NBER WORKING PAPER SERIES

MEASURING KNOWLEDGE AND LEARNING

James J. Heckman
Jin Zhou

Working Paper 29990
http://www.nber.org/papers/w29990

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
April 2022, Revised November 2022

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 the suggestions of Flavio Cunha made at Rice University in November 2021. Matthew Wiswall made very useful comments. Alejandra Campos, Fanmei Xia, and Haihan Tian contributed highly competent and insightful research assistance. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Measuring Knowledge and Learning
James J. Heckman and Jin Zhou
NBER Working Paper No. 29990
April 2022, Revised November 2022
JEL No. C81,I21,J71

ABSTRACT

Empirical studies in the economics of education, the measurement of skill gaps, the impacts of interventions on skill formation, and the value-added literature rely on psychometrically validated test scores. Test scores are taken as measures of an invariant scale of human capital compared over time and people. We examine if conventional skill measures are comparable on mastery of specific task knowledge. An unusually rich dataset from an early childhood intervention is used to test the assumption of scale invariance. We reject the scale invariance hypothesis for multiple skills and cast doubt on the uncritical use of test scores in research.

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

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

An online appendix is available at http://cehd.uchicago.edu/measuring-knowledge

1 Introduction

A crucial assumption maintained in the literature on skill formation, racial, ethnic and gender 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 for the same people and across 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] The assumption of such measures motivates studies of skill gaps across demographic groups (Cunha et al., 2021), value-added models in education (Konstantopoulos, 2014; Rivkin et al., 2005; Hill, 2009; Rockoff, 2004), and studies of skill formation (Agostinelli and Wiswall, 2022) charting the development of children. Education policies are often assessed by PISA scores that are based on this assumption. This paper tests for and rejects the hypothesis that such invariant measures exist for prototypical achievement and assessment tests using a unique Chinese data set.

Test scores are psychometric creations (see, e.g., 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; Cunha et al., 2010; Agostinelli and Wiswall, 2022; Freyberger, 2021; Cunha et al., 2021). This paper examines the

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

foundations of this approach.

The crucial assumption in this paper is that mastery of tasks *within a well-defined skill level* is an accurate measure of knowledge. In the China REACH intervention that we analyze, the curriculum design supports this assumption. It is based on a widely used framework for measuring knowledge designed by Uzgiris and Hunt (1975) and Palmer (1971) (henceforth UHP). In this framework, performance on tasks of the same knowledge content are evaluated multiple times. Using these scales, we can chart mastery of skills within the same knowledge levels and can compare knowledge and its growth across children on a common micro-scale. We can also measure transitions across levels and hence can determine, at least in an ordinal sense, whether or not there is growth in knowledge because the tasks across levels are clearly ordered. Cardinalizing the magnitude of that growth is another matter, unless an invariant scale across levels is assumed.

The scales used in our tests are intuitively valid. Children can either perform a task or not. We use this basic measure to assess the validity of 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 summaries of binary task performance as measures of knowledge that can be used to create meaningful cardinal comparisons across people, time, or skill levels.

The UHP measures we use are based on the performance of children on common tasks of equal difficulty. The *weekly* tasks we analyze within levels are well defined and clearly classified into developmental levels. Within narrowly defined levels, tasks have the same knowledge content. A child's mastery of these tasks within a level is

a precisely defined measure of knowledge fully comparable across children and over time. An ordinal measure of learning is the mastery of progressively more difficult tasks. The question is whether, across levels, the scales measure growth of the same thing ("human capital").

Mastery can be measured in multiple ways. In addition to investigating traditional measures (i.e., aggregated passing rates across levels), we examine other measures that might be used to capture knowledge of skills. For example, time to first mastery captures how quickly children master task content within levels. The measure can also be compared across levels, although no cardinal scale necessarily exists. Instability (backsliding) is another measure that captures the persistence of skill mastery after the first success. We examine agreement among them using non-parametric methods. These alternative measures are correlated among themselves and with traditional measures in the expected direction, though far from perfectly. They capture different aspects of knowledge and learning.

A precise definition of mean age invariance of measures of skill was introduced to the literature in Agostinelli and Wiswall (2022). It requires existence of a common mean scale across people of different ages who command identical mastery of tasks. Our analysis shows that, in our data, prototypical tests for language, cognitive, and motor skills are not well described by a common scale.

Given the widespread use of cognitive tests we applied work in economics, this finding is of great importance. Test scores depend on the age when measured and not just mastery of task content. Accurate skill measurement requires more fine grained approaches. Conventional measures that assume invariance are fragile and

should be used with caution, if at all. In contrast, anchoring relates test scores to objectively comparable outcomes (e.g., wages, education, criminality, employment) generates interpretable scales.

This paper is organized as follows. Section 2 discusses our data. 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 test the existence of age invariance, i.e., a constant-unit measuring stick. We reject such hypothesis. Section 6 concludes.

2 Our Data of Knowledge and Skill

This section describes our data source. We document the background of the China REACH program, the curriculum designed in the program, and the implementation of the assessment.

