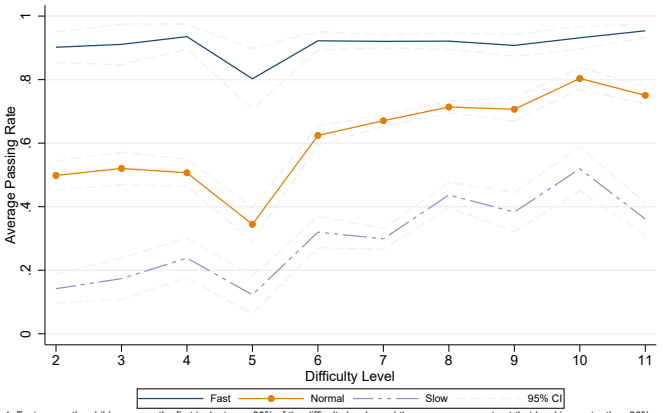
of times that children get the wrong answer after a first correct answer (a measure of instability in performance) for cognition, language, and the other skills. See Figure 7. We next use these micro-based measures of knowledge to evaluate the invariance of aggregated test scores, as previously defined.

Table 7: Ability Categories (Measured over All Levels)

Fast group	Pass the first task for more than 80% of difficulty levels, and pass all skill-specific tasks at an average rate of more than 80%.
Normal group	Pass the first task for less than 80% of difficulty levels, and the pass rate is greater than 50%; or pass the first task for more than 80% of difficulty levels, and the average passing rate of all skill-specific tasks is between 50% and 80%.
Slow group	The average passing rate of all skill-specific tasks is less than 50%.

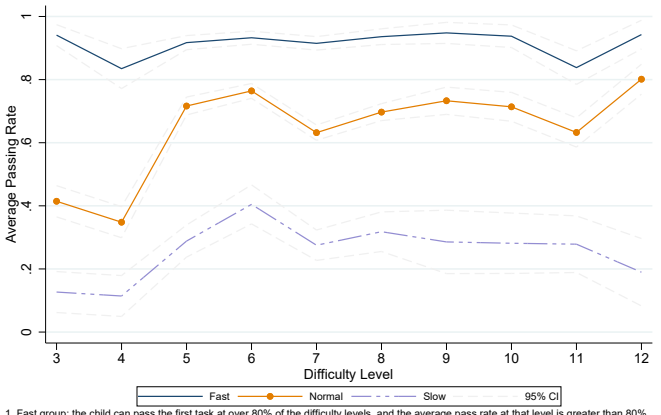
Figure 5: Average Passing Rate by Ability Category and Level

(a) Language Tasks



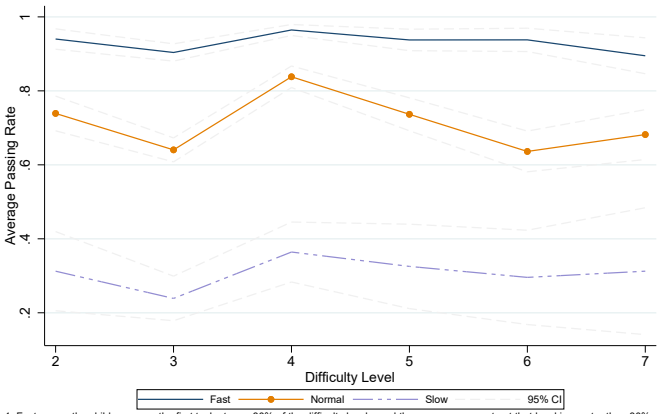
1. Fast group: the child can pass the first task at over 80% of the difficulty levels, and the average pass rate at that level is greater than 80%. Normal group: the child doesn't pass the first task, and the pass rate is greater than 50%; or the child passes the first task, and the pass rate is between 50% and 80%. Slow group: the average pass rate is less than 50%. 2. 95% confidence intervals are shown for three groups.

(b) Cognitive Tasks



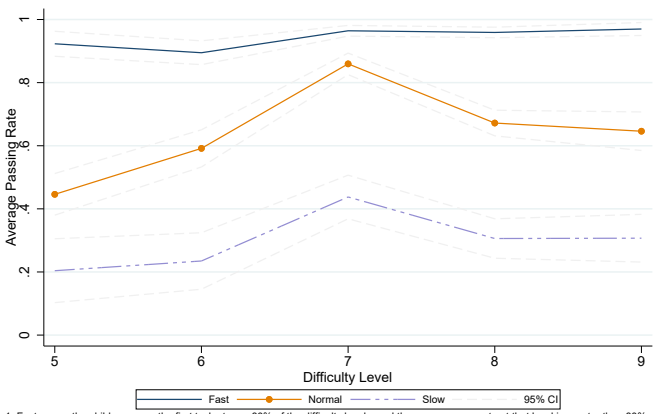
1. Fast group: the child can pass the first task at over 80% of the difficulty levels, and the average pass rate at that level is greater than 80%. Normal group: the child doesn't pass the first task, and the pass rate is greater than 50%; or the child passes the first task, and the pass rate is between 50% and 80%. Slow group: the average pass rate is less than 50%. 2. 95% confidence intervals are shown for three groups.

(c) Fine Motor Tasks



1. Fast group: the child can pass the first task at over 80% of the difficulty levels, and the average pass rate at that level is greater than 80%. Normal group: the child doesn't pass the first task, and the pass rate is greater than 50%; or the child passes the first task, and the pass rate is between 50% and 80%. Slow group: the average pass rate is less than 50%. 2. 95% confidence intervals are shown for three groups.

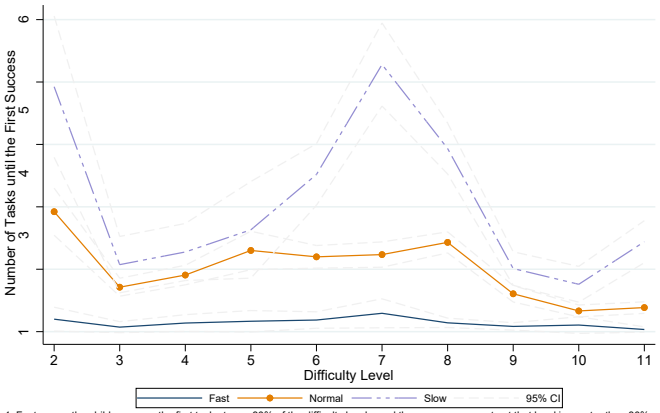
(d) Gross Motor Tasks



1. Fast group: the child can pass the first task at over 80% of the difficulty levels, and the average pass rate at that level is greater than 80%. Normal group: the child doesn't pass the first task, and the pass rate is greater than 50%; or the child passes the first task, and the pass rate is between 50% and 80%. Slow group: the average pass rate is less than 50%. 2. 95% confidence intervals are shown for three groups.

