NBER WORKING PAPER SERIES

MEASURING THE GROWTH OF SKILLS

James J. Heckman
Haihan Tian
Zijian Zhang
Jin Zhou

Working Paper 34737
http://www.nber.org/papers/w34737

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
January 2026

This research was supported in part by NIH grants NICHD R01HD107079 and R01HD103666, Hong Kong RGC ECS Grant 9048312, and the China Development Research Foundation. AEA registry number: AEARCTR-0007119. The content of this article is solely the responsibility of the authors and does not necessarily represent the official views of the funders. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Measuring the Growth of Skills
James J. Heckman, Haihan Tian, Zijian Zhang, and Jin Zhou
NBER Working Paper No. 34737
January 2026
JEL No. C18, J24

ABSTRACT

This paper discusses a fundamental problem in measuring the growth of knowledge and comparing the skills of people. New skills emerge that are not just more of the previously acquired skills. Psychometric convention forces these skills into arbitrarily constructed scales, which can severely distort measurement. To formally address this problem, we measure skills using a novel measurement scheme, estimate a stochastic learning process and reject the common scale assumption across levels for language and cognitive skills. Furthermore, we estimate dynamic complementarity without imposing arbitrary scales for skills.

James J. Heckman
University of Chicago
and IZA
and also NBER
jjh@uchicago.edu

Haihan Tian
The University of Chicago
hhtian@uchicago.edu

Zijian Zhang
Columbia University
zz3263@columbia.edu

Jin Zhou
City University of Hong Kong
jzhou269@cityu.edu.hk

A randomized controlled trials registry entry is available at https://www.socialscienceregistry.org/trials/7119
A Replication data in the Harvard Dataverse is available at https://doi.org/10.7910/DVN/GQBPPO

I Introduction

Economists use test scores to measure learning, treating scores as cardinal and commensurable across people and ages. This practice is untenable. Test scores are at best ordinal, and different normalizations generate different rankings and treatment effects.[1] Arbitrarily scaled scores are widely used in empirical work on child development and education.

This paper makes a new point: skills do not evolve as higher levels of a fixed set of abilities. New skills emerge with development, and they are often qualitatively different from the ones measured at earlier ages. There is no constant-unit measure of cognition or language that is valid across ages for the same people. Forcing distinct skills into a common scale distorts measures of both levels and growth. This insight can explain key anomalies in the literature, including the apparent fadeout of early treatment effects.

Using weekly data from a large-scale home-visiting program in rural China, we measure skills precisely within age-specific intervals. This novel measurement scheme allows scale-free tests of dynamic complementarity, learning, cross-complementarity, and catch-up without relying on psychometric fiat (Heckman et al., 2025).

II The Technology of Skill Formation

The technology of skill formation is a multistage technology for which a *vector* of skills at age $a+1(\boldsymbol{K}(a+1))$ depends on previous skills and investment $I(a)$:

[1] See Jacob and Rothstein (2016); Cunha et al. (2021), and Freyberger (2025).

$$\boldsymbol{K}(a+1) = \boldsymbol{f}^{(a)}(\boldsymbol{K}(a), \boldsymbol{I}(a)) \tag{1}$$

where vector $\boldsymbol{I}(a)$ is investment at age a broadly defined (parenting, schools, etc.). $\boldsymbol{K}(a+1)$ and $\boldsymbol{K}(a)$ might be measured on a standard scale as is common in the literature, but that is not required.

Dynamic complementarity is the notion that investment at age a raises the productivity (rate of learning) of later investment (i.e., $\boldsymbol{I}(a+j)$, $j > 0$);

$$\frac{\partial^2 \boldsymbol{K}(a+j+1)}{\partial \boldsymbol{I}(a) \partial \boldsymbol{I}'(a+j)} > 0. \tag{2}$$

This parameter does not require any assumption that $\boldsymbol{K}$ at different ages is measured in the same units.