2.1 Background of China REACH Program

The measurement and development of multiple skills in young children has been extensively studied. The UHP measures are collected to evaluate an early childhood intervention program in China.

The China REACH program that we analyze is adapted from the Jamaican Reach Up and Learn program, which was designed using UHP as a framework to understand and support child growth and development.[2] The tasks children confront in China

[2]See Grantham-McGregor et al. (1997).

REACH cover four domains of skill: fine motor, gross motor, language, and cognitive skills.

The program was implemented in 2015 in a large-scale randomized control trial. It enrolled 1,500 participants aged 9-30 months (about 700 participants in the treatment group) in 111 villages in Huachi county, Gansu province, one of the poorest areas in China (Zhou, Heckman, Liu, and Lu, 2022). Trained home visitors visit each treated household weekly and provide one hour of parenting or child caregiving guidance. Multiple skills are fostered and tested. The program 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.

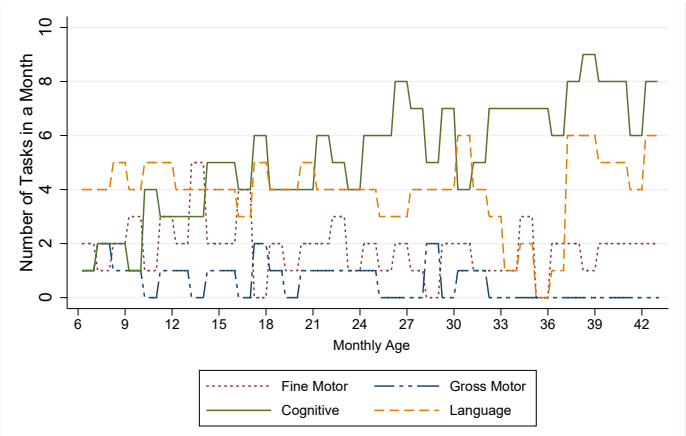
2.2 Curriculum and Assessment Design and Implementation

Three or four different skills (from a group of gross motor, fine motor, language, and cognitive skills) are taught and examined each week. They are organized within homogenous skill levels and across hierarchies of knowledge. Figure 1 displays a crude schematic of the curriculum for the skills taught and measured at each age. We later discuss the specific skills taught and how they are measured.

Difficulty levels are ordered by average child performance (see Palmer (1971)).[3] The curriculum is designed based on the child weekly ages. All children of a given

[3]Palmer's team conducted an intervention on 240 African American males in Harlem in 1964 to boost their skill development. They found long-term effects for the children trained by their materials and procedures. The measures are calibrated against that sample. There is broad agreement in the child development community of a common general pattern of child development, although individual children may deviate from it. See Ertem et al. (2018); Fernald et al. (2017); WHO Multicentre Growth Referece Study Group and Onis (2007).

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



age confront the same tasks. In the field, home visitors strictly follow the design of the curriculum, which means that regardless of the child's performance, home visitors provide training in the curriculum based on the child's actual weekly age.[4] This strict delivery method provides an ideal environment to analyze child progress. We describe the details of cognitive and fine motor skill task content in the following subsections. Detailed descriptions of the assessments of other skills are presented in Appendices A and B.

[4]In the future intervention, we plan to compare this strict implementation delivery method to a more flexible delivery method.

Table 1: Difficulty Levels for 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.

2.3 Cognitive Skills

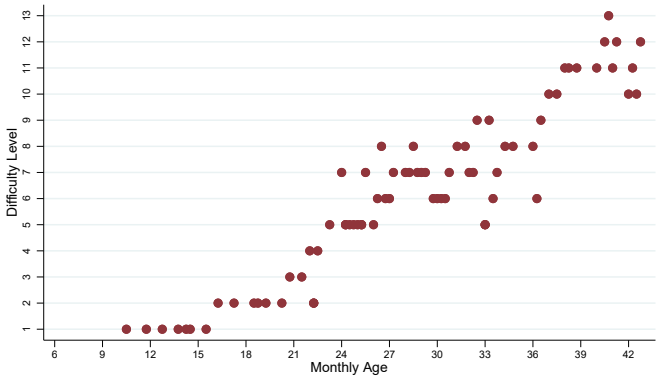
There are thirteen difficulty levels for cognitive skills (see Table 1). Cognitive skills have different dimensions. In the curriculum, cognitive skills taught cover spatial skills, knowledge of objects and object functions, order and number, etc. We use knowledge of objects and object functions as an example of the teaching and assessment curriculum. Cognitive skill difficulty levels are defined based on the abstract concepts shown in Table 1, such as the child's proficiency in performing common tasks. Across the thirteen ordered difficulty levels, there are seventy-four lessons.[5] The lessons cover the process of how the child learns to know an object and understand its function.

The cognitive knowledge of objects tasks progress 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 subsequent lessons, which increase with difficulty as 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.

[5]The difficulty level has ordinal meaning only, not necessary cardinal meaning. Age invariance assumes cardinal meaning.