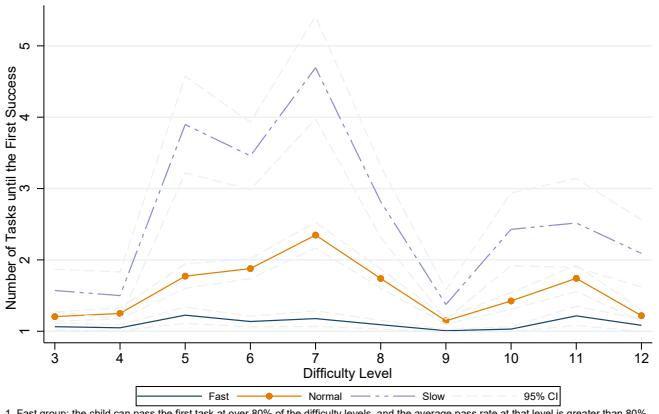
Figure 6: Time to First Mastery Measures by Ability Category and Level

(a) Language Tasks



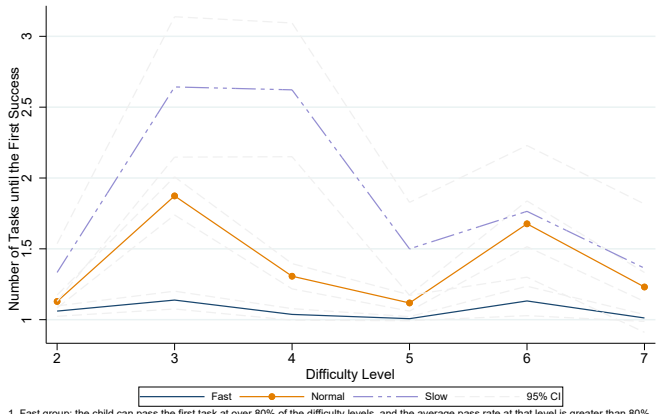
1. Fast group: the child can pass the first task at over 80% of the difficulty levels, and the average pass rate at that level is greater than 80%. Normal group: the child doesn't pass the first task, and the pass rate is greater than 50%; or the child passes the first task, and the pass rate is between 50% and 80%. Slow group: the average pass rate is less than 50%. 2. 95% confidence intervals are shown for three groups.

(b) Cognitive Tasks



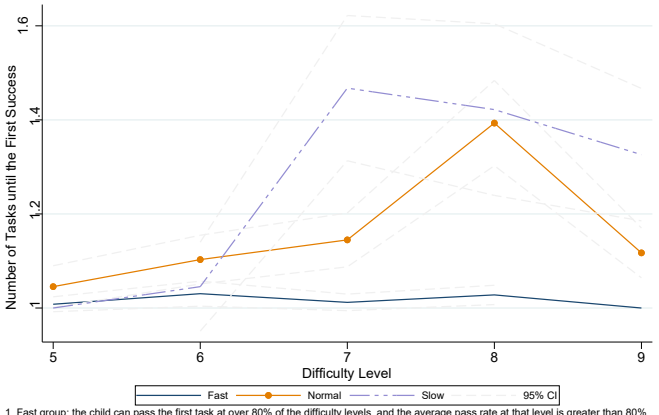
1. Fast group: the child can pass the first task at over 80% of the difficulty levels, and the average pass rate at that level is greater than 80%. Normal group: the child doesn't pass the first task, and the pass rate is greater than 50%; or the child passes the first task, and the pass rate is between 50% and 80%. Slow group: the average pass rate is less than 50%. 2. 95% confidence intervals are shown for three groups.

(c) Fine Motor Tasks



1. Fast group: the child can pass the first task at over 80% of the difficulty levels, and the average pass rate at that level is greater than 80%. Normal group: the child doesn't pass the first task, and the pass rate is greater than 50%; or the child passes the first task, and the pass rate is between 50% and 80%. Slow group: the average pass rate is less than 50%. 2. 95% confidence intervals are shown for three groups.

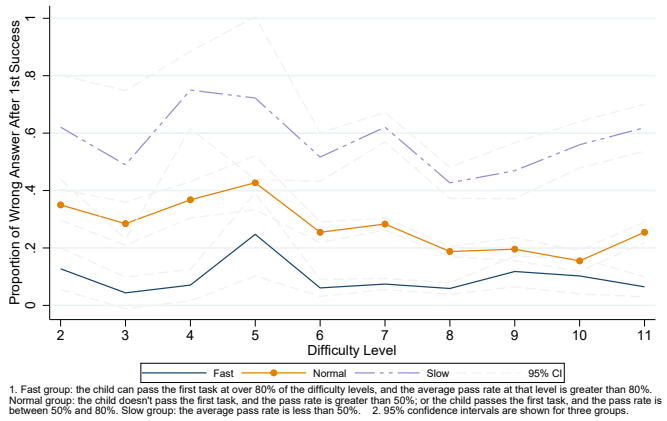
(d) Gross Motor Tasks



1. Fast group: the child can pass the first task at over 80% of the difficulty levels, and the average pass rate at that level is greater than 80%. Normal group: the child doesn't pass the first task, and the pass rate is greater than 50%; or the child passes the first task, and the pass rate is between 50% and 80%. Slow group: the average pass rate is less than 50%. 2. 95% confidence intervals are shown for three groups.