III Our Strategy

Key to our strategy is access to repeated lessons on skills at specific levels administered to people of the same age. The outcomes (learned or not) of each sequence of identical lessons can be measured. They are well-defined measures of skill comparable across otherwise identical people exposed to the same sequence of lessons and can meaningfully measure growth within levels. These measures need not cumulate additively across levels, as they would if a constant-unit assumption across sequences of lessons were appropriate. This evidence is consistent with more formal tests reported in Heckman and Zhou (2026).

Children enter the program we study at different ages. They face a structured schedule of repeated instruction and assessments of age-specific skills that applies to *all* participants of the same age, irrespective of their previous experience in the program. We examine whether the learning rate of skills is accelerated by higher levels of skills present at the start of each stage (level) of instruction. Dynamic complementarity enables low-ability children to partially close the gap with high- and normal-ability children.

One solution to the problem of lack of a uniform scale across levels of a skill is to carry over items across administrations of the assessments across levels, and fix the scale on previously tested items. However, Heckman et al. (2025) show that this solution is problematic. When the *same* item is tested on occasions separated by a year, children do *better* on the second occasion. There is *appreciation* of knowledge with the passage of time. The scale shifts upward on the second test. The gain in knowledge is consistent with the complementarity of inputs received after the first set of lessons and the consolidation of knowledge documented in neuroscience.[2]

IV Our Data

Our data come from a large scale, experimental evaluation of *China REACH* conducted by the China Development Research Foundation (CDRF). For details, see Zhou et al. (2026) and Heckman et al. (2025).

The intervention cultivates multi-dimensional skill development through home-visiting. Trained home visitors at the same level of education as the mothers of the children studied, visit each treated household weekly, and provide one hour of age-

[2]See Dehaene (2020).

specific caregiving guidance to caregivers. Children are evaluated weekly on their knowledge. Zhou et al. (2026) show that the intervention significantly improves skill development (e.g., language and cognitive, fine motor, and social-emotional skills).

The program teaches *caregivers* to interact with their children. Tasks in three different skill categories (fine motor, language, and cognitive) are taught each week. Skills taught are ordered by difficulty levels following the profiles developed by Palmer (1971) and Uzgiris and Hunt (1975), henceforth UHP, which are widely used in the literature on child development. We use scales of skills that describe levels of knowledge with content that are **the same** within each level and across all children of the same age at that level.

There are eleven UHP difficulty levels for language skills. Language skill difficulty levels are based on the concepts shown in Table 1. The language skill tasks increase in difficulty with the expectation that the child will learn to identify and use expressive language to indicate understanding.

Standard psychometric practice imposes a common numerical scale across these levels and constructs a cardinal scale of "language comprehension" that is claimed to be comparable across levels. The activities at higher levels are assumed to produce knowledge that is more of the same thing measured at lower levels.[3]

[3]See, e.g., Lord and Novick (1968) for a discussion of the artifice of creating test scores.

Table 1: Overview of Language Task Child Should Perform (Learning Words)

Level 1	Caregiver and child make sounds to each other to interact
Level 2	Caregiver tells child the things she does in the house
Level 3	The child recognizes people's names
Level 4	The child learns movements that show intimacy: clapping, bye-bye, and thank you
Level 5	Caregiver and child look at the pictures together, and let the child vocalize and touch the pictures
Level 6	The child recognizes at least one body part
Level 7	The child identifies and/or names ordinary objects
Level 8	The child points to the pictures which are being named, names one or more pictures, mimic the sound of the objects
Level 9	The child points to the pictures which are being named, names two or more pictures, mimic the sound of the objects
Level 10	The child points at 7 or more than 7 pictures and talk about them
Level 11	Teach the child some simple descriptive words and the child names objects at home, and tells the usage of those objects

Within levels, the skills taught and measured are essentially identical. For example, Table 2 presents detailed information about the five tasks (and assessments) in difficulty level 3 directed to children 9-18 months old. All tasks relate to the activity of teaching the child to recognize people's names and are essentially identical. There is no hierarchy of tasks assigned *within* levels.