Figure 2 displays the evolution of the assessments of the levels of cognitive skills (knowing objects and understanding object functions). The number of lessons varies across difficulty levels. As children age and advance across difficulty levels, they confront more demanding tasks.[6]

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.

Table 2 presents detailed information about the seven lessons (and assessments) in difficulty level 1 directed to ten-month-old to fifteen-month-old children. In Table 2, although the learning materials are different (e.g., Picture book A and B), 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. In addition, all the evaluation rules are the same for these tasks. There is no hierarchy of tasks within levels. In fact, some tasks are exactly repeated while others are slightly altered. Appendix B documents the task content for each difficulty level for all of the skills.

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

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: baby makes sound when looking at the pictures

2.4 Fine Motor Skills

As another example, consider fine motor drawing lessons which have seven difficulty levels.[7] 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 progresses from simple shapes to representative drawing (see Table 3).

In addition to tasks of different difficulty levels, the curriculum features multiple lessons and assessments *within* the same difficulty level. Table 4 presents the content of all tasks at level one for fine motor skills. 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. See Tables B.18-B.20 in Appendix B.

[7]The standard generating the difficulty levels is based on an understanding of the content in the tasks assigned.

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

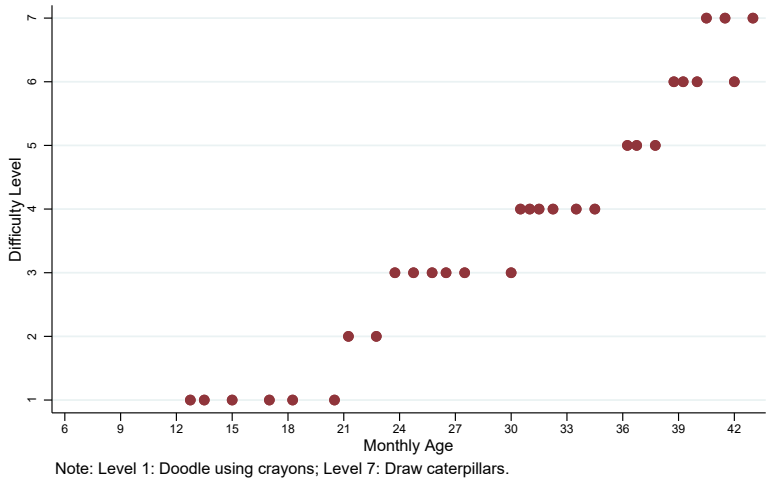
Difficulty Level	Task Content
1	Doodle using crayons
2	Mimic circles
3	Mimic circles and draw straight lines
4	Draw a circle, vertical line, and horizontal line
5	Draw circles, many lines, and crossed lines
6	Draw a cross (or T), curves, and zigzag curves
7	Draw caterpillars

Table 4: Fine Motor Task Content (Drawing) Level One

Difficulty Level	Difficulty Level Aim	Month	Week	Learning Materials	Task Aim and Content
Level 1	Doodle using crayons	12	3	Crayon and paper	Child doodles on the paper.
Level 1	Doodle using crayons	13	2	Crayon and paper	Child doodles on the paper.
Level 1	Doodle using crayons	14	4	Crayon and paper	Child doodles on the paper.
Level 1	Doodle using crayons	16	4	Crayon and paper	Child doodles on the paper.
Level 1	Doodle using crayons	18	1	Crayon and paper	Child scribbles on the paper.
Level 1	Doodle using crayons	20	2	Crayon and paper	Child doodles on the paper.

Figure 3 displays 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 overlaps across difficulty levels. When fine motor lessons at difficulty level 7 start, students can still receives lessons and assessments at difficulty level 6. Circling back is a strategy designed to solidify a child's understanding of a concept. Appendices A and B discuss in detail all of the skills we measure.

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



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

2.5 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 within each level of skill as our fundamental measure of knowledge. Knowledge is acquired in real time. By design, knowledge content is the same within each difficulty level. It is also the same across all children of the same age. It may be forgotten or retained as children advance through the curriculum. There are different measures of knowledge, which we next present.

3 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 PISA, ACT, SAT, Iowa Test, 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 one might measure knowledge and learning, it is helpful to introduce some notations.

Let $\mathcal{S}$ be the set of skills taught. Let $\ell(s,a)$ be the level of skill s (index of a set of tasks) taught at age a; $\ell(s,a) \in \{1,\ldots,\overline{L}_s\}$, an ordered set. $\overline{L}_s$ is the maximal difficulty level taught for each skill s. Mastery of skill s at level ℓ at age a

is characterized by a binary threshold crossing model:

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

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 Item Response Theory (IRT) model in educational psychology (Lord and Novick, 1968; van der Linden, 2016). In our case, $\bar{K}(s,\ell)$ indicates whether a child can perform tasks relative to skill s at level ℓ at age a.