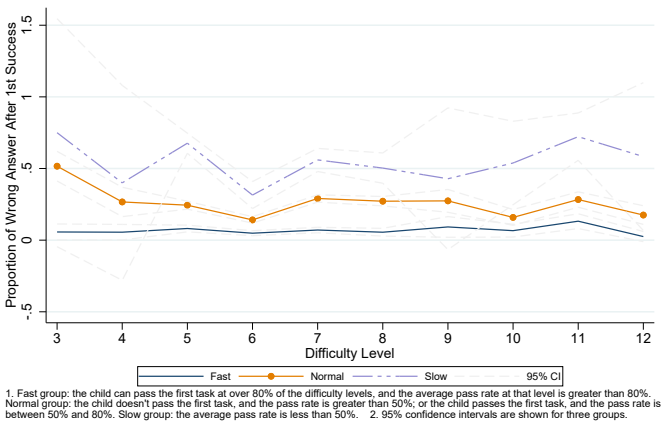
Figure 7: Instability (Proportion of Wrong Answers after First Success) Measures by Ability Category and Level

(a) Language Tasks



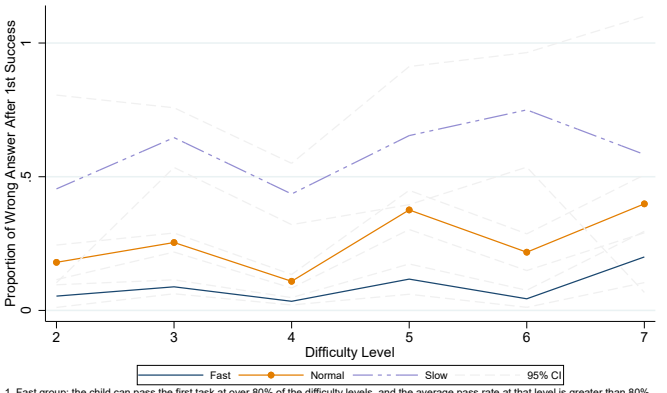
1. Fast group: the child can pass the first task at over 80% of the difficulty levels, and the average pass rate at that level is greater than 80%. Normal group: the child doesn't pass the first task, and the pass rate is greater than 50%; or the child passes the first task, and the pass rate is between 50% and 80%. Slow group: the average pass rate is less than 50%. 2. 95% confidence intervals are shown for three groups.

(b) Cognitive Tasks



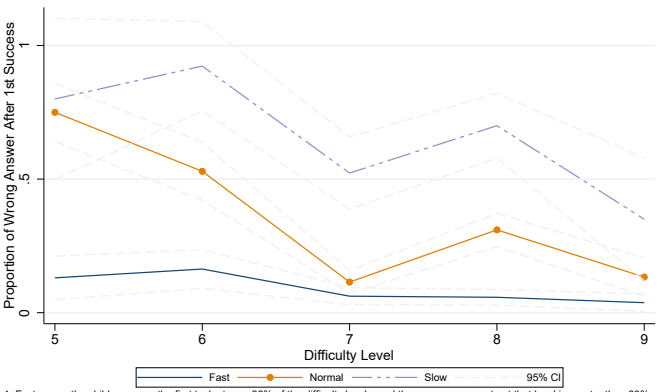
1. Fast group: the child can pass the first task at over 80% of the difficulty levels, and the average pass rate at that level is greater than 80%. Normal group: the child doesn't pass the first task, and the pass rate is greater than 50%; or the child passes the first task, and the pass rate is between 50% and 80%. Slow group: the average pass rate is less than 50%. 2. 95% confidence intervals are shown for three groups.

(c) Fine Motor Tasks



1. Fast group: the child can pass the first task at over 80% of the difficulty levels, and the average pass rate at that level is greater than 80%. Normal group: the child doesn't pass the first task, and the pass rate is greater than 50%; or the child passes the first task, and the pass rate is between 50% and 80%. Slow group: the average pass rate is less than 50%. 2. 95% confidence intervals are shown for three groups.

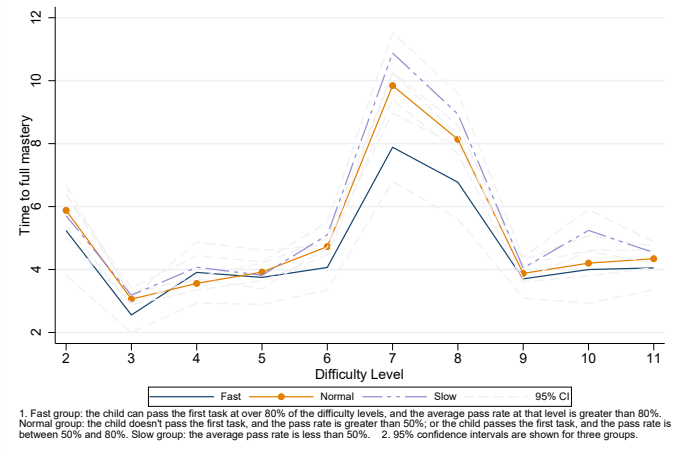
(d) Gross Motor Tasks



1. Fast group: the child can pass the first task at over 80% of the difficulty levels, and the average pass rate at that level is greater than 80%. Normal group: the child doesn't pass the first task, and the pass rate is greater than 50%; or the child passes the first task, and the pass rate is between 50% and 80%. Slow group: the average pass rate is less than 50%. 2. 95% confidence intervals are shown for three groups.

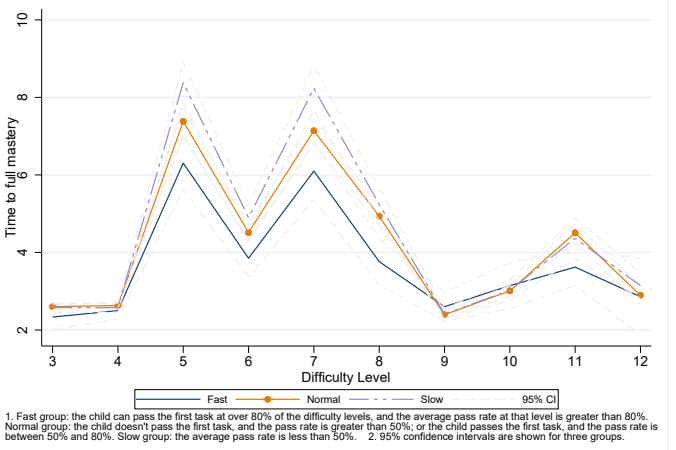
Figure 8: Time to Full Mastery for Language Tasks by Ability Category and Level

