"At Three Years of Age, We Can See a Child's Future" Online Supplementary Materials.

Link 1 (as referenced in the introduction):

In order to track the lifecycle of cognitive development over different age groups, we used data from four different papers. While the data from these studies comes from different times in the life cycle and from different parts of China, all are randomly selected and all come from rural areas from Central and Western China. The exact list of papers, their data sources, and study areas can be found in Table 1S. This supplement will describe more in detail the sampling methods, cognitive measures, and math score measures in these studies and therefore our study.

Sample selection for each age cohort

Data used to analyze the development of cognition among infants and toddlers—that is, at the first developmental stage—come from 5 projects which each included surveys of infants and toddlers. These data were combined into a single data set to create a larger sample comprised of 3,353 infant/toddler and caregiver pairs. The data, all randomly drawn, come from four different rural sub-populations (western China rural communities, central China rural communities, migrant communities, and resettlement communities) across four different provinces (Shaanxi, Hebei, Henan, Yunnan, and Guizhou). Survey teams also measured cognition in samples of rural households in migrant communities in the urban areas of Beijing, Zhengzhou, and Xi'an.

Although data come from five different project surveys, the protocols for sampling the respondent-families were similar. First, poor counties were randomly chosen from the four provinces. Then, after excluding the county seat townships (as these townships tend to be relatively richer and urban), single townships were randomly selected. Villages within townships were also randomly selected. Birth records were then obtained from officials in the village or township. These records were used to identify all household with infants/toddlers 6-30 months old. The sample generated and used by this study cover sub-populations that account for approximately 69% of all Chinese rural infants and toddlers.

The young children in the second age cohort (ages 4 and 5) are from Southern Shaanxi province. The respondents are in fact the same children as were sampled and surveyed when the respondents were infants/toddlers in the Shaanxi dataset described in the paragraph above. In other words, the 4 to 5 year olds in this study are part of a cohort that the research team followed over at three year period, which consisted of 512 children. The key to putting the sample together for this part of the study were the activities undertaken to track the young children who at that time were almost all (above 90%) in preschools. Some preschools were in villages; some were in towns. After a one-month tracking period, the survey teams accounted for 92.6% of the original infant/toddler sample.

To collect data on primary school children, as a way to analyze the relationship between cognition and academic success of children in rural elementary schools, we chose a sample of school-aged children in Anhui and Henan provinces. The same sampling procedure was followed in each of the provinces. We first chose a set of five counties in Anhui and five counties in Henan. After excluding the schools in the county seat (as most students were urban residents), we randomly chose six schools per county. Once inside the schools, we focused on two groups of students: students whose parents lived and worked in the countryside and students whose parents worked as migrants. Taken together, these children are referred to as rural students in the remainder of the paper. The research team was able to survey successfully a total of 6,390 rural primary school students in 60 rural primary schools.

To analyze the relationship between cognition and academic performance of youth in middle school, we collected panel data on seventh graders in two prefectures across two provinces in Northwest China. Both these provinces are relatively poor, with per capita GDP ranked in the bottom half of China's provinces. Surveyed schools were selected by randomly selecting schools out of all the junior high schools in sample areas, excluding those located in the county seat township (as these schools are relatively more rich and urban) or those that had too few seventh graders (twenty students or fewer). Individual students were selected by first randomly selecting one grade seven class then randomly selecting half the students in a class for inclusion. Three waves of data were collected, with the first wave occurring when students just entered seventh grade, the second wave occurring at the end of the academic year, and a third wave at the start of the following academic year. In the end, this study was able to survey 3,012 students in 160 different schools.