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, $1+\bar{a}(s,\ell)-\underline{a}(s,\ell)$ is the length of the run (# of lessons measured on skill s at level ℓ) starting at age $\underline{a}(s,\ell)$. In our data, there is no hierarchy of tasks *within levels*. For level ℓ of skill s, we collect the indicators of knowledge in spell ℓ:

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

This records the age-by-age mastery of tasks at level ℓ for skill s.

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. This implicitly assumes that different levels capture the

same or similar content. A more refined measure that recognizes heterogeneity in knowledge across levels is defined *within a difficulty level* (s, ℓ) for skill s. The passing rate on skill s at level ℓ is:

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

The overall passing rate is:

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

which weights all items across all difficulty levels equally and tends to put more weight on difficulty levels with more tested items. This conventional measure does not standardize for the level of difficulty at level ℓ, the sampling frequency of items in ℓ, or the retention of knowledge, or the speed of acquisition to be discussed next.

There are other possible measures of knowledge and knowledge acquisition. 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 ***age at full mastery*** as $\tilde{a}(s,\ell) = \min_a [D(s,\ell,a) = 1, \forall a \geq \tilde{a}(s,\ell)]$.[8] ***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

[8] We define time to first mastery using the number of tasks a child attempts 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.

$\hat{a}(s,\ell)$, etc. ***Instability*** (Backsliding) at level ℓ for skill s is:

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

Speed of learning is sometimes used to assess IQ (van der Linden, 2016), whereas full mastery captures retention of acquired knowledge and backsliding captures forgetting.

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 our measure of traditional scores (Appelbaum, 1978). 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. Denver tests are commonly used in clinical examinations for early childhood skill development. It has an established observer validity and reliability (see Frankenburg and Dodds (1967) and Frankenburg et al. (1971)). Denver scores are very closely related to Bayley scores used to measure child development (Rubio-Codina and Grantham-McGregor, 2020; Rubio-Codina et al., 2016).[9] Tables 5a–5d present the correlations between the Denver scores at midline 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 level at which the Denver

[9] Ryu and Sim (2019) report that the Denver test is more accurate than the Bayley test in detecting the delay of language development.

test is administered. *"Up to Denver Endline"* age means that when we calculate the passing rate, time to mastery, and instability measures, we use all the treated children's weekly task performance data from the time the children enrolled into the program to the endline of the intervention.

The comparisons of the Denver scores with the measures introduced in the previous section are made in the following way. Endline measures of skill s are taken at the end of level s, $\overline{L}_s$, by design. The midline measures are taken at L_s^*. We could compare our measures of knowledge at level $\overline{L}_s$ or L_s^*, or else, based on an average measure $p(s)$ as defined in Equation (2), an unweighed average over all items and levels. The average passing rate up through midline, $p(s, L_s^*)$, is based on the measures up through L_s^* in Equation (2) rather than through $\overline{L}_s$. Such average measures are traditional.

Less traditional are average measures *at* endline and midline for the other measures. We replace $p(s, \ell)$ in Equation (2) with the measures previously defined for each level. We use the same weights by level as used for the average passing rate but replace $p(s, \ell)$ with the measures previously introduced. The average can be defined respectively as before through $\overline{L}_s$ or L_s^*. Tables labeled "Up to midline Denver ages", report the correlations of the averages of these measures with Denver scores evaluated at the dates July 2016 for midline and July 2017 for endline.

Denver scores are negatively correlated with the time the child takes to achieve first success and negatively correlated with the proportion of failed attempts after the first success. Compared to fine and gross motor scores, the language and cognitive scores have more statistically significant correlations with the corresponding Denver

counterparts measured by Denver passing rates. 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 comparable but weaker than the correlation with conventional passing rate, as measured by p values. This weaker correlation suggests that Denver scores do not capture other dimensions of knowledge as well as it does the conventional passing rate, but even that correlation with conventional passing rate is not especially strong.

Table 5a: Correlation between Average Passing Rate (Up to Midline/Endline Measurement Level) 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

Notes: 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 calculated using the tasks evaluated from the time of enrollment up to Denver midline measurement age and for the Denver score (endline) rows, it reports the correlations between the endline Denver scores and average passing rates calculated using the tasks evaluated from the time of enrollment up to Denver endline measurement age. 3. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

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

In addition to correlating knowledge measured over intervals up through the time the Denver test was administered, it is useful to measure knowledge at the exact level

Table 5b: Correlation between the Average of the Times to First Mastery (Up to Midline/Endline Measurement Level) 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

Notes: 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 calculated using the tasks evaluated from the time of enrollment up to Denver midline measurement age and for the Denver score (endline) rows, it reports the correlations between the endline Denver scores and measures of time to first mastery calculated using the tasks evaluated from the time of enrollment up to Denver endline measurement age. 3. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 5c: Correlation between Instability (Up to Midline/Endline Measurement Level) 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

Notes: 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 for the Denver score (endline) rows, it reports the correlations between the endline Denver scores and measures of instability calculated using the tasks evaluated from the time of enrollment up to Denver endline measurement age. 3. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 5d: Correlation between Time to Full Mastery (Up to Midline/Endline Measurement Level) 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

Notes: 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 for the Denver score (endline) rows, it reports the correlations between the endline Denver scores and measures of time to full mastery calculated using the tasks evaluated from the time of enrollment up to Denver endline measurement age. 3. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

the Denver tests are taken (i.e., the passing rate at that level). Tables C.1–C.4 in Appendix C report such correlations. The contemporaneous measures of knowledge are much more weakly correlated with the Denver scores. Cumulative measures are more predictive.