(a) Language Tasks



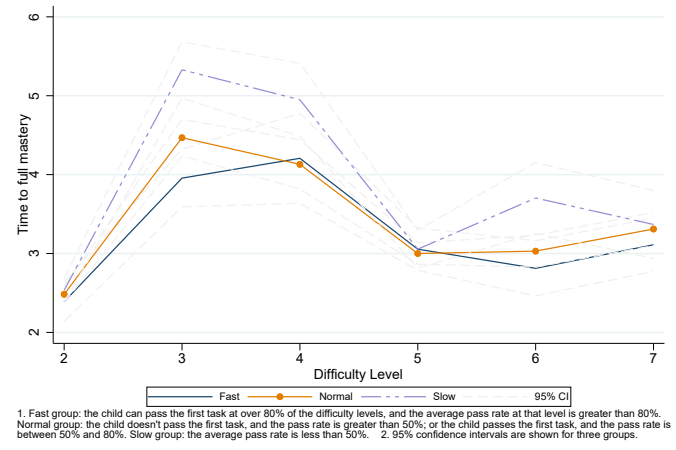
1. Fast group: the child can pass the first task at over 80% of the difficulty levels, and the average pass rate at that level is greater than 80%. Normal group: the child doesn't pass the first task, and the pass rate is greater than 50%; or the child passes the first task, and the pass rate is between 50% and 80%. Slow group: the average pass rate is less than 50%. 2. 95% confidence intervals are shown for three groups.

(b) Cognitive Tasks



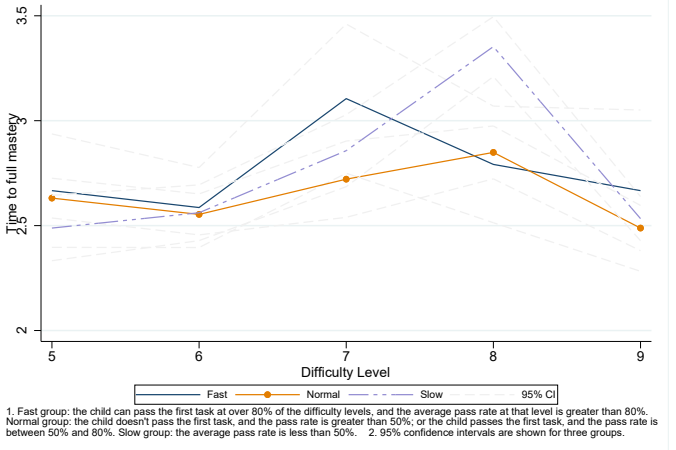
1. Fast group: the child can pass the first task at over 80% of the difficulty levels, and the average pass rate at that level is greater than 80%. Normal group: the child doesn't pass the first task, and the pass rate is greater than 50%; or the child passes the first task, and the pass rate is between 50% and 80%. Slow group: the average pass rate is less than 50%. 2. 95% confidence intervals are shown for three groups.

(c) Fine Motor Tasks



1. Fast group: the child can pass the first task at over 80% of the difficulty levels, and the average pass rate at that level is greater than 80%. Normal group: the child doesn't pass the first task, and the pass rate is greater than 50%; or the child passes the first task, and the pass rate is between 50% and 80%. Slow group: the average pass rate is less than 50%. 2. 95% confidence intervals are shown for three groups.

(d) Gross Motor Tasks



1. Fast group: the child can pass the first task at over 80% of the difficulty levels, and the average pass rate at that level is greater than 80%. Normal group: the child doesn't pass the first task, and the pass rate is greater than 50%; or the child passes the first task, and the pass rate is between 50% and 80%. Slow group: the average pass rate is less than 50%. 2. 95% confidence intervals are shown for three groups.

5 Testing Measured Skill Invariance

Agostinelli and Wiswall (2021) raise important questions about the existence of invariant measures of skill. They define mean measures of skill invariance. ***Mean measured skill invariance*** (our term) for measure $Z(s,a)$ of skill s at age a requires:

$$E(Z(s,a) \mid K(s,\ell,a) = \tau) = E(Z(s,a') \mid K(s,\ell,a') = \tau) \tag{3}$$

for $a \neq a'$; i.e., at the same *true skill level* τ, the measures of skill s at ages a and a' should coincide for all $a, a' \in [\underline{a}(\ell), \bar{a}(\ell)]$.

To conduct this test, we need to find groups with the same latent skill levels $K(s,\ell,a)$ at different ages and then measure the child task performance $Z(s,a)$ for the different age groups. For the treated children in the China REACH program, we have the task performance measures at each age and difficulty level ℓ for each skill. We also have conventional Denver test measures.

UHP define "true knowledge" at level ℓ for skill s as $K(s,\ell,a)$ for $a \in [\underline{a}(l), \bar{a}(l)]$. $K(s,\ell,a) \geq \bar{K}(s,\ell)$ is a measure of mastery *at that level* at age a for skill s. We use different measures of knowledge: average passing rate, time to mastery, and instability.

Consider using the average passing rate at each difficulty level as the measure of true skill for testing (3). The logic for other measures is the same, although as we have seen, they measure different aspects of knowledge. In this paper, we mainly focus on the test based on average passing rate.

5.1 Finding Groups with Same τ but Different a

For all children in the intervention, we calculate average passing rates at each difficulty level for each skill throughout the entire intervention. To avoid small cells for our measures of knowledge, we array the data by quantiles of passing rates in the order of difficulty level. Table 8 uses passing rates on language skills at level ℓ and skill s-specific disaggregated UHP measures to test the condition $K(s, \ell, a) = K(s, \ell, a') = \tau$ (equal passing rates), a precondition for a test of measure invariance comparing age a and a' aggregated Denver scores. Based on the average passing rate at each difficulty level, we group the children with similar task performance in the same group. At difficulty level 2, the children at the lowest quantile (τ_1) have the lowest passing rate (i.e., the passing rate is zero) and the children at quantile 4 (τ_4) have the highest passing rate (i.e., the passing rate is 100%).

We then order the children's enrollment age within each τ group. For example, in quantile τ_1, there are 117 children at level 2, and we order them by their ages at the time of enrollment. Ages are in $[\underline{a}_s(\ell), \bar{a}_s(\ell)]$. The "young" group for quantile τ_1 is the group of children who are in the bottom 50% of the ages. The "old" group rank in the top 50% by age.

For example, the mean of the passing rate for the group of younger children in quantile group 2 (τ_2) at difficulty level 3 is about 0.513, and the mean for the older group of children in quantile group 1 (τ_1) is about 0.514. A p-value for a test of equality is 0.97. Therefore, we do not reject the hypothesis that, for this group, $K(s, \ell, a) = K(s, \ell, a')$. However, within the same knowledge group, there are statistically significant age differences. For example, in quantile group 2 (τ_2) at

difficulty level 3, the mean age for the younger group is about 10 months old, and the mean age for the older group is about 14 months old. In Appendix B, Tables B.1–B.3 show comparable partitions for higher levels of language skill. Tables B.4–B.8 show comparable partitions for other skills across levels. For the vast majority of groups for all skills across all levels, there are groups with similar levels of knowledge but children of different ages. These are inputs into our test of (3).

Table 8: Test of the Condition That $K(s, \ell, a) = K(s, \ell, a')$ for **Language Skill** Using UHP Difficulty Levels (Up to Denver Endline Age)[3]

Level	Category	τ_1	τ_2	τ_3	τ_4
2	**Average Passing Rate**				
	Young	0	0.283	0.723	1
	Old	0	0.321	0.656	1
	Test $K(s, \ell, a) = K(s, \ell, a')$: p-value		0.148	**0.004**	
	N	117	112	112	108
	Latent Skill Range	[0, 0]	[0.077, 0.5]	[0.5, 0.917]	[1, 1]
	Age at Enrollment (Months)				
	Young	12.432	10.267	10.049	13.611
	Old	17.909	13.940	13.871	18.352
	Test $a = a'$: p-value	**0.000**	**0.000**	**0.000**	**0.000**
	Average Starting Age at Level 2				
	Monthly Age (Young)	13.186	10.543	10.179	14.676
	Monthly Age (Old)	19.103	13.991	14.478	20.000
	Curriculum Age Range for Level 2: [6.75, 20]				
3	**Average Passing Rate**				
	Young	0	0.513	1.000	
	Old	0	0.514	1.000	
	Test $K(s, \ell, a) = K(s, \ell, a')$: p-value		0.969		
	N	122	136	134	
	Latent Skill Range	[0, 0]	[0.2, 0.8]	[1, 1]	
	Age at Enrollment (Months)				
	Young	12.162	10.147	11.715	
	Old	17.140	13.866	16.480	
	Test $a = a'$: p-value	**0.000**	**0.000**	**0.000**	
	Average Starting Age at Level 3				
	Monthly Age (Young)	14.035	11.638	13.352	
	Monthly Age (Old)	17.671	15.310	17.286	
	Curriculum Age Range for Level 3: [9.5, 18.25]				

1. Groups are categorized by the passing rate for each skill. τ_1 is for the children with the lowest passing rate and τ_4 is for the children with the highest passing rate.
2. Within each group, we sort the children based on their monthly ages at the time of enrollment and generate two equal size subgroups named "young" and "old." The children whose enrollment ages are in the top 50% are categorized into the old group.
3. All the measures in the table are evaluated from the time of enrollment to the Denver endline measurement age.

5.2 Testing Measured Skill Invariance

We next test the hypothesis that the aggregate Denver tests for s-comparable skills satisfy the criterion $E(Z(s,a) \mid K(s,\ell,a) = \tau) = E(Z(s,a') \mid K(s,\ell,a') = \tau)$ for different skills. Our Denver test endline measures are comparable to other commonly used achievement and assessment tests such as the Bailey tests.

Tables 9a–9b report tests of whether the means of raw Denver scores are different (e.g., young vs. old) for each partition of τ at each difficulty level. We find that, for raw Denver scores, the old group's performance at the same level of measured knowledge is consistently better than the young group's performance; i.e., condition (3) is almost always violated, so the condition $E(Z(s,a) \mid K(s,\ell,a) = \tau) = E(Z(s,a') \mid K(s,\ell,a') = \tau)$ does not hold, even though the disaggregated measures of skill are the same. Measured skill invariance is rejected. Other factors beside pure knowledge of s, as we measure it, affect Denver tests. We report similar findings for cognitive and fine motor skill tests (see Tables 10a, 10b, and 11).

5.2.1 Language Skill

Table 9a: Tests of the Mean Differences of Raw Denver Language Score $Z(s,a)$ Conditional on Language τ Groups by Difficulty Levels (Up to Denver Endline Age)[3]

Denver	Category	τ_1	τ_2	τ_3	τ_4
		UHP Language Level 2			
Endline	Young	26.271	24.306	24.447	26.486
	Old	29.956	28.056	28.159	29.237
(Language and Cognitive)	p-value	**0.000**	**0.000**	**0.000**	**0.004**
		UHP Language Level 3			
Endline	Young	26.180	24.081	25.813	
	Old	28.786	28.191	27.957	
(Language and Cognitive)	p-value	**0.002**	**0.000**	**0.012**	
		UHP Language Level 4			
Endline	Young	26.949	24.580	23.882	25.872
	Old	29.278	27.889	27.553	28.892
(Language and Cognitive)	p-value	**0.023**	**0.000**	**0.000**	**0.000**
		UHP Language Level 5			
Endline	Young	24.966	23.940	25.250	
	Old	28.848	26.357	26.750	
(Language and Cognitive)	p-value	**0.000**	**0.000**	0.313	
		UHP Language Level 6			
Endline	Young	29.323	25.467	25.440	27.385
	Old	32.321	30.427	30.292	31.742
(Language and Cognitive)	p-value	**0.011**	**0.000**	**0.000**	**0.000**

1. Groups are categorized by the passing rate for each skill by level. τ_1 is for the children with the lowest passing rates and τ_3, τ_4, or τ_5 are for the children with the highest passing rates (according to the level). Levels 7 and 8 are divided into 5 equally sized groups sorted by the passing rate. Levels 9 and 11 are divided into three groups: τ_3 corresponds to children with passing rate one, and τ_1 and τ_2 are equally divided according to the rest of the sample. Level 10 is divided into four groups: τ_1 corresponds to children with passing rate zero, τ_4 to children with passing rate one.
2. Within each group, we sort the children based on their monthly ages at the time of enrollment and generate two equal size subgroups named "young" and "old." The children whose enrollment ages are in the top 50% are categorized into the old group.
3. All the measures in the table are evaluated from the time of enrollment to the Denver endline measurement age.

Table 9b: Tests of the Mean Differences of Raw Denver Language Score $Z(s,a)$ Conditional on Language τ Groups by Difficulty Levels (Up to Denver Endline Age)[3]

Denver	Category	τ_1	τ_2	τ_3	τ_4	τ_5
		UHP Language Level 7				
Endline	Young	27.148	27.518	26.183	26.182	25.532
	Old	30.300	32.145	31.067	31.725	31.042
(Language and Cognitive)	p-value	**0.003**	**0.000**	**0.000**	**0.000**	**0.000**
		UHP Language Level 8				
Endline	Young	26.942	27.000	26.102	28.237	25.339
	Old	29.333	31.442	32.526	32.320	30.600
(Language and Cognitive)	p-value	**0.025**	**0.000**	**0.000**	**0.000**	**0.000**
		UHP Language Level 9				
Endline	Young	27.500	29.516	25.773		
	Old	31.525	32.247	30.615		
(Language and Cognitive)	p-value	**0.000**	**0.000**	**0.000**		
		UHP Language Level 10				
Endline	Young	25.579	28.048	30.756	27.692	
	Old	28.300	29.692	32.886	32.136	
(Language and Cognitive)	p-value	0.163	0.151	**0.005**	**0.000**	
		UHP Language Level 11				
Endline	Young	27.129	27.519	26.063		
	Old	30.609	32.218	31.072		
(Language and Cognitive)	p-value	**0.000**	**0.000**	**0.000**		

1. Groups are categorized by the passing rate for each skill by level. τ_1 is for the children with the lowest passing rates and τ_3, τ_4, or τ_5 are for the children with the highest passing rates (according to the level). Levels 7 and 8 are divided into 5 equally sized groups sorted by the passing rate. Levels 9 and 11 are divided into three groups: τ_3 corresponds to children with passing rate one, and τ_1 and τ_2 are equally divided according to the rest of the sample. Level 10 is divided into four groups: τ_1 corresponds to children with passing rate zero, τ_4 to children with passing rate one.

2. Within each group, we sort the children based on their monthly ages at the time of enrollment and generate two equal size subgroups named "young" and "old." The children whose enrollment ages are in the top 50% are categorized into the old group.

3. All the measures in the table are evaluated from the time of enrollment to the Denver endline measurement age.

5.2.2 Cognitive Skill

Table 10a: Tests of the Mean Differences of Raw Denver Score $Z(s,a)$ Conditional on Cognitive τ Groups by Difficulty Levels (Up to Denver Endline Age)[3]

Denver	Category	τ_1	τ_2	τ_3	τ_4	τ_5
		UHP Cognitive Level 2				
Endline	Young	28.815	26.474	24.986	24.871	
	Old	31.023	30.186	29.621	30.323	
(Language and Cognitive)	p-value	**0.026**	**0.000**	**0.000**	**0.000**	
		UHP Cognitive Level 3				
Endline	Young	26.750	26.063	25.167		
	Old	30.479	29.533	29.317		
(Language and Cognitive)	p-value	**0.000**	**0.000**	**0.000**		
		UHP Cognitive Level 4				
Endline	Young	26.917	25.161	25.605		
	Old	30.767	30.219	30.301		
(Language and Cognitive)	p-value	**0.000**	**0.000**	**0.000**		
		UHP Cognitive Level 5				
Endline	Young	27.234	26.723	26.010		
	Old	31.536	31.725	31.241		
(Language and Cognitive)	p-value	**0.000**	**0.000**	**0.000**		
		UHP Cognitive Level 6				
Endline	Young	27.277	27.536	25.761		
	Old	31.162	32.408	30.560		
(Language and Cognitive)	p-value	**0.000**	**0.000**	**0.000**		
		UHP Cognitive Level 7				
Endline	Young	27.018	26.863	26.981	26.389	25.845
	Old	30.085	32.235	32.132	31.063	30.870
(Language and Cognitive)	p-value	**0.003**	**0.000**	**0.000**	**0.000**	**0.000**

1. Groups are categorized by the passing rate for each skill by level. τ_1 is for the children with the lowest passing rates, and τ_3, τ_4, or τ_5 are for the children with the highest passing rates (according to the level). Level 2 is divided into four groups: τ_1 corresponds to children with passing rate zero, τ_4 to children with passing rate one, and τ_2 and $\tau 3$ are equally divided according to the rest of the sample. Levels 8–12 are divided into three groups. Levels 9, 11, and 12 are constructed as follows: τ_1 corresponds to children with passing rate zero, τ_3 to children with passing rate one, and τ_2 to the rest. For Level 8, τ_3 corresponds to children with passing rate one, and τ_1 and τ_2 are divided into equal sizes. Finally, Level 13 is divided into 2 groups: τ_1 for children with passing rate zero, and τ_2 for children with passing rate one.

2. Within each group, we sort the children based on their monthly ages at the time of enrollment and generate two equal size subgroups named "young" and "old." The children whose enrollment ages are in the top 50% are categorized into the old group.

3. All the measures in the table are evaluated from the time of enrollment to the Denver endline measurement age.

Table 10b: Tests of the Mean Differences of Raw Denver Score $Z(s,a)$ Conditional on Cognitive τ Groups by Difficulty Levels (Up to Denver Endline Age)[3]

Denver	Category	τ_1	τ_2	τ_3
		UHP Cognitive Level 8		
Endline	Young	27.071	28.011	25.786
	Old	30.627	32.704	30.625
(Language and Cognitive)	p-value	**0.000**	**0.000**	**0.000**
		UHP Cognitive Level 9		
Endline	Young	25.676	26.842	28.229
	Old	29.517	31.649	32.061
(Language and Cognitive)	p-value	**0.011**	**0.000**	**0.000**
		UHP Cognitive Level 10		
Endline	Young	26.360	30.328	28.909
	Old	28.947	31.750	32.355
(Language and Cognitive)	p-value	0.076	0.120	**0.000**
		UHP Cognitive Level 11		
Endline	Young	27.647	29.605	29.735
	Old	30.231	32.271	31.686
(Language and Cognitive)	p-value	0.140	**0.000**	**0.005**
		UHP Cognitive Level 12		
Endline	Young	27.154	30.767	30.838
	Old	30.077	32.367	32.207
(Language and Cognitive)	p-value	0.217	0.141	**0.013**
		UHP Cognitive Level 13		
Endline	Young	29.564	30.984	
	Old	31.833	32.085	
(Language and Cognitive)	p-value	**0.029**	0.100	

1. Groups are categorized by the passing rate for each skill by level. τ_1 is for the children with the lowest passing rates, and τ_3, τ_4, or τ_5 are for the children with the highest passing rates (according to the level). Level 2 is divided into four groups: τ_1 corresponds to children with passing rate zero, τ_4 to children with passing rate one, and τ_2 and $\tau 3$ are equally divided according to the rest of the sample. Levels 8–12 are divided into three groups. Levels 9, 11, and 12 are constructed as follows: τ_1 corresponds to children with passing rate zero, τ_3 to children with passing rate one, and τ_2 to the rest. For Level 8, τ_3 corresponds to children with passing rate one, and τ_1 and τ_2 are divided into equal sizes. Finally, Level 13 is divided into 2 groups: τ_1 for children with passing rate zero, and τ_2 for children with passing rate one.

2. Within each group, we sort the children based on their monthly ages at the time of enrollment and generate two equal size subgroups named "young" and "old." The children whose enrollment ages are in the top 50% are categorized into the old group.

3. All the measures in the table are evaluated from the time of enrollment to the Denver endline measurement age.

5.2.3 Fine Motor Skill

Table 11: Tests of the Mean Differences of Raw Denver Score $Z(s,a)$ Conditional on Fine Motor τ Groups by Difficulty Levels (Up to Denver Endline Age)[3]

Denver	Category	τ_1	τ_2	τ_3	τ_4	τ_5
		UHP Fine Motor Level 1				
Endline	Young	22.000	20.891	20.771	21.123	
	Old	23.000	22.879	22.333	23.141	
(Fine Motor)	p-value	**0.008**	**0.000**	**0.000**	**0.000**	
	N	46	79	75	159	
		UHP Fine Motor Level 2				
Endline	Young	21.848	21.326	21.305		
	Old	23.767	23.125	23.256		
(Fine Motor)	p-value	**0.000**	**0.000**	**0.000**		
	N	63	78	270		
		UHP Fine Motor Level 3				
Endline	Young	21.659	22.089	21.457	22.111	21.130
	Old	23.825	24.118	23.811	23.805	23.089
(Fine Motor)	p-value	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
	N	84	79	83	86	182
		UHP Fine Motor Level 4				
Endline	Young	21.378	22.196	21.695		
	Old	23.590	23.886	23.626		
(Fine Motor)	p-value	**0.000**	**0.000**	**0.000**		
	N	84	95	311		
		UHP Fine Motor Level 5				
Endline	Young	22.412	22.806	22.460		
	Old	23.767	24.036	23.926		
(Fine Motor)	p-value	**0.000**	**0.000**	**0.000**		
	N	64	59	219		
		UHP Fine Motor Level 6				
Endline	Young	22.533	22.968	22.921		
	Old	23.667	24.290	24.000		
(Fine Motor)	p-value	**0.002**	**0.000**	**0.000**		
	N	60	62	159		
		UHP Fine Motor Level 7				
Endline	Young	22.667	23.031	23.172		
	Old	23.800	23.897	24.067		
(Fine Motor)	p-value	**0.059**	**0.006**	**0.000**		
	N	30	61	124		

1. Groups are categorized by the passing rate for each skill by level. τ_1 are fore children with the lowest passing rates, and τ_3, τ_4, or τ_5 are for children with the highest passing rates (according to the level). Level 1 is divided into four groups: τ_1 corresponds to children with passing rate zero, τ_4 correspond to children with passing rate one, and τ_2 and τ_3 are equally divided according to the rest of the sample. Levels 2 and 4–7 are divided into three groups. Levels 2 and 7 are constructed as follows: τ_1 corresponds to children with passing rate zero, τ_3 to children with passing rate one, and τ_2 to the rest. For Levels 4–6, τ_3 corresponds to children with passing rate one, and τ_1 and τ_2 are divided into equal sizes. Level 3 is divided into 5 groups with equal sizes.
2. Within each group, we sort the children based on their monthly ages at the time of enrollment and generate two equal size subgroups named "young" and "old." The children whose enrollment ages are in the top 50% are categorized into the old group.
3. All the measures in the table are evaluated from the time of enrollment to the Denver endline measurement age.

5.2.4 Up to Midline Measures

Appendix D reports comparable tests using Denver midline scores (i.e., all measures are evaluated from the time of the child's enrollment to the time of the Denver midline test). Tables D.1–D.3 present the test of $K(s, \ell, a) = K(s, \ell, a') = \tau$ up to the Denver midline measurement age. For each difficulty level, we only consider the tasks that are conducted before the Denver midline measurement age. We reach the same conclusion as obtained for the endline measures.

5.3 Denver Language Test Results

The previous tests report tests of hypothesis (3) using combined Denver language and cognitive tests. Scores are combined because there are few Denver test items for cognition. Our rejections for the Denver tests may be a consequence of these scores combining conceptually distinct skills.

We conduct a similar series of tests using only language tests. In Tables 12a–12b, we continue to reject the skill invariance assumption for language skill even after only considering the Denver language items.

Table 12a: Tests of the Mean Differences of Raw Denver Language Score $Z(s,a)$ Conditional on Language τ Groups by Difficulty Levels (Up to Denver Endline Age)[3]

Denver	Category	τ_1	τ_2	τ_3	τ_4
		UHP Language Level 2			
Endline	Young	22.229	20.652	21.463	22.405
	Old	24.622	23.976	22.789	24.026
(Language)	p-value	**0.000**	**0.000**	**0.009**	**0.011**
		UHP Language Level 3			
Endline	Young	22.220	20.774	21.958	
	Old	23.667	23.489	23.191	
(Language)	p-value	**0.012**	**0.000**	**0.032**	
		UHP Language Level 4			
Endline	Young	22.744	20.902	21.059	21.974
	Old	24.056	23.143	23.132	23.757
(Language)	p-value	**0.056**	**0.000**	**0.000**	**0.001**
		UHP Language Level 5			
Endline	Young	21.458	20.700	21.750	
	Old	23.909	22.167	22.500	
(Language)	p-value	**0.000**	**0.000**	0.455	
		UHP Language Level 6			
Endline	Young	24.387	21.987	21.713	22.949
	Old	26.536	25.123	24.623	26.097
(Language)	p-value	**0.009**	**0.000**	**0.000**	**0.000**

1. Groups are categorized by the passing rate for each skill by level. τ_1 is for the children with the lowest passing rates, and τ_3 or τ_4 are for the children with the highest passing rates (according to the level). Levels 2, 4, and 6 are divided in four groups: τ_1 corresponds to children with passing rate zero, τ_4 to children with passing rate one, and τ_2 and τ_3 are equally divided by the rest of the sample. Levels 3 and 5 are divided into three groups: τ_1 corresponds to children with passing rate zero, τ_3 to children with passing rate one, and τ_2 to the rest.

2. Within each group, we sort the children based on their monthly ages at the time of enrollment and generate two equal size subgroups named "young" and "old." The children whose enrollment ages are in the top 50% are categorized into the old group.

3. All the measures in the table are evaluated from the time of enrollment to the Denver endline measurement age.

Table 12b: Tests of the Mean Differences of Raw Denver Language Score $Z(s,a)$ Conditional on Language τ Groups by Difficulty Levels (Up to Denver Endline Age)[3]

Denver	Category	τ_1	τ_2	τ_3	τ_4	τ_5
		UHP Language Level 7				
Endline	Young	22.833	22.911	22.361	22.056	21.729
	Old	24.980	26.309	25.659	26.000	25.447
(Language)	p-value	**0.004**	**0.000**	**0.000**	**0.000**	**0.000**
		UHP Language Level 8				
Endline	Young	22.712	22.672	22.276	23.210	21.673
	Old	24.286	25.977	26.526	26.479	25.109
(Language)	p-value	**0.032**	**0.000**	**0.000**	**0.000**	**0.000**
		UHP Language Level 9				
Endline	Young	23.333	24.355	21.883		
	Old	25.750	26.476	25.198		
(Language)	p-value	**0.000**	**0.000**	**0.000**		
		UHP Language Level 10				
Endline	Young	21.842	23.698	24.953	23.154	
	Old	23.500	24.675	26.886	26.311	
(Language)	p-value	0.187	0.202	**0.003**	**0.000**	
		UHP Language Level 11				
Endline	Young	22.803	23.013	22.099		
	Old	25.217	26.385	25.505		
(Language)	p-value	**0.000**	**0.000**	**0.000**		

1. Groups are categorized by the passing rate for each skill by level. τ_1 is for the children with the lowest passing rates, and τ_3 or τ_4 are for the children with the highest passing rates (according to the level). Levels 2, 4, and 6 are divided in four groups: τ_1 corresponds to children with passing rate zero, τ_4 to children with passing rate one, and τ_2 and τ_3 are equally divided by the rest of the sample. Levels 3 and 5 are divided into three groups: τ_1 corresponds to children with passing rate zero, τ_3 to children with passing rate one, and τ_2 to the rest.

2. Within each group, we sort the children based on their monthly ages at the time of enrollment and generate two equal size subgroups named "young" and "old." The children whose enrollment ages are in the top 50% are categorized into the old group.

3. All the measures in the table are evaluated from the time of enrollment to the Denver endline measurement age.

5.4 Robustness to Age of Entry

A feature of China REACH is that all children of the same age are taught and examined on the same task. The late entrants have fewer lessons and may not be at the same level of knowledge due to dynamic complementarity of knowledge (see, e.g., Heckman and Zhou, 2022). However, we condition on knowledge $K(s, \ell, a)$ attained, so this consideration does not affect our analysis. Nonetheless, we conduct a series of robustness checks and find that our conclusions are not affected by alternative treatments of late entrants. See Appendix E.

6 Conclusion

This paper tests and rejects a key assumption invoked in the economics of education and in the analysis of skill formation: the existence of invariant measures of skill across different levels of the same skill ("human capital"). This assumption underlies a large body of research in the social sciences. Value-added measures are widely used to measure the output of schools. Aggregate test scores are used to measure gaps in skills across demographic groups.

This paper shows that this practice is unwise. The aggregate measures used to chart student gains, child development, and the contribution of teachers and caregivers to student development are not comparable over time and persons except, possibly, for narrowly defined measures of skill. Our results on the nonexistence of globally valid invariant scales for skills such as cognition, language, and fine arts reported in this paper are consistent with results obtained from the model of Heckman

and Zhou (2021). Accurate skill measurement requires much more disaggregated approaches, and conventional measures that assume invariance are fragile and should be used with caution if at all.

References

Agostinelli, F., M. Saharkhiz, and M. J. Wiswall (2019). Home and school in the development of children. *National Bureau of Economic Research* (Paper No. 26037).

Agostinelli, F. and M. Wiswall (2021). Estimating the technology of children's skill formation. Revision requested, Journal of Political Economy.

Appelbaum, A. S. (1978). Validity of the revised denver developmental screening test for referred and nonreferred samples. *Psychological Reports 43*(1), 227–233.

Cawley, J., J. J. Heckman, and E. J. Vytlacil (1999, November). On policies to reward the value added by educators. *Review of Economics and Statistics 81*(4), 720–727.

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.

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

Heckman, J. and J. Zhou (2022). Nonparametric tests of dynamic complementarity. Unpublished Paper, 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 d l'INSEE 30–31*, 227–269. Special Issue.

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 J. Zhou (2021). Interactions as investments: The microdynamics and measurement of early childhood learning. Unpublished Paper, University of Chicago.

Hill, H. C. (2009). Evaluating value-added models: A validity argument approach. *Journal of Policy Analysis and Management 28*(4), 700–709.

Konstantopoulos, S. (2014). Teacher effects, value-added models, and accountability. *Teachers College Record 116*(1).

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

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

Rivkin, S. G., E. A. Hanushek, and J. F. Kain (2005). Teachers, schools, and academic achievement. *Econometrica 73*(2), 417–458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review 94*(2), 247–252.

Rubio-Codina, M., M. C. Araujo, O. Attanasio, P. Muñoz, and S. Grantham-McGregor (2016). Concurrent validity and feasibility of short tests currently used to measure early childhood development in large scale studies. *PLoS ONE 11*(8), 1–17.

Rubio-Codina, M. and S. Grantham-McGregor (2020). Predictive validity in middle childhood of short tests of early childhood development used in large scale studies compared to the Bayley-III, the Family Care Indicators, height-for-age, and stunting: A longitudinal study in Bogota, Colombia. *PLoS ONE 15*(4), 1–20.

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. Hunt (1975). Assessment in infancy: Ordinal scales of psychological development. *University of Illinois Press*.

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

Zhou, J., J. Heckman, B. Liu, and M. Lu (2021). The impacts of a prototypical home visiting program on child skills. Working Paper.

Zhou, J., J. Heckman, F. Wang, and B. Liu (2021). Tests of early childhood skill learning patterns. Revision requested, *Journal of Community Psychology*.