Cognitive Outcome Measures

While we have measures for cognition for all of our age cohorts, each uses a slightly different metric. To measure the cognition (and other skills—language; socio-emotional skills; motor skills) for the infant/toddler group, the research team used the Bayley Scales of Infant and Development (BSID-III), which is an internationally recognized and reputable test for measuring infant cognitive development. The scale was implemented in the respondents home by trained enumerators who directly interact with the infant/toddler, however, the caregiver (typically the mother or paternal grandmother) is present. The cognitive scale, the main variable of interest, assesses information processing, counting and numeric skills. BSID-III scores for each category are compiled, then compared to scores taken from reference populations, which are the exact some age in days). For cognition, our main variable of interest, a mean score for a normal population is expected to be 105 (“Are infant/toddler delays a problem across rural China”, Wang et al., 2019). Cognitive delay is defined as any infant/toddler that scores below -1 SD (or below an 85) (“Are infant/toddler delays a problem across rural China”, Wang et al., 2019; Zhao et al., 2019).

To measure the level of cognitive development of the young children of preschool age, the research team used the fourth edition of the Weschler Preschool and Primary Scale of Intelligence (WPPSI-IV). The WPPSI was developed as an IQ test, but many of its questions also measure executive function, which is a similarly important measure for cognitive development. Because of this, the cognition scale includes tasks that measures skills relative to five specific categories: verbal comprehension, visual-spatial competence, fluid reasoning, working memory, and processing speed. Unlike the BSID scale, which was given to infants/toddlers with the main caregiver present, enumerators complete the WPPSI with the young children in a quiet one-on-one environment—typically at home or at preschool (in a quiet room).

The Raven Standard Progressive metrics test (henceforth referred to as the “Raven test” or “Raven’s”) was used to measure cognitive development in older children and youths. The Raven test is a non-verbal intelligence test that uses pictorial questions to test spatial reasoning and pattern making skills. Scores on the Raven test are then calculated to generate an IQ score based on established norms. The Raven test is one of the most widely used tests in the world, and has previously been used in China.

In addition to the Raven test, the research team also used the Fourth Edition of the Wechsler Intelligence Scale for Children (WISC) to measure cognition in rural middle schoolers.

WISC and WPPSI-IV were both designed by Weschler, and are thus extremely similar: the main difference is that WISC is for older children (6-16 year olds) while WPPSI-IV is designed for preschoolers. By using both Raven's and WISC to collect cognition data for youth, these results gathered by the research team become more robust. However, it is important to note that sample sizes are different: while Raven's was administered to every student (resulting in 2,507 observations), WISC IQ tests were only administered to three randomly chosen students from each sampled class, thus only collecting 472 observations.

Math Score Measures

Math test scores are used to measure levels of academic achievement in older children and youths. For older children, all sampled third and fourth grade students were administered a standardized math test. Different versions of this test were administered to third and fourth graders. Test questions were chosen from the Trends in International Mathematics and Science Study (TIMSS) test data bank and were tailored to be appropriate for the local third and fourth grades, respectively. Youth academic achievement data was collected by administering a 35-minute math test to surveyed seventh graders. Test questions were selected according to local curriculum standards and validated by local education experts and psychometricians. Enumerators proctored test-taking students to ensure testing rules were followed. For all children who took math tests, raw test scores were standardized to create a standardized measure for academic achievement.

Link 2 (as referenced in the *Persistence of Cognitive Delays* subsection in the Results):

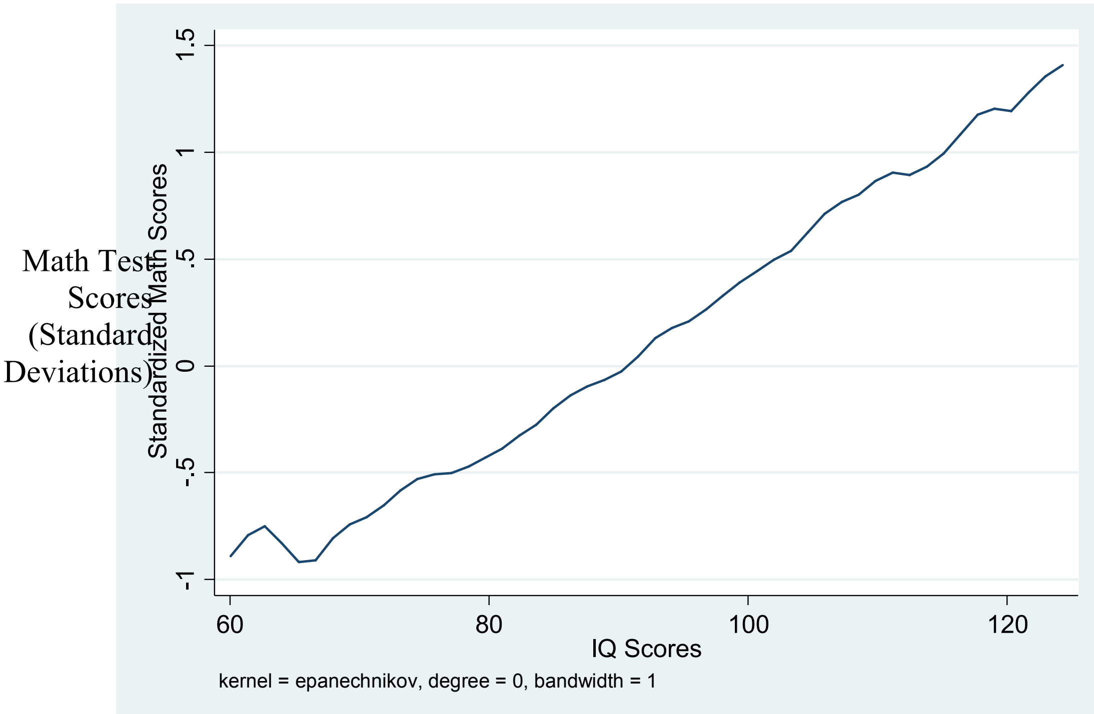
Because the findings for infants/toddlers and young children are both based on the same cohort of children that are followed over this three to four year period, it is possible to demonstrate actual persistence of cognitive delay (and cognitive improvement) during this period of the life cycle ("The Persistence and Fade-out Paradox", Wang et al., 2019). According to their findings (that use the exact same sample that we have used), there is persistence over time. To show this the authors of Wang et al. (2019) show how they divided their sample of infants/toddlers into two groups with nearly identical levels of cognitive development. They then implemented an intervention to one group that was successful in raising the average level of cognitive skills of the infants/toddlers in the intervention group relative to those in the control group (made up of those individuals that did not get the intervention). When the authors compare the level of working memory (a key component of healthy cognitive development) between the young children (who are now 4 to 5 years old) in the intervention and control groups, they find working memory levels are higher in the intervention group than in the control. Clearly, these results in a causal manner that if cognition is higher when children are in the infant/toddler stage of development, the cognition of the group remains high when they move into the next phase of the life cycle (as a young child, 4 to 5 years old). In this way, the data shows that cognitive delays persist without intervention.

Link 3 (as referenced in the *Cognition and Academic Performance* subsection in the Results):

The striking correlations between IQ and academic achievement can be seem in Figures 1S, 2S, and 3S. Figure 1S illustrates the relationship between cognition and academic achievement for primary school students—in this graph, we can see that an increase of 20 IQ points from 80 points to 100 points results in an almost 1 SD increase in math scores (Figure 1S). In Figures 2S and 3S, we see this same strong connection for junior high students. Figure 2S shows the relationship between cognition (as measured using WISC-IQ scores) and math scores over two waves: the blue line shows the relationship at the start of 7th grade, and the red line shows the relationship at the end of 7th grade. As in Figure 1S, we see that IQ and academic achievement are highly correlated: the difference in achievement between a student with 80 IQ points and 100 IQ points is almost 0.5 SDs (Figure 2S). This relationship is also seen in Figure 3S, in which Raven's IQ scores are used to track changes in academic achievement across all three waves. Like in Figures 1S and 2S, we see that IQ and academic achievement are closely linked: the lower the IQ of a student, the worse their performance (Figure 3S).

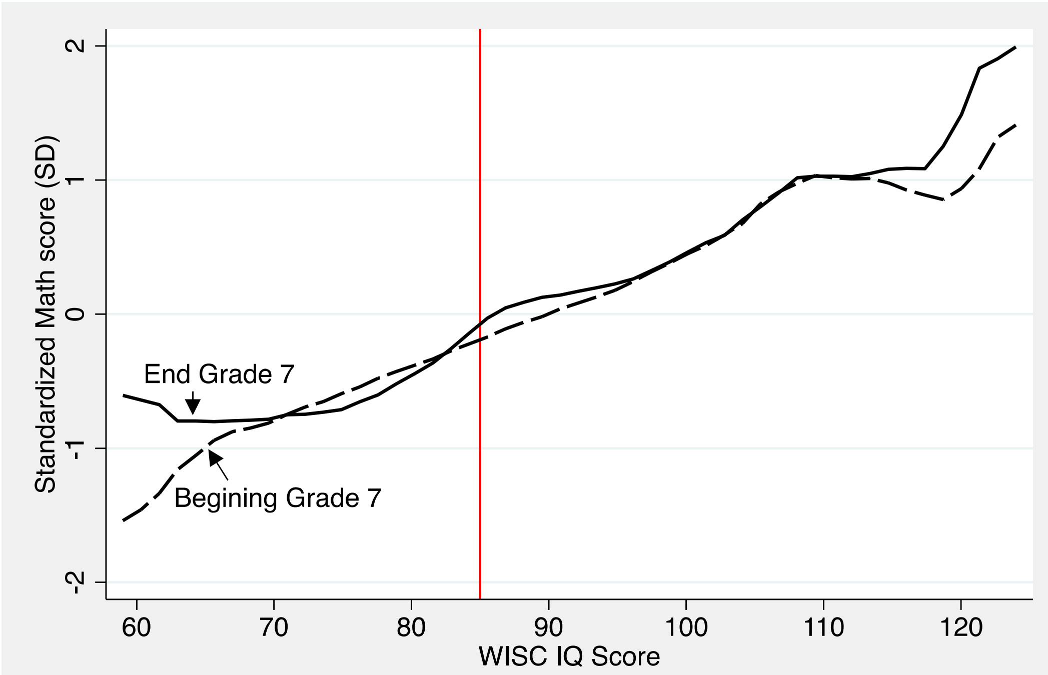
Table 1S: Summary of cohorts and papers used in this study

Paper Source	Survey Time	Sample Size	Sample Population	Age Range	Sample Area	Cognitive Outcome Measures	Author	Published Year
The Asia-Pacific Journal | Japan Focus 17(16)(3),1-16.	2015-16; 2017;	3,353	Infants and toddlers	6-30 months old	Shaanxi; Hebei, Yunnan; Guizhou; Henan; Beijing; Zhengzhou; Xi'an, Shaanxi;	The third edition of the Bayley Scales (Bayley-III)	Wang et al.	2019
(Rural Education Action Project Working Paper) Stanford CA: Stanford University.	2017	517	Young children	4-5 years old	Southern Shaanxi province	The Weschler Preschool and Primary Scale of Intelligence (WPPSI).	Wang et al.	2019
(Rural Education Action Project Working Paper) China Economic Review (55), 199-217.	2017	6,309	Primary School students in 3rd and 4th grade	10-11 years old	Anhui and Henan provinces	The Raven Intelligence Quality (IQ) scale	Zhao et al.	2019
(Rural Education Action Program Working Paper). Stanford, CA: Stanford University.	2016, 2017	3,012	Middle School Students	13-14 years old	Two provinces in Northwest China	The Wechsler Intelligence Scale for Children (WISC IQ) and Raven's Standard Progressive Matrices (Raven IQ or SPM-IQ)	He et al.	2019



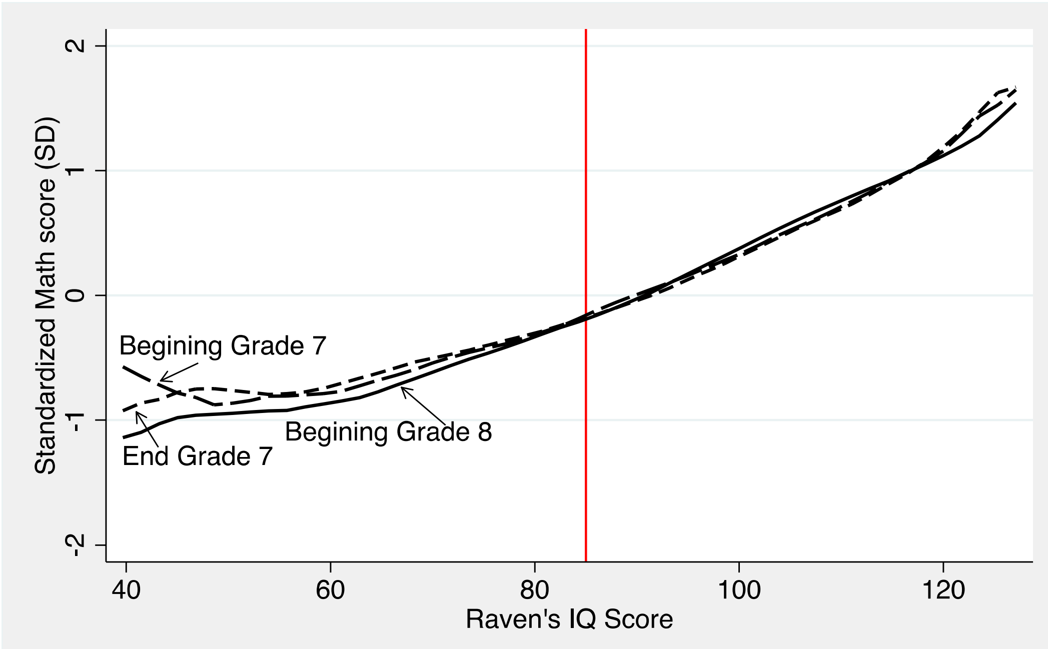
Data Source: Author's survey from Zhao et al. 2019 “Better Cognition, Better School Performance: Evidence from Primary Schools in China.”

Figure 1S: The Graphical Relationship Between IQ and Academic Achievement in Primary School Children



Data Source: He et al, 2019 "IQ, Grit, and Academic Achievement: Evidence from Rural China"

Figure 2S: The Graphical Relationship Between WISC IQ Score and Standardized Math Test Scores for Junior High Students Across Two Waves



Data Source: He et al, 2019 “IQ, Grit, and Academic Achievement: Evidence from Rural China”

Figure 3S: The Graphical Relationship Between Raven IQ and Standardized Math Test Scores For Junior High Students Across Three Waves

Works Cited

He, Xinyue, Fang Chang, Huan Wang, Sarah-Eve Dill, Scott Rozelle, and Prashant Loyalka. 2019. "IQ, Grit, and Academic Achievement: Evidence from Rural China." Working Paper. Stanford, CA: Stanford University.

Wang, Lei, Wilson Liang, Siqi Zhang, Laura Jonsson, Mengjie Li, Cordelia Yu, Yonglei Sun, et al. 2019. "Are Infant/Toddler Developmental Delays a Problem across Rural China?" *Journal of Comparative Economics* 47 (2): 458–69. https://doi.org/10.1016/j.jce.2019.02.003.

Wang, Lei, Yiwei Qian, Scott Rozelle, Chuyu Song, Nele Warrinnier, and Sean Sylvia. 2019. "The Persistence and Fade-out Paradox: New Evidence on Medium Run Effects of an Early Childhood Development Intervention." Working Paper. Stanford, CA: Stanford University.

Zhao, Qiran, Xiaobing Wang, and Scott Rozelle. 2019. "Better Cognition, Better School Performance? Evidence from Primary Schools in China." *China Economic Review* 55 (June): 199–217. https://doi.org/10.1016/j.chieco.2019.04.005.