3.2.2 The Measures Capture Different Aspects of Knowledge

Table 6 shows the correlations between different measures of knowledge. While all the correlations are in the expected direction, different measures are far from perfectly correlated, suggesting that they capture different aspects of knowledge.[10]

[10] 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.

The correlations reported in Table 6 are weighted by the number of levels over which they are measured.

Average (across levels) time to first mastery is strongly negatively correlated with passing rates but much more weakly correlated with knowledge retention. Average instability (backsliding) across levels is at best weakly correlated with speed (time to mastery). Different measures of knowledge capture aspects of knowledge and learning. Time to first mastery, sometimes taken as a measure of IQ, is strongly correlated with the traditional measure based on average passing rates.

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. 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. 3. Instability is defined as the proportion of fails after the first success at each difficulty level by each skill type. 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

3.3 The Dimensionality of the Knowledge Measures

The strong correlations across some of measures analyzed in Table 6 suggest a possible one dimensional model of skill, although correlations of these measures with the Denver measures differ greatly. In this section, we examine how distinct these

measures are.

Appendix D reports estimates of principal components of the number of dimensions of these skills. A recurrent belief in the literature is that ability or skill is one-dimensional.[11] Human capital is often treated as a one-dimensional skill despite volumes of evidence against that assumption in the empirical literature (e.g., Heckman and Sedlacek, 1985). We examine this proposition within and across skills defined by our task measures. For the same skill, the dimensionality is multiple, although there is one dominant factor for each skill. Across skills, there are four and plausibly five dimensions. We present these results in Appendix D.

Our estimates of the dimensionality of these measures indicate that 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 valid summary of multiple knowledge is called into question.

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 these categories. We define ability by the speed of mastery of tasks (time to the first correct answer). As previously noted, it is conventional to measure ability by the learning speed, while knowledge is defined by eventual mastery of tasks.

Table 7 defines the categories. We experiment with other definitions and find similar results (see Appendix E). Figures E.1a–E.1d show that passing rates are

[11] This is what Willis and Rosen (1979) call the "one-factor" model.

persistent. Appendix Figures E.2a–E.2d and E.3a–E.3d show similar persistence in other measures of knowledge. The full mastery measure is quite noisy (see Figures E.4a–E.4d). The speed ability measure predicts the proportion of times that children get the wrong answer after the first correct answer (a measure of instability in performance) for all skills. See Figure E.3. Thus, within our survey, ability measures are persistent. Although they measure different aspects of knowledge, they capture traits that are not ephemeral. We next use these micro-based measures of knowledge to test the hypothesis of mean measured skill invariance.

Table 7: Ability Categories (Measured across All Levels)

Fast group	Pass the first task for more than (or equal) 80% of across all difficulty levels for each skill, and pass all skill-specific tasks at an average rate more than 80%.
Normal group	Pass the first task less than 80% across difficulty levels; the pass rate is greater than 50% within each level; 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%.

5 Testing Mean Measured Skill Invariance

In a fundamental paper, Agostinelli and Wiswall (2022) raise important questions about the existence of invariant measures of skill. They define mean measures of skill invariance. Formally, ***Mean Measured Skill Invariance*** (our terminology

but their idea) for a pair of ages a, a', 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 $a, a' \in [\underline{a}(\ell), \bar{a}(\ell)]$. The concept can be broadened for all skill levels.

This section conducts a test of the mean age invariance assumption for ages in the supports of our sample. Specifically, we examine whether Equation (3) holds: individuals with the same level of latent skills have, on average, the same test scores. We reject that hypothesis across all levels.

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

We use the weekly task performance based on the UHP protocols to define "true knowledge" at level ℓ for skill s as $K(s,\ell,a)$ for $a \in [\underline{a}(l), \bar{a}(l)]$. Recall that $K(s,\ell,a) \geq \bar{K}(s,\ell)$ is a binary measure of mastery *at level* ℓ and age a for skill s. We use the average passing rate. Averages are less sensitive to measurement errors. For traditional early childhood test measures, we use Denver scores as the measure of Z. As previously discussed, Denver tests are commonly used to measure early childhood skill development. (see Frankenburg and Dodds (1967) and Frankenburg et al.

(1971)) Rubio-Codina and Grantham-McGregor (2020); Rubio-Codina et al. (2016) show that the Denver test has better performance in evaluating the early childhood's skills among various evaluation tools used in different interventions across the world.

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

5.1 Finding Groups with Same ($\tau = K(s, \ell, a)$) but Different Ages (a) with levels

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, we array the data by quantiles of passing rates in the order of difficulty levels. 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 testing mean invariance at a and a' of skill s and level ℓ for Denver tests. Using the average passing rate at each difficulty level, we group children with similar passing rates in the same group. At difficulty level $\ell = 2$, children at the lowest quantile (τ_1) have the lowest passing rate (i.e., the passing rate is zero) and children at quantile 4 (τ_4) have the highest passing rate (i.e., the passing rate is 100%).

We then order child enrollment rates by 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 in the bottom 50% of the ages. The "old" group rank in the top 50% by age.

For example, the mean passing rate for the group of younger children in group 2 (τ_2) at difficulty level 3 is about 0.513, and the mean for the older group of children in group 2 (τ_2) 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 level of s, there are statistically significant age differences. For example, in group 2 (τ_2) at difficulty level 3, the mean age for the younger group is about 10 months, and the mean age for the older group is about 14 months.

In Appendix F, Tables F.1–F.4 show the partitions for higher levels of language skill. Tables F.5–F.9 show the comparable partitions for other skills across levels. 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 Hypothesis (3).

5.2 Testing Mean Measured Skill Invariance

We next systematically test the hypothesis that the Denver tests for 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 ages, levels, and skills. We present all test results in Appendix G. Tables G.1–G.2 report tests of whether the means of raw Denver language and cognitive scores are different (e.g., young vs. old) for each partition of τ at each difficulty level. The hypothesis of mean skill invariance is rejected. For raw Denver scores, the old group's performance at the same level of measured knowledge is consistently better than the

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

Level	Category	τ_1	τ_2	τ_3	τ_4
	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]
2	**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
	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]	
3	**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	

Notes: 1. Groups are categorized by passing rate for each skill by level. τ_1 is for children with the lowest passing rate, and τ_3 or τ_4 is for children with the highest passing rate. 2. Within each group, we sort children based on their monthly ages at the time of enrollment and generate two equal size subgroups named "Young" and "Old." Children whose enrollment ages are in the top 50% are categorized into the old group. 3. All measures in the table are evaluated from the time of enrollment to the Denver endline measurement age. 4.* When controlling for late entrants, we cannot reject the hypothesis. See Appendix J.

young group's performance; i.e., condition (3) is almost always violated. Therefore, 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. Other factors besides pure knowledge of s, defined by the ability to perform the same task, affect Denver tests. We report similar findings for cognitive and fine motor skill tests (see Tables G.3, G.4, and G.5).

We summarize the results for the tests of equality of Mean Differences in Denver Score $Z(s,a)$ conditional on τ groups by difficulty levels in Figures 4-6. For example, in Figure 4, we use different shading to indicate the p-values of the tests by difficulty levels at the given language τ group. The light gray color means that the tests reject with p-values less than 0.05, and the dark gray regions are with larger p-values.

Figure 4: Tests of the Mean Differences of Endline Raw Denver Language and Cognitive Score $Z(s,a)$ Conditional on Language τ Groups by Difficulty Levels

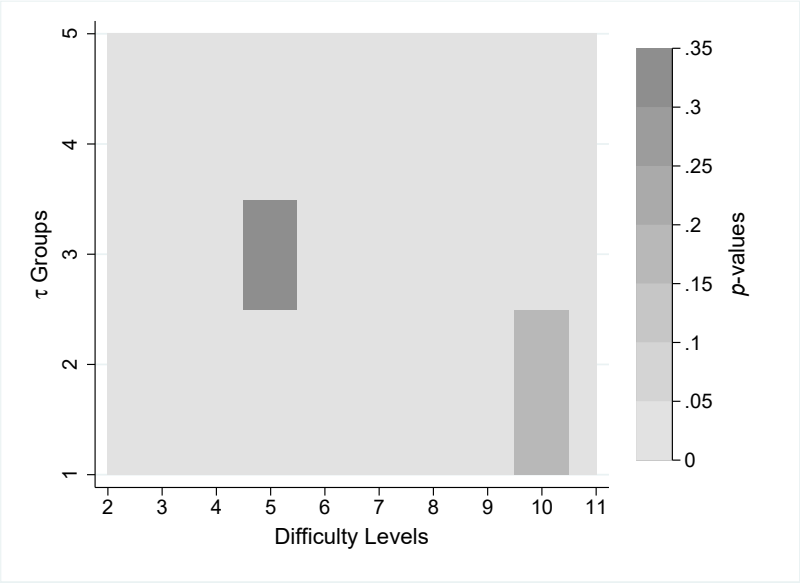


Figure 5: Tests of the Mean Differences of Endline Raw Denver Score $Z(s, a)$ Conditional on Cognitive τ Groups by Difficulty Levels

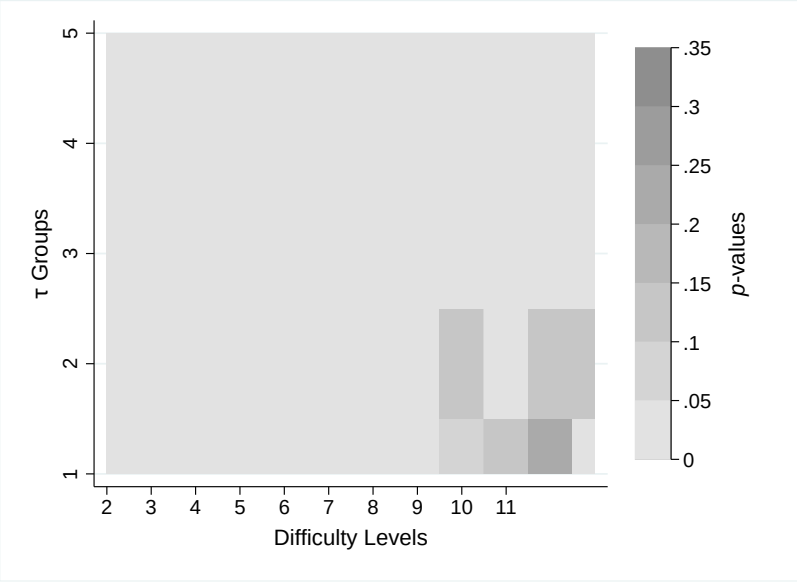
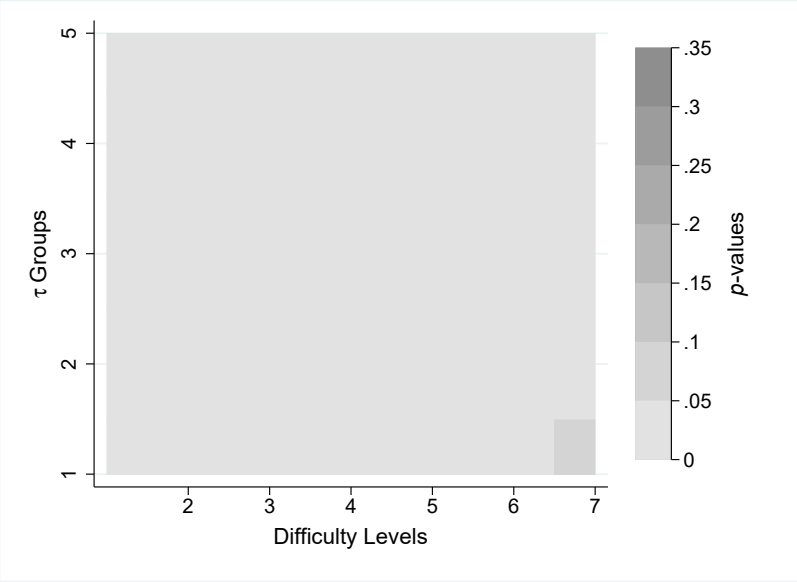


Figure 6: Tests of the Mean Differences of Endline Raw Denver Score $Z(s, a)$ Conditional on Fine Motor τ Groups by Difficulty Levels



We find that almost all the regions are in the light gray color, which means that for almost all tests we reject the null hypothesis that young age group has the same raw Denver scores as the old group.[12] We also report similar findings for cognitive and fine motor skills in Figures 5-6.

5.2.1 Up to Midline Measures

Appendix H 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 H.1–H.3 present tests of $K(s, \ell, a) = K(s, \ell, a') = \tau$ for 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: mean skill invariance condition is rejected.

5.3 Denver Language Test Results

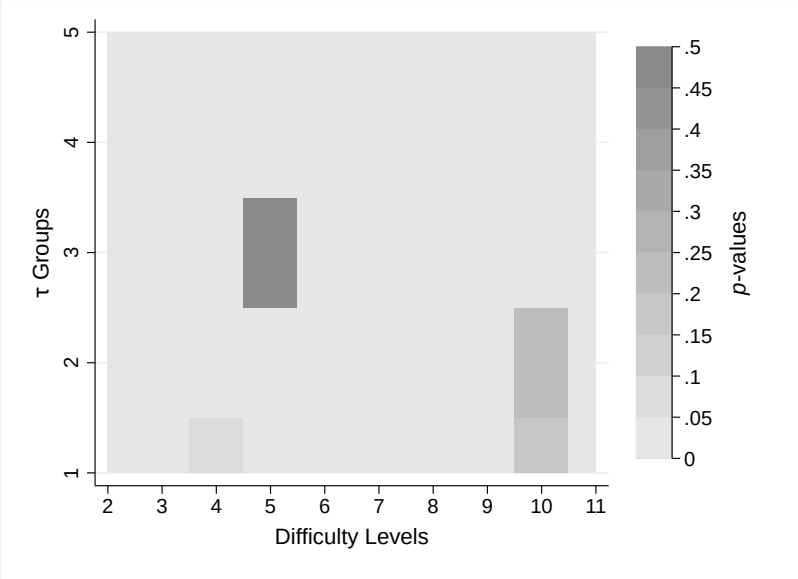
The preceding analysis reports tests of the hypothesis of Equation (3) using combined Denver language and cognitive tests. Scores are combined because there are few Denver test items for cognition. Our rejections for the Denver tests may be a consequence of combining conceptually distinct skills.

We conduct a similar series of tests using only language tests. These results are reported in Appendix I. Figure 7 summarizes the test results, and the results continue to reject the skill invariance assumption for language skill even after only considering

[12] Only three tests do not reject the null hypothesis that young age group has the same raw Denver scores as the old group.

the Denver language items. Details of the tests are presented in Tables I.1–I.2. There are too few cognitive tests to test the hypothesis for cognitive scores alone.

Figure 7: Tests of the Mean Differences of Endline Raw Denver Language Score $Z(s, a)$ Conditional on Language τ Groups by Difficulty Levels



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 tasks. Late entrants have fewer lesson experiences and may not be at the same level of knowledge due to dynamic complementarity of knowledge (see, e.g., Heckman and Zhou, 2022b). However, we condition on knowledge $K(s, \ell, a)$ attained, so this consideration should 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 J.

6 Conclusion

This paper examines the foundations of measurement of knowledge and learning. We use a rare data set for which we can measure weekly growth in the knowledge of a large sample of young children who execute identical tasks at identical ages for different skills. The data also contain measures on standard age-adapted, prototypical achievement test scores widely used in the educational, child development, and value added literature to compare across students, teachers, schools, and even entire countries to measure learning and evaluate intervention programs.

We test and reject a key assumption invoked in these literature: the existence of invariant measures of skill across different levels of tasks designed to measure the magnitudes of the same skill ("human capital") and to explore the growth of knowledge.

This paper shows that the standard measures used to chart student gains, child development, and the contribution of teachers and caregivers to student development are not comparable over ages and persons.[13] Conventional, widely-used measures, like PISA scores in (Organisation for Economic Co-operation and Development (OECD), 2014) that assume invariance are fragile and should be used with caution, if at all.

Our micro-based, alternative measures of knowledge also allow us to compare widely used passing rate measures with plausible alternatives: (a) speed at mastery tasks; (b) persistence in mastery; and (c) forgetting. The correlations among some

[13]Our results on the nonexistence of globally valid invariant scales are consistent with results obtained from the analysis of Heckman and Zhou (2022a).

of the alternatives are, at best, weak. These measures capture different notions of learning and knowledge. We find evidence of temporally persistent ability groups in terms of speed of mastery of tasks for skills across ascending levels of difficulty. There is persistence across levels of difficulty in all the measures we examine. Children separate into distinct ability groups for all the measures we investigate and they remain in those groups.

There are multiple measures of ability. At this stage of our study, we do not know which are predictive of school achievement, although in future work we will be able to do so. There are meaningful measures of learning and knowledge based on objective real work outcomes like education, wages and employment. Cunha et al. (2010) show that analysis using different anchors do not necessarily display common development patterns. But they are interpretable as comparable scales and warrant application.

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*.

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.

Ertem, I. O., V. Krishnamurthy, M. C. Mulaudzi, Y. Sguassero, H. Balta, O. Gulumser, B. Bilik, R. Srinivasan, B. Johnson, G. Gan, et al. (2018). Similarities and differences in child development from birth to age 3 years by sex and across four countries: A cross-sectional, observational study. *The Lancet Global Health 6*(3), e279–e291.

Fernald, L. C., E. Prado, P. Kariger, and A. Raikes (2017). A toolkit for measuring early childhood development in low and middle-income countries. *Strategic Impact Evaluation Fund, The World Bank*.

Frankenburg, W. K., B. W. Camp, and P. A. Van Natta (1971). Validity of the Denver developmental screening test. *Child Development*, 475–485.

Frankenburg, W. K. and J. B. Dodds (1967). The Denver developmental screening test. *The Journal of Pediatrics 71*(2), 181–191.

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

Grantham-McGregor, S., S. Walker, S. Chang, and C. Powell (1997). Effects of early childhood supplementation with and without stimulation on later development in stunted Jamaican children. *American Journal of Clinical Nutrition 66*(2), 247–253.

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

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

Heckman, J. J. and G. L. Sedlacek (1985, December). Heterogeneity, aggregation, and market wage functions: An empirical model of self-selection in the labor market. *Journal of Political Economy 93*(6), 1077–1125.

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.

Organisation for Economic Co-operation and Development (OECD) (2014, February). PISA 2012 results: What students know and can do - student performance in mathematics, reading and science. Technical report, PISA, OECD Publishing. Volume 1, Revised Edition.

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.

Ryu, S. H. and Y.-J. Sim (2019). The validity and reliability of DDST II and Bayley III in children with language development delay. *Neurology Asia 24*(4), 355–361.

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.

WHO Multicentre Growth Referece Study Group and M. Onis (2007). Assessment of Sex Differences and Heterogeneity in Motor Milestone Attainment among Populations in the WHO Multicentre Growth Reference Study: Assessment of Differences in Motor Development. *Acta Paediatrica 95*, 66–75.

Willis, R. J. and S. Rosen (1979, October). Education and self-selection. *Journal of Political Economy 87*(5, Part 2), S7–S36.

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.