Unpacking the Black Box of Cognitive Ability
A novel tool for assessment in a population based survey

Amar Hamoudi
Duke University

Margaret Sheridan
University of North Carolina, Chapel Hill

November 2015

This manuscript is part of a wider project undertaken in partnership with the Young Lives Study. We gratefully acknowledge the invaluable contributions of our colleagues and collaborators with the Young Lives, Peru project—Sofya Krutikova, Alan Sánchez, Santiago Cueto, Guido Melendez, and Mary Penny, as well as the entire field staff at GRADE, Peru and the Young Lives program leadership at Oxford University. This manuscript has also benefited from discussions with our co-panelists and participants at Session 5 (“Cognition and Demography”) at the 2015 Annual Meeting of the Population Association of America, and with workshop participants at the population centers of McGill University, University of Maryland, University of Wisconsin-Madison, the University of Pennsylvania, Duke, and UNC. Any remaining errors are ours.

*Sanford School of Public Policy, Box 90312, Durham, NC 27708-0413. 919.613.9343. amar.hamoudi@duke.edu
†Department of Psychology & Neuroscience, Davie Hall 248, 235 E. Cameron Ave, Chapel Hill, NC 27599-3720. 919.962.5365. sheridan.margaret@unc.edu
Abstract

Academic achievement is an important form of human capital, but non-academic skills and capabilities are also of central to individual, family, and community well-being. In this study, we complement tools and models from population science with tools and models from cognitive sciences in order to explicate and measure population-level variation in cognitive skills that are knowledge-domain-general— that is, their relevance is not restricted to any single circumscribed topic area like reading or mathematics. We developed a novel application for tablet PCs that builds in the key elements of standard measurement tools used in cognitive neuroscience and cognitive psychology. The application assesses four discrete domain-general cognitive skills: implicit learning, which is the capacity to learn without conscious awareness; declarative memory, which is the capacity to encode, retain, and retrieve information that is not an object of continuous attention; inhibitory control, which is the capacity to override counterproductive impulses and resist distraction by irrelevant information; and working memory, which is the capacity to hold in mind and manipulate information that is an object of continuous attention. The assessment does not rely on respondents having had any prior exposure to specific information, techniques, or algorithms. To prove the concept, we deployed it in a population-based sample of Peruvian adolescents in the context of the Young Lives, Peru study. Our findings indicate that domain-general cognitive skill measurement can be done effectively and at manageable cost within broad population-based socioeconomic surveys, even if respondents have minimal prior exposure to the technology used in the testing.
Academic achievement is an important form of human capital, but non-academic skills and capabilities are also central to individual, family, and community well-being. Accordingly, capabilities like IQ, fluid intelligence, organization, self-control, perseverance, and socioemotional skills have had a high and rising profile in empirical social science research (Kimball, 2015; McArdle and Willis, 2012; Smith et al., 2010; Duncan et al., 2007; Heckman, 2007; Bowles et al., 2001). Explicating the expansive array of skills and capabilities comprising everything other than knowledge acquired through formal schooling has become an important objective in the empirical social sciences. In this study, we combine tools and models from the population sciences with those from cognitive sciences in order to contribute to that goal. We focus on a subset of nonacademic capabilities that are cognitive, but knowledge domain general. They are cognitive because they enable individuals to accomplish tasks that require sustained mental effort, and domain general because they are not specific to any circumscribed topic area. An example of such a skill is inhibitory control, which is the capacity to override counterproductive impulses and to resist distraction by irrelevant information. Research in psychology and neuroscience has investigated how discrete skills like these complement and determine composite measures that are now standard in the empirical social sciences, like fluid intelligence, crystallized intelligence, or overall IQ.1

Measurement challenges are a significant barrier to population research on domain-general cognitive skills. As we will discuss, there has been substantial progress in incorporating measurement of broad, composite constructs like fluid intelligence into population surveys. That progress has opened avenues for population scientists to understand factors affecting and affected by non-academic capabilities. Similarly, there is great potential in building measurement of more specific skills like inhibitory control into population research. For example, the skills we measured in this study are defined using a taxonomy that is borrowed from the cognitive sciences, which is based in part on physiology (that is, skills are grouped together in part on the basis of the brain regions on which they rely). Population scientists interested in understanding how environmental exposures affect cognitive development or cognitive aging may benefit from tools to separately measure and analyze skills with separate physiological underpinnings. This paper introduces such a novel measurement tool.

Cognitive scientists measure domain-general cognitive skills using specialized computer applications, which are presented in the form of games. Each game challenges respondents to complete some task that specifically requires the skill being assessed. Computer technology is powerful in this context because the performance measures require precise recording of respondents’ behavior as they engage with the test. Two other design characteristics are critical to the effectiveness of these tests. First, assessments are designed to be as domain-general as possible. Because the skills being assessed apply across multiple domains of knowledge, it is

1For reviews and examples, see Blair (2006); Nisbett et al. (2012), and the references cited therein.
important that performance not be influenced by prior familiarity with any information, techniques, or algorithms. Capabilities like literacy, numeracy, or language should not affect how a respondent performs; for example, if numerals or letters are used in any of the tasks, then familiarity with those stimuli may affect performance, obscuring measurement of the true underlying cognitive skill. Therefore, cognitive scientists build very general paradigms into the assessments, including for example using only symbolic cues and prompts. Second, they are designed to be as skill-specific as possible. If a test is designed to assess inhibitory control, then performance should not be influenced by other domain-general skills and capabilities, like memory or openness to engaging with new technology. Cognitive scientists achieve this design objective by building two separate types of tasks into the same test. One type (“baseline”) is affected by statistical noise factors like comfort with the testing environment, ability to understand instructions, and so on. The second type (“challenge”) is designed to capture all the same statistical noise factors, plus the specific skill being tested. The presence of the baseline provides an analyst with information that can be used to sweep out the noise, and extract the signal about the relevant skill.