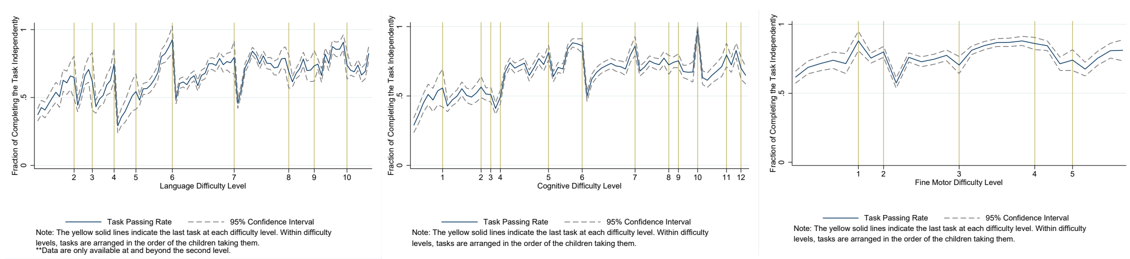
Table 2: Language Task Content (Learning Words) Level Three

Difficulty Level	Difficulty Level Aim	Month	Week	Learning Materials	Task Aim and Content
Level 3	To teach child to recognize people's names	9	2	Language – correct nouns:	To teach child to recognize people's names: Mother teaches child the names of family members
Level 3	To teach child to recognize people's names	11	4	Language – correct nouns:	To teach child to recognize people's names
Level 3	To teach child to recognize people's names	14	2	Language – correct nouns:	To teach child to recognize people's names: Mother teaches child to say the names of the family members
Level 3	To teach child to recognize people's names	16	3	Language – correct nouns:	To teach child to recognize people's names: Mother teaches the child to identify the names of family members.
Level 3	To teach child to recognize people's names	18	1	Language – correct nouns:	To teach child to recognize people's names: Mother teaches the child the names of family members.

For each item at each level of each skill, we measure mastery of the item at each age. We compute passing rates by age, by level, and by skill as a measure of learning. Figure 1 plots the growth of knowledge in language, cognitive, and fine motor skills across levels. Average (across people) passing rates by age within each difficulty level for language and cognitive tasks increase with the number of lessons, a pattern consistent with learning. When individuals transition to higher difficulty levels of nominally the same skill, initial age-specific passing rates decline at entry into the next level, consistent with the notion that new skills are being taught at each level. After initial declines, age-specific passing rates within levels increase as learning of new skills ensues. At most levels of fine motor skills, there is—at best—modest learning.

Figure 1: Average Task Passing Rates by Order and Level

(a) Language (b) Cognitive (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. Source: See primary data and the plots in Heckman and Zhou (2026).

We also examine the rate of learning by ability level. High ability children have the highest performance across levels, normal the second, and low the third. Normal ability children catch up at later skill levels. At later levels of instruction, there is strong evidence that low performers at earlier stages can catch up in learning. The distinction between ability groups is sharp at earlier levels but gradually narrows as children progress across levels.

Using the dynamic reinforcement learning model developed in Heckman and Zhou (2026), which dynamizes the IRT model to allow for stochastic skill growth, and level-specific shocks, we reject the hypothesis of a common language skill scale across levels. The same holds for cognitive skills, but not fine motor skills.

This evidence calls into question a standard tool in the economics of education. Value-added measures assume that psychometric scales measure the same skill at different levels with higher levels being more of the same thing measured at lower levels. So do long-term evaluations of skill formation programs. Without a common scale assumption across levels of nominally the same skill, the empirical literature rests on shaky ground.

Fadeout, much discussed in the evaluation literature (see, e.g., Bailey et al., 2020), is one possible consequence of the emergence of new skills. Comparisons of different skills over time are not meaningful. Heckman and Zhou's (2026) estimates of a stochastic process for learning show a pattern of rises and declines within levels. Variances of shocks differ across levels. This produces a pattern of fadeout, recovery, and fadeout.

V A Scale-Free Test of Dynamic Complementarity

In our treatment sample, almost all children receive lessons starting from level 6, but heterogeneity in exposures prior to level 6 gives considerable variability in earlier investment. Roughly 40% take the entire curriculum. The others enter at later stages of the curriculum ladder.

We separate treatment group children by their enrollment status. One group receives the full curriculum, and another group enrolls at level 6 with no prior investment. We survive balance tests on baseline background variables between two cohorts with different program exposure.

Table 3 compares children's rate of learning using the number of trials they take to achieve their first success at the same UHP level. At UHP Language levels 8, 9, and 11, children with full exposure to the curriculum require significantly fewer trials until their first success within each level. Otherwise, similar children with more early investment are learning more than those without early investment. Children raised by grandparents are slower to learn.

Table 3: Time to First Mastery by Enrollment Cohort (Language)

	UHP Language Levels ℓ				
$d(s,\ell)$	$\ell = 7$	$\ell = 8$	$\ell = 9$	$\ell = 10$	$\ell = 11$
Full Exposure of Curriculum	2.487	2.079	1.496	1.404	1.348
Enrolled at Language Level 6	2.518	2.367	1.686	1.454	1.698
p-value	0.870	**0.024**	**0.017**	0.608	**0.000**
stepdown p-value	0.864	**0.059**	**0.059**	0.827	**0.000**

Notes: 1. This table presents the learning speed within the same UHP Language level by enrollment cohorts. Time to first mastery is the number of trials a child takes until the first success at each difficulty level during the intervention for each skill type. 2. See Heckman et al. (2025) for further detail.

Results are comparable for cognitive skills. There is no evidence of dynamic complementarity for fine motor skills. For initially low-ability groups in both cohorts, the difference in learning rates at later levels is even more pronounced (Heckman et al., 2025). Disadvantaged children benefit the most in the early stages through dynamic complementarity and reduce the gap at later stages with higher ability children and children in better social learning environments.

VI Summary

Using unusually fine granular data on learning, we document the emergence of new skills at different levels of nominally the same skill. This calls into question current procedures that impose common scales on the nominally the same skills used to assess growth in skills and compute value-added measures. Comparing the incomparable can also explain fadeout.

The lack of a common scale across levels of nominally the same skill does not prevent analysts from identifying crucial aspects of the technology of skill formation if meaningful scales can be found *within* levels. We use such scales to identify dynamic complementarity and catch-up of the less able. Careful design of future studies enables analysts to avoid imposing arbitrary scales on the data and computing meaningless value-added measures.

References

Bailey, D. H., G. J. Duncan, F. Cunha, B. R. Foorman, and D. S. Yeager (2020). Persistence and fade-out of educational-intervention effects: Mechanisms and potential solutions. *Psychological Science in the Public Interest 21*(2), 55–97.

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.

Dehaene, S. (2020). *How We Learn: The New Science of Education and the Brain*. Penguin UK.

Freyberger, J. (2025). Normalizations and misspecification in skill formation models. Forthcoming, *The Review of Economic Studies*.

Heckman, J. J., H. Tian, Z. Zhang, and J. Zhou (2025). Dynamic complementarity. Under Revision, *American Economic Journal: Applied Economics*.

Heckman, J. J. and J. Zhou (2026, January). The microdynamics of early childhood learning. Forthcoming, *The Journal of Political Economy*.

Jacob, B. and J. Rothstein (2016). The measurement of student ability in modern assessment systems. *Journal of Economic Perspectives 30*(3), 85–108.

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.

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

Zhou, J., J. J. Heckman, B. Liu, and M. Lu (2026). The impacts of a prototypical home visiting program on child skills. *Journal of Labor Economics*.