It would expensive and impracticable to directly transpose the standard testing applications from cognitive science research into population studies, because the applications require specialized equipment and pose substantial respondent burden. Therefore, we developed a novel application for tablet PCs that builds in the key elements of these tools: computer technology, domain-general paradigms, and a baseline/challenge design to achieve skill-specificity. Our assessment application is cheap, portable, requires no specialized training to administer, and poses manageable respondent burden. As such, it is feasible to include in a broader population-based survey. To prove the concept, we deployed it in a population-based sample of adolescents in the context of the Young Lives, Peru Survey.

Our findings indicate that even though most of the study sample was unfamiliar with touchscreen tablet computers, our assessment tool captured population-level variation in four separate domain-general cognitive skills. A fundamental condition expected of these assessments is that respondents perform more poorly on challenge trials than on baseline trials. We observe that pattern in the study sample as a whole, and in every subgroup that we examine. Evidence from cognitive science research gives reason to believe a priori that young children have lower skill levels than adolescents; we confirm that our measurement tool establishes that pattern in our sample. One challenge to interpreting the patterns that we observe is that performance in these assessments is reported in units that are not straightforward to interpret terms of outcomes with which population scientists are usually concerned. In order to provide a basis to benchmark our measures, and also as a further validity check, we examine how our skill measures relate to schooling progress in the study sample. One of the skills that we chose to measure was selected because there is no a priori reason to expect it to be related to school progress; we observe a precisely estimated zero relationship between measured levels.
of that skill and school progress. The other skills are related to school progress, in keeping with a priori expectation.

These results imply that our novel measurement tool can support research bridging gaps between the population sciences and psychology, in order to allow population scientists to tap into the growing literature on relationships between neural function and individual, family, and community characteristics. It can also facilitate taking cognitive psychology research out of the lab, allowing psychologists to benefit from the unique features of large population-based studies in order to better understand cognitive function at the level of the population.

1 Background: Domain General & Domain Specific Skills

We developed tools to measure a set of cognitive skills that are not specific to any single circumscribed topic area, or “knowledge domain.” Classifying skills on the basis of how specifically they apply to circumscribed knowledge domains is common in psychology and education (see, for example, Tricot and Sweller (2014) and references cited therein). An example of a more domain specific skill might be arithmetic, which applies to the circumscribed domain of mathematics. By contrast, cognitive inhibition is more domain general. Cognitive inhibition is the capacity to control attention or behavior, and override counterproductive impulses. A person might use this skill to focus on solving a problem on a mathematics test, but it can also be used in many other contexts that are unrelated to mathematics. Domain specific skills require prior exposure to information, techniques, or concepts. Literacy skills require prior exposure to the symbolic meanings of letters, the definitions of words, and the rules of grammar; arithmetic skills require prior exposure to symbolic meanings of numerals and operators, and basic algorithms. By contrast, the development of more domain-general skills like cognitive inhibition does not rely on prior exposure to any particular set of information or techniques. This does not mean that these capabilities cannot be shaped or trained; many, if not all, do improve with training and practice (Diamond et al., 2007).

Economists have recognized the human capital role of skills which are specific to circumscribed knowledge domains at least since the seminal work of T.W. Schultz, who structured and organized modern human capital theory by citing the central role that “the acquisition of knowledge and skills through schooling” could play in alleviating poverty (Schultz, 1980). A contemporaneous discussion among sociologists was focused on distinguishing the extent to which schooling drives labor market outcomes by imparting domain-specific knowledge, versus by acculturating young people to their future work environment (Bowles et al., 2001; Bowles and Gintis, 2002; Farkas, 2003). Across the social sciences, there is a long tradition of empirical
investigations into sources of and returns to domain specific skills like reading or mathematics. Some of these analyses rely on skill measures involving administrative data on performance on school subject tests (Weinberger, 2014, for example) or standardized tests administered by the government (Crawford et al., 2014; Lundborg et al., 2014; Figlio et al., 2014, for example). Others rely on standardized subject tests administered in the context of a broader survey, like the Armed Forces Qualification Test (AFQT) (Aughinbaugh and Rothstein, 2015; Hall and Farkas, 2011; Farkas et al., 1997), the Peabody Individual Achievement Test (Guo and Harris, 2000), the Inter-American Reading and Comprehension Test (Behrman et al., 2014), or the Program for International Student Assessment (PISA) tests (Evans et al., 2014; Jerrim and Micklewright, 2014). Several large-scale publicly available population surveys support this line of research by administering subject tests to their respondents— for example, the Child Development Supplement to the Panel Study on Income Dynamics (PSID-CDS) includes the Woodcock-Johnson Assessment, which measures reading and math skills, among others; the National Longitudinal Survey of Youth, 1979 cohort (NLSY79) included an administration of the AFQT; and the Early Childhood Longitudinal Studies (ECLS-B and ECLS-K) include reading and mathematics tests.

Complementing the literature on domain-specific skills as human capital, over the last half-century economists and sociologists have expanded traditional theories and models regarding capabilities that matter for long-run success. These expanded theories and models pay more attention to characteristics that apply across a wider array of contexts and knowledge domains— including IQ, organization, self-control, perseverance, and socioemotional skills (Kimball, 2015; Bowles et al., 2001; Duncan et al., 2007; Heckman, 2007). Empirical research into some of these models and theories has relied on assessments of skills using components of standardized tools like the Wechsler Preschool and Primary Scale of Intelligence (WPPSI) (Duncan et al., 1994; Brooks-Gunn et al., 1993, 1996), the Stanford-Binet Intelligence Scale (Brooks-Gunn et al., 1996), or the Bayley Scales of Infant Development (Attanasio et al., 2015). Each of these instruments involves asking respondents to answer questions or to complete tasks that are cognitively demanding, but that do not require much specific topical knowledge. For example, one part of the WPPSI requires respondent children to look at several pictures, and identify a subset that has a common characteristic. To perform well on this part of the WPPSI, a child must have well-developed reasoning capabilities. The task is not meant to distinguish children in any way based on whether or not they have previously been exposed to any specific information, techniques, or algorithms. Nonetheless, these tasks are not perfectly domain-general. For example, a typical question might involve showing a respondent a set of pictures including a pencil and a crayon, in order to test whether they can identify those two as having a common characteristic (they are both writing implements). Accomplishing this task does not require literacy (since all the prompts are symbolic), but it does require prior exposure to pencils and crayons, and familiarity with how those two objects are used in many cultures.
Measurement of abstract reasoning skills (which is general across many domains of knowledge) will be confounded by any variation in prior knowledge about pencils and crayons; the confounding is ignorable only as long the test is administered on a sample in which everyone has had equal exposure to this prior knowledge.

Other approaches to domain-general skill measurement involve asking survey respondents to complete tasks which require the types of skills that researchers are trying to assess. For example, the Raven's Progressive Matrices task requires respondents to infer a rule linking one picture to the next in a series, and then extend that inference to identify the picture that would come next in the series (Raven, 1958); population scientists have used it to assess fluid intelligence, which is closely related to domain general cognitive capabilities (Behrman et al., 2014; Beuermann et al., 2015). Similarly, the Forward/Backward Digit Span task involves asking respondents to listen to a string of numbers being read to them, and then repeat the string back either in the order in which it was originally presented, or in reverse order. Social scientists have used that task to assess short term memory capacity (Crawford et al., 2014). The Health and Retirement Study includes a series of tasks like these. In many of these tasks, domain generality is likely to be imperfect. For example, a Forward/Backward Digit Span task is likely to be easier for respondents who are more familiar with the names and meanings of numerals.

Skills that are specific to circumscribed knowledge domains can be coherently characterized on the basis of those domains. Reading skills, arithmetic skills, linguistic skills, and so on can be modeled, measured, and analyzed as separate coherent constructs. Knowledge-general skills require an alternative taxonomic approach. One such approach aggregates many discrete skills and capacities in very broad global constructs like IQ or fluid intelligence. Traditions in cognitive neuroscience and developmental psychology recommend more fine grained distinctions, so that global constructs like IQ are modeled as comprising many constituent skills. In this approach, skills are defined in part on the basis physiology– they are grouped together or distinguished on the basis of the brain regions on which they rely. Skills are also defined in part on the basis of their function– they are grouped together or distinguished on the basis of the roles they play in behavior. We adopted the same taxonomy to identify four distinct domain general skills, which we measured for this project:

1. **Implicit learning** skill is the ability to learn without conscious awareness. This skill is sometimes described as “muscle memory.” When a person learns how to ride a bicycle, they learn a complex set of fine motor responses to maintain an even distribution of their weight around the bicycle frame. This learning can be described as occurring “without conscious awareness” since the person would be unable to explicitly describe all of these responses, or the cues that guide how and when they must be deployed. Implicit learning is a basic skill which develops very early in life, and is conserved in
many species with simpler brains than humans (Squire, 2004; Reber, 1989; Packard and Knowlton, 2002). This form of learning is critical to very basic skill acquisition, including for example an infant’s learning to recognize language sounds (Saffran et al., 1999). Given that the skill and the brain regions on which it depends develop so early in life, it is likely that implicit learning will be less sensitive to environmental conditions than the other skills we measured. Thus, this skill was included in the assessment tool because we had reason to believe a priori that it would be less systematically variable across population subgroups, and so could make a valuable placebo test. If we were to observe subgroup variation in this skill of similar scale to that of the other skills we measured, we would have reason to be concerned that our assessment tool is not appropriately sensitive to the specific skills we were targeting.

2. **Declarative memory** skill is the ability to encode, retain, and retrieve information that is not an object of continuous attention. For example, declarative memory supports an individual’s capacities to learn new facts, techniques, or algorithms, and to build the database of information derived from experience by which they make sense of their environment. Declarative memory skill, and the brain regions on which it depends, develop later in life than implicit learning (Ofen et al., 2007; Buckner, 2003; Ghetti et al., 2010; Gogtay et al., 2006). Furthermore, the cognitive neuroscience literature contains substantial evidence that those brain regions are particularly vulnerable to negative effects from exposure to chronic stress in the environment (Lupien et al., 2009); this implies that disadvantaged subgroups, who may be exposed to more chronic stress, may be expected to have poorer declarative memory skill.

3. **Inhibitory control** is the ability to override counterproductive impulses and resist distraction by irrelevant information. For example, the children’s game of “Simon says” challenges players’ inhibitory control abilities. Inhibitory control may play an important role in self-control.

4. **Working memory** skill is the ability to hold in mind and manipulate information that is an object of continuous attention. For example, an individual would use working memory when working through a multistep algorithm, in order keep track of each step in the process and update plans of action while proceeding.

Working memory and inhibitory control are key elements of **executive function**, which is a set of skills that are critical for controlling behavior, and ensuring that higher level abstract goals are not supplanted by lower-level, more immediate goals. For example, an individual would use executive function to attend to a conversation (a higher-level goal) even despite being hungry (a lower-level goal) (Miyake and Friedman, 2012). Executive function supports the formulation and execution of complex plans and the orientation of
behavior toward a goal (Diamond, 2013). These are among the latest developing cognitive skills, and the brain regions on which they rely may continue to develop well into the second decade of life (Luna et al., 2010; Best and Miller, 2010; Gogtay et al., 2004; Sowell et al., 2001; Aron et al., 2014; D’Esposito et al., 1995).

In cognitive neuroscience, the principle of “developmental plasticity” implies that skills are most responsive to environmental conditions during the time when the neural systems on which they rely are growing and developing (Nelson and Sheridan, 2011; Fox et al., 2010; Wiesel and Hubel, 1965a,b). These periods of growth and development span nearly three decades for inhibition and working memory, and more than a decade for declarative memory; therefore, for example, socioeconomic variation in these skills may be especially pronounced. In addition, declarative memory, inhibitory control, and working memory represent a logical starting point for domain-general measurement because they play central roles in many hypotheses about human capital, as we discuss in the next section.

2 Motivation: Cognitive Human Capital

Domain-general cognitive skills like declarative memory and executive function are components of intelligence and are crucial to effective decision-making, and as such they fit neatly into an economic model of human capital (Kimball, 2015). Our measures are designed to improve population scientists’ capacity to empirically analyze the human capital roles that these skills may play. Like other forms of human and non-human capital, evidence indicates that these skills bear two characteristics which are of central interest to empirical social scientists. First, they tend to endure over time, although their stock can be increased through the investment of resources or depleted as a result of adverse exposure. Second, they produce a flow of returns in the form of outcomes that individuals have reason to value. Together, these characteristics imply that cognitive human capital, like many other forms of capital, supports a cycle of value.

Experimental evidence in psychology, neuroscience, and child development indicates that some domain-general skills, including especially executive function, grow in response to investments of time and effort on the part of parents and teachers (Holmes et al., 2009; Klingberg, 2010; Diamond et al., 2007). Furthermore, there is observational evidence that stressful, challenging, or deprived conditions may impede these skills’ development and hasten their decay (Lupien et al., 2009; Hackman et al., 2010; Noble et al., 2005; Nelson and Sheridan, 2011; Brooks-Gunn et al., 1996; Mullainathan and Shafir, 2013; Shonkoff and Garner, 2012; McLaughlin et al., 2014; Sheridan and McLaughlin, 2014; Sheridan et al., 2012; Evans and Schamberg, 2009; Sheridan et al., 2013). This evidence would imply that family and community resources, if invested to
enrich learning environments and insulate individuals from excessive stress, can be used to expand the stock of at least some forms of cognitive human capital.

There is also evidence that cognitive human capital generates returns. Canonical models in economics propose that some forms of human capital, like facility with reading or mathematics, generate a flow of resources in the form of increased wages. For example, workers who are facile in reading or arithmetic can be more productive. This is likely also true for domain-general skills like declarative memory. A business owner who can encode, retain, and recall more information than their competitors may have a competitive advantage in the marketplace, and an employee who can retain and recall more complex instructions may be able to command a higher wage. However, the flow of resources generated by domain-general skills is likely to be broader than just increases in labor productivity. For example, improving declarative memory makes it easier to learn basic skills like reading and to retain facts. In this way, declarative memory functions as an input to domain-specific knowledge, which in turn is another form of capital. The flow of resources produced by executive functioning skills are likely to be especially broad, since these skills support the creation and execution of complex future oriented plans, which is a fundamental and essential capability. In convenience samples, measured executive functioning skills have been observed to correlate with initial school readiness, academic success, reduced risk of alcohol and drug abuse in adolescence and adulthood, and reduced risk of incarceration for crimes involving impulsive behavior (Mischel, 2014; Diamond, 2013; Duncan et al., 2007).

Domain-general cognitive abilities, like other forms of capital, support a cycle of value if they improve in response to investment and if they generate a flow of resources. The capital stock today (a person’s executive functioning capability) generates a flow of resources (benefits of effective planning). Some of those resources can be reinvested (more effective parenting) to expand the capital stock tomorrow (better executive functioning capabilities in offspring). This produces a bigger flow of resources in the future, some of which can be reinvested, and so on. This cyclical character, in turn, may have important implications for inequality, and for the intergenerational transmission of poverty.

There is likely to be value in distinguishing among discrete domain-general skills. Different skills may respond differently to investment or to environmental conditions. For example, the physiology of declarative memory is distinct from the physiology of executive function; sensitive periods in their development do not perfectly overlap, so negative exposures at some points in the life-course may be more harmful for one than for the other. School curricula or parental behaviors that facilitate the development of executive function may not be as productive of declarative memory, and vice-versa. Specific types of negative exposures may be more harmful to some domain-general skills than others.

These dynamics motivate many hypotheses involving potential human capital roles played by discrete domain-
general cognitive skills. Are there specific behaviors which are especially productive for each separate skill? What kinds of early life experiences shape skill development? How do these effects vary by skill? In what respects can different domain-general cognitive skills complement or substitute for domain-specific knowledge in producing socioeconomic well-being? In what respects can one specific cognitive skill complement or substitute for another? Questions like these are of central interest to empirical social scientists as well as cognitive scientists. Because of the complex real-world interdependencies involved, there are likely to be important contributions that could be made using empirical approaches that are central to the population sciences. These include population representative sampling, long follow-up, comparable measurements collected on multiple family members, and techniques that generate or exploit exogenous sources of variation. Cost-effective direct measurement of the relevant skills in population surveys will enhance the power of all of these standard demographic tools. Our assessment tool was designed to facilitate that measurement. We describe the specifics of the design in the next section.

3 Measurement Approach

3.1 Sample: Young Lives, Peru Study

We measured domain-general cognitive skills in a population-based sample of Peruvian children born in 2001 and 2002, in the context of the fourth wave of the Young Lives, Peru survey. The broader Young Lives study tracks over 12,000 children in four developing countries: Peru, Vietnam, Ethiopia and India. Respondents include the Young Lives children themselves, a coresident sibling, a key adult informant from the child’s household, and community leaders.

Table 1 describes our analytical sample for this study. Column 1 describes the original Young Lives, Peru sample, which comprises 2052 children who were born between December 2000 and June 2002. At the time of the first wave of the survey, these children were aged 6-12 months. The sample was representative of the communities in which respondents lived; the communities themselves were chosen purposively to span Peru’s geographic, ethnic, and socioeconomic diversity. The sample was first interviewed in 2002/2003, and has been reinterviewed every three years since. The fourth wave was conducted from February to December 2013. A total of 91% of the original sample (1869 children) was successfully reinterviewed. Further analysis (results not shown) indicates that loss to followup was more likely among rural and more asset poor respondents than among urban and less asset poor respondents.2 Of the 1869 children who were successfully

2More information on the Young Lives study, including more detailed analyses of sample composition and sample attrition, is available at http://discover.ukdataservice.ac.uk/series/?sn=2000060.
interviewed for round 4, all but 12 completed our domain general cognitive assessment; the dozen failures were due to technical glitches in the tablet PC-based assessment tool. The median age of original sample respondents at the time of assessment was 11 years and 8 months.

In 2009/2010, as part of the third wave of the study, Young Lives expanded its sample to include not only the original cohort of children described in column 1 of table 1, but also each sample child’s next younger coresident sibling, if they had one. A total of 1045 next-younger coresident siblings were identified and added to the sample. All younger siblings older than 5 years at the time of the wave 4 interview were also targeted to complete the domain general cognitive assessment. We would expect older children to be more skilled than younger children, on average, on each of the dimensions that we assessed. To evaluate the plausibility of our assessment tool, we tested whether it uncovered these expected skill gradients by comparing the main sample to those younger siblings who were born between March 2005 and September 2006 (aged 6-8 years at the time of assessment). We excluded siblings younger than this group because they would have been too young to engage with the assessment tool. We excluded siblings older than this group in order to maintain age separation between the “main cohort” and “younger sibling” groups that is wide enough that we would expect measurable developmental differences in cognitive skill. Column 2 of table 1 describes the younger sibling sample. Of the 455 age-eligible younger siblings identified in round 3, 398 (87%) were successfully reinterviewed in the fourth wave. Two did not complete the cognitive assessment due to technical glitches with the assessment tool.

In addition to the cognitive assessments, respondents and their coresident adult caregivers were interviewed on a variety of domains. The interview included a module on household socioeconomic characteristics that was modeled on World Bank Living Standards Measurement Surveys. Other modules covered children’s experiences, including school attendance and attainment. The survey also tracks the schooling attainment of all household members, and involves anthropometric measurement of children and their coresident parents (Sánchez et al., 2015).

3.2 Measurement of Knowledge-Domain-General Cognitive Skills

In general, a skill assessment is presented as a game consisting of multiple rounds, which cognitive psychologists call “trials.” Each of the four skills is measured in its own game. In each game, there are two types of trials, “baseline” and “challenge.” A respondent’s performance on a baseline trial depends on many skills and competencies, including for example familiarity with touchscreen tablet computers or computer games. The game is designed so that challenge trials require exactly the same set of skills and competencies,
plus the specific skill being assessed. During the course of the game, a respondent will encounter baseline trials and challenge trials, but will not be told about any distinction. This design allows an analyst to compare challenge trial performance against baseline performance to isolate variation in specific skills. In our analyses, we measure skills using the difference in within-individual mean performance on challenge trials versus baseline trials. We have also measured them using the difference in median performance; results are substantively the same. More generally, the assessment tool we built reports all the raw data that can be used to calculate within-individual distributions of key performance measures. Therefore, in other applications it would also be possible to use other quantiles or distributional statistics, according to the specific application and the judgement of the analyst.

Depending on the specific game/assessment, key performance measures may involve precise measures of the time taken by respondents to complete some task (down to the level of $10^{-5}$ seconds), or precise measures of the location where respondents touched the screen (measured in screen pixels). As a result, skill measures are in units like milliseconds or screen pixels, which are difficult to quantify in terms of outcomes that population scientists are familiar with. In order to benchmark magnitudes, we compare average skill levels for older (11-12 year old) respondents in the main Young Lives cohort against younger (6-8 year old) children in the sibling sample. As an alternative benchmarking exercise, we also compare children on the basis of their progress through school. This allows us to translate units like milliseconds in more intuitive terms like years of age or years of schooling.

For each game, we estimate differences between baseline and challenge in a regression context. In computing estimation error around our key parameters, we use the “clustered sandwich” estimator of the variance/covariance matrix (Arellano, 1987; Liang and Zeger, 1986; White, 2001) to account for the fact that each game involves multiple observations of each respondent. We have also used a bootstrap estimator, resampling respondents; results are substantively identical.

The total respondent burden for the four games averaged about 20 minutes. Below, we describe each of our four games and the specifics of our analytical approach.

**Measuring Implicit Learning**

In figure 1a we show a schematic representation of our implicit learning game. It is based on the “Serial Reaction Time” paradigm, which is standard in cognitive science (Ruitenberg et al., 2015; Lum et al., 2013; Nissen and Bullemer, 1987). Our version is presented to respondents as a game called “chase the dot.” A total of 175 dots are presented one at a time in rapid succession, and the respondent is asked to touch each
dot as quickly as possible after it appears. Each presentation of a dot is a trial. The dot appears in one of four locations on the screen, and disappears after 1 second, or as soon as the respondent touches it (whichever comes first). Trials are divided into 4 blocks. Block $A$ comprises the first 35 trials; block $B_1$, the next 70; block $A'$, the next 35; and finally, block $B_2$ comprises the last 35 trials.

Baseline trials are those in blocks $A$ and $A'$. In these blocks, the dots move among the four screen locations in a non-patterned sequence. The challenge trials are in blocks $B_1$ and $B_2$. During these blocks the dots move in a repeating cycle. Specifically, the first dot appears at screen location 2, the next at location 4, then 2, and so on, in a repeating “2-4-2-4-1-1-4” pattern. The game moves seamlessly from block to block, so to respondents it looks like one continuous activity. A respondent would be unlikely to consciously note the “2-4-2-4-1-1-4” pattern in the challenge blocks because the trials succeed each other so quickly and the pattern is reasonably long. Respondents would, however, be able to learn the pattern through unconscious “muscle memory,” or implicit learning.

The key performance measure in this game is a respondent’s reaction time—elapsed time from the appearance of dot $d$ until respondent $i$ touches it ($t_{id}$). Figure 1b illustrates how average performance varies over the course of the game. The vertical axis is the number of elapsed milliseconds between the appearance of the dot and the respondent’s touch. Each point in the graph is average response time among the 1857 main cohort respondents (first column of table 1). The solid line represents a locally smoothed nonparametric regression of response time against trial number. The dashed lines represent block-wide averages, and the shaded areas represent 95% confidence intervals around those averages. Overall, respondents went quicker as the task went on, even during the first few trials of block $A$, during which there was no pattern. However, they accelerated steeply during block $B_1$, as they became unconsciously acquainted with the repeating “2-4-2-4-1-1-4” pattern. When the pattern was removed (block $A'$), they decelerated because they could no longer rely on muscle memory. When the pattern was re-established (block $B_2$), they were able to accelerate again.

The speed advantage that a respondent acquires from having implicitly learned the pattern reflects their capacity to learn without conscious awareness. We assess that speed advantage based on the difference in a respondent’s mean response time on challenge trials versus baseline trials. For our preferred measure of implicit learning ability, we focus on differences in average response times between block $A'$ and block $B_2$, which occur after respondents have had time to learn the pattern. We also report differences between all challenge trials (blocks $B_1$ and $B_2$) and all baseline trials (blocks $A$ and $A'$).

We estimate the difference by regressing response time against an indicator identifying the challenge trials ($C_d$). We also include an individual respondent fixed effect ($\alpha_i$) and a dot location fixed effect ($\lambda_d$), as well as flexible controls for accuracy ($a_{id}$) in order to capture variation in the degree of care that respondents may
have taken in playing the game. We compute accuracy using the Euclidean distance between the Cartesian coordinates of the dot and the location where the respondent touched.

$$t_{id} = \beta_0 + \beta_1 C_d + \alpha_i + \lambda_d + f(a_{id}) + \varepsilon_{id}$$ (1)

In this regression equation, \(\beta_1\) represents average implicit learning ability in the sample.

In order to induce “muscle memory,” it is important to put respondents under perceptible time pressure, which is achieved by having the dot disappear quickly even if the respondent is not able to touch it in time. As a result, \(t_{id}\) will not be observed if it is longer than 1000ms; when this occurs, respondent \(i\) is said to have “timed out” of trial \(d\). In our main cohort, 11% of trials in block \(A'\) and 9% in block \(B_2\) were timed out. In estimating and interpreting equation (1), we used five alternative approaches to account for the truncation of response times. All five approaches give substantively similar results. Our preferred approach is illustrated in figure 1c. In each block, we fit a normal distribution to have the same median and the same quantile at 1 second as the observed truncated distribution. We then impute to each timed out trial a random draw from the portion of the fitted distribution that lies to the right of 1 second. Figure 1c shows the observed cumulative distributions of response times in each block; the dotted portions beyond 1 second are imputed.

**Measuring Declarative Memory**

Our declarative memory assessment/game was divided in two parts, bookending the other tasks. So, it was the first game that each respondent started, and also the final one that they completed. The design was based on the “Paired Associates Learning” paradigm, which is well-known in cognitive science (Gabrieli, 1993; Hannula et al., 2006).

We grouped 12 shapes arbitrarily into 6 pairs. The paired associate for shape “A” would be shape “1”; for shape “B,” shape “2,” and so on. We use these labels for the purposes of discussion here only; the shapes were not labeled in any way for the respondent. The game is divided into 24 trials. In each trial, at the top of the screen, a respondent sees one shape, and across the bottom they see four options, one of which is the “associate” of the shape at the top. An example is shown in figure 2a. The respondent is asked to complete the pair by touching the correct paired associate at the bottom of the screen. When they choose the correct option, the shape they touch moves up to the top of the screen beside its paired associate, a box is drawn around the pair, and the pair dances in a short animation sequence.

Each pair is presented to each respondent four times. For example, in the first trial they might see shape “A,” and guess among the options until they correctly identify shape “1.” Then after the box is drawn around the
“A”/“1” pair and the animation sequence is finished, the next trial will start and they will see shape “B” and will search among the options until they find shape “2,” and so on. After several trials they will see shape “1” at the top of the screen, and shape “A” along with three other options at the bottom. If they have excellent declarative memory, they will immediately recognize “A” as the associate of “1,” and touch it without having to search. Therefore, the first appearance of each pair is a baseline trial, and subsequent appearances are challenge trials. Performing better on the challenge trials than on the baseline trials requires declarative memory skill.

In comparing baseline to challenge trials, there are two alternative performance measures that may be analyzed. Our preferred measure \( \pi_{ipa} \) is binary, equal to 1 if, upon being presented with pair \( p \) for the \( a \)th time, respondent \( i \) correctly completed the pair with their first choice, and equal to 0 otherwise. Figure 2b illustrates that accuracy by this measure improves with increasing exposure to pairs, reflecting respondents’ declarative memory abilities. A second measure would be the total number of choices they made before arriving at the correct paired associate \( n_{ipa} \). This second measure is bounded from below by 0, and is censored from above by 3 since any respondent who made more than 3 incorrect guesses must have selected at least one incorrect shape more than once. This truncation and censoring complicates analyses using the second measure. In this sample, 2.2% of trials involved more than 3 incorrect guesses, and 5 trials (out of more than 22,000) involved 10 or more incorrect guesses. We also report results of analyses based on this secondary measure.

We estimate differences in performance between baseline and challenge trials by regressing our performance measure \( \pi_{ipa} \) or alternatively \( n_{ipa} \) against indicators for challenge trials \( C_{pa} = 1 \) for the \( a \)th appearance of pair \( p \), and 0 for all other appearances). We include an individual respondent fixed effect \( (\alpha_i) \) and a shape-pair fixed effect \( (\gamma_p) \).

\[
\pi_{ipa} = \beta_0 + \sum_{a=2}^{4} \beta_a C_{pa} + \alpha_i + \gamma_p + \varepsilon_{ipa} \tag{2}
\]

\[
n_{ipa} = \beta_0 + \sum_{a=2}^{4} \beta_a C_{pa} + \alpha_i + \gamma_p + \eta_{ipa} \tag{3}
\]

The coefficients \( \beta_a \) in each regression represent measures of average long term memory skill in the sample. We report estimates of \( \beta_4 \) only, in the interest of succinctness.
Measuring Inhibitory Control

In figure 3a we show a schematic representation of our inhibitory control assessment/game, which is based on the “Simon Task,” a well known game in cognitive (Simon and Rudell, 1967; O’Leary and Barber, 1993). In each trial, a respondent is presented with a dot on either the left or the right side of the screen. On the baseline trials, the presented dot is solid, which is the cue to simply touch the dot, as close as possible to its center. On the challenge trials, the presented dot is striped, which is the cue to touch the opposite side of the screen, as close as possible to where the center of the dot’s mirror image would be. The dot disappears after 2.5 seconds, or as soon as the respondent touches the screen (whichever comes first).

When each new dot appears, it draws the respondent’s attention, and so their first impulse is to direct their touch to the dot itself. On the baseline trials, they simply follow that impulse. By contrast, on the challenge trials they must control their behavior and override that impulse, redirecting their touch toward the opposite side of the screen. Therefore, differences in performance between baseline and challenge trials reflect a respondent’s inhibitory control ability. There are two coequal measures of performance. The first is a respondent’s reaction time–elapsed time from the appearance of dot d until respondent i touches it (t_{id}). Figure 3b illustrates that the distribution of response times is shifted right on challenge trials, relative to baseline trials, reflecting the burden that inhibitory control imposes on respondents. The second performance measure is accuracy. We measure respondent i’s accuracy in trial d (a_{id}) by comparing the screen column where the respondent touched to the correct location for that trial. On baseline trials, the correct location is the screen column of the center of the presented dot; on challenge trials, it is the screen column where the center of the mirror image would be. Figure 3c indicates that respondents are much more likely to err specifically toward the center of the screen on challenge trials, compared with baseline trials. This is likely because of failure to perfectly override the impulse to touch the presented dot.

Response time and accuracy are likely to be jointly determined. For example, a respondent who responds as quickly on challenge trials as on baseline trials may be paying little attention to the stimuli, and therefore responding inaccurately. Figure 3d represents a flexible regression of accuracy against response times. On average, in both types of trials the fastest and slowest responses are the least accurate. We address the problem of joint determination by conditioning each performance measure on the other. Specifically, when our performance measure is accuracy (a_{id}), we de-mean response times for challenge trials (C_{d} = 1) and baseline trials (C_{d} = 0) separately, and estimate differences in performance between baseline and challenge trials with the following regression:

\[ a_{id} = \beta_0 + \beta_1 C_{d} + \alpha_i + f(t_{id}, C_{d}) + \varepsilon_{ipa} \] (4)
The \( f(\cdot) \) function is a flexible condition-specific spline, with knots at the 10th, 25th, 50th, 75th, and 90th percentiles.

Similarly, when our performance measure is response time \((t_{id})\), we de-mean accuracy measures for each trial type separately, and estimate differences in performance between baseline and challenge trials with the same flexible condition-specific spline:

\[
t_{id} = \beta_0 + \beta_1 C_d + \alpha_i + f(a_{id}, C_d) + \epsilon_{ipa}
\]

### Measuring Working Memory

Our working memory assessment/game is based on a “spatial delayed match to sample” task, which is widely used in cognitive science (Thomason et al., 2009; Goldman-Rakic, 1996). The game consists of 42 trials. In each trial, a respondent is shown a screen that has 1, 2, or 3 dots on it; each dot can occur in one of 9 screen locations. The respondent has 2 seconds to look at the screen and take note of where the dots are located. Then, the screen goes blank and there is a “delay” period during which respondent must wait while holding in mind the locations of the dots; then the screen “opens” and the respondent may touch where the dots used to be. This is represented schematically in figure 4a. On half the trials, the delay period is 0.1 seconds long, and on the other half it is 3 seconds.

Performance on the task can be scored in two alternative ways. In one approach, the baseline trials are those with the 0.1 second delay period, and the challenge trials have the 3 second delay period. In order to perform well in the challenge trials relative to the baseline, respondents must be able to hold the dot locations in mind for a longer period of time; this requires working memory skill. The alternative approach restricts comparisons to the 3-second-delay trials only, but treats trials where only 1 dot appears as the baseline, and trials where multiple dots appear as the challenge. In this approach, performing well in the challenge trials relative to the baseline entails holding more information in mind, which also requires working memory skill. For both scoring approaches, even the baseline requires some working memory skill. For example, even when the delay period is only 0.1 seconds, respondents must still hold in mind some dot locations for much longer than that, since it takes time to make the actual screen touches. Respondents take on average about 2 seconds to finish a 3-dot trial; during those 2 seconds, they must hold in mind the location of the third dot. Similarly, when longer delay 1-dot trials are used as the baseline, respondents must hold in mind the location of that dot, which requires working memory skill. As a result, in either scoring approach the baseline will sweep out some signal, along with the noise generated by differences in non-working-memory...
skills and competencies like familiarity with computer games. The different scoring approaches will likely be appropriate for analyzing different parts of the skill distribution.

The key performance measure in each trial is the deviation from their touch to the center of the nearest displayed dot. This represents a continuous measure of accuracy; it reflects not simply whether a respondent could hold in mind the location of the dot, but also how precisely they could hold that location in mind. We compute the deviation of respondent \(i\)'s touch from the location of dot \(d\) in trial \(t\) \(D_{idt}\) using the Euclidean distance from touch location to the center of the dot. Higher measures of \(D\) would indicate greater distance from touch to target, and so would be indicative of poorer performance. Figure 4b illustrates that the distribution of deviation is shifted right on long delay trials, relative to short delay trials, reflecting the challenge of holding the dot locations in mind for a longer period of time. Figure 4c indicates a similar pattern for the multiple dot trials relative to single dot trials.

We estimate differences in performance between baseline and challenge trials by regressing log deviation \((\ln(D_{idt}))\) against an indicator for whether the dot appeared during a challenge trial \((C_d = 1)\) or a baseline trial \((C_t = 0)\), as well as an individual child fixed effect \((\alpha_i)\) and a fixed effect for each screen location \((\lambda_d)\).

\[
\ln(D_{idt}) = \beta_0 + \beta_1 C_d + \alpha_i + \lambda_d + \varepsilon_{idt} \tag{6}
\]

The coefficient \(\beta_1\) indicates average working memory ability in the sample.

**Subgroup Comparisons**

We also compare assessed cognitive skills in the main cohort versus younger sibling samples, and between subgroups based on their rate of progress through school. It would be inappropriate to approximate those comparisons by simply estimating differences in \(\beta_1\) across these groups, because of imperfect overlap in the baseline performance distributions between the subgroups. For example, as figure 5a illustrates, the distribution of response times in baseline trials on the inhibitory control game differs substantially for younger and older children. The median response time on baseline trials among the younger children is 1.03 seconds; among the older children, that represents the 30th percentile. If challenge effects vary across the baseline distribution, the difference between global average challenge effects among the younger siblings to global average effects among the older siblings will in part reflect differences in the distribution of baseline response times between the two subgroups. This analytical challenge is analogous to those posed by imperfect between-group overlap in propensity score distributions (Dehejia and Wahba, 2002). Each respondent completed 30 baseline trials; in figure 5b, we report the discretized distribution of within-individual averages.
over those thirty trials. Among the younger siblings, 12.9% averaged 1210-1260 milliseconds in their baseline trials. By contrast, among the main cohort, the modal response bin is 960-1010 milliseconds (11.6% of main cohort respondents). The importance of the between subgroup differences in the baseline distribution is illustrated in figure 5c. We computed average challenge effects and average baseline response times for each respondent, and plotted a locally smoothed nonparametric regression over the region of common support. Respondents who tended to be slower on baseline trials tended to show smaller challenge effects. Therefore, a direct comparison of global average challenge effect between the subgroups will underestimate the average skill differentials, since more of the younger siblings are represented among the slower baseline trial response times. In general, we observed this pattern for almost all performance measures—poorer baseline performance tended to corresponded with smaller challenge effects.

We address this analytical challenge using an approach analogous to a “caliper matching” technique in propensity score applications (Dehejia and Wahba, 2002). Specifically, we group respondents on the basis of their average baseline performance in each game, as shown for example in figure 5b for the inhibitory control game. We then include a challenge condition by baseline performance group fixed effect 

\[ C_d \xi_i \]

in all regressions comparing cognitive skills across subsamples of respondents.\(^3\) For example, this is implemented using the following extension of regression equation (5), in which \( S_i = 1 \) for younger siblings and \( S_i = 0 \) for the main cohort:

\[
t_{id} = \beta_0 + \delta C_d S_i + C_d \xi_i + \alpha_i + f(a_{id}, C_d) + \varepsilon_{ipa}
\]

(7)

We extend each of equations (1)-(6) in this way. In each regression, \( \delta \) represents the difference in assessed skill between subgroups.

4 Findings

4.1 Overall Performance

Table 2 reports overall performance in the main cohort. We report sample-wide average performance on each type of trial for each game, and the regression-adjusted differences based on equations (1)-(6).

Panel A describes average performance in the implicit learning game. Overall, respondents took about three quarters of a second to touch each dot. However, when the dots were moving around the screen in a pattern that they had learned by “muscle memory,” they were able to make contact on average 36 milliseconds

\(^3\)Note that the pure baseline performance group fixed effect, \( \xi_i \), is absorbed into the individual fixed effect, \( \alpha_i \).

Page 18
faster. That time difference would not be perceptible to the naked eye, but can be captured using the precise chronometry of a tablet computer. This time advantage provides a measure of average implicit learning skill in the sample overall.

Panel B describes average performance in the declarative memory game. Overall, respondents stood a 22% chance of correctly completing the pair on its first appearance; however, after they had seen the pair three times, they were able to complete it correctly on the first try 47% of the time. This 25 percentage point improvement represents a quantifiable measure of average declarative memory skill in the overall sample. Similarly, first row in the panel provides an alternative unit of measure of average declarative memory skill, 0.65 fewer guesses.

Panel C describes average inhibitory control performance. When respondents can simply follow their impulse and touch the dot as soon as it appears, they are able to do so, on average, in just over 1 second, and they tend to touch about 31 pixels toward the middle of the screen from the center of the dot. When they must override that impulse, and instead redirect their touch away from the dot, they take about 37.6 (\(= 1112.5 - 1074.9\)) milliseconds longer to do so, and touch about 60 pixels closer to the center of the screen from the dot’s mirror image. Since time and accuracy are jointly determined, adjusting each for the other uncovers larger inhibition effects than are implied by the unadjusted differences. For example, the result in the third column, second row of panel C indicates that the average respondent took 53.1 milliseconds longer to make a touch of comparable accuracy in a challenge trial, compared with a baseline trial. This provides two alternative units for measuring average inhibitory control capability in the sample—53.1 milliseconds, or 64.2 screen pixels.

Panel D describes average working memory performance. When respondents are allowed to touch the screen almost immediately after the dots disappear (section D1), they touch on average 49 pixels (\(= e^{3.9}\)) away from the correct location. When they have to hold in mind the location of the dots for a longer time (3 seconds), they touch on average 0.28 logs, (32%) farther away. The results using the alternative scoring approach are shown in section D2. That approach restricts attention to long delay trials only and uses one dot trials for baseline and multiple dot trials for challenge. The working memory effect of having more dots to hold in mind is larger than the effect of waiting; respondents are on average 0.62 logs (86%), less accurate when they have to hold in mind the locations of 2 or 3 locations for 3 seconds, compared with only holding in mind 1 location for 3 seconds. Comparing the baseline levels of error in the two alternative scoring approaches indicates that the short delay baseline (section D1) more difficult for the average respondent than the one dot baseline (section D2), since the average respondent makes 0.37 logs (44%) greater error in the former.

Taken together, these results indicate that performance on our assessment tool is broadly consistent with patterns observed in cognitive psychology labs in developed countries, even though our sample of respondents
has had relatively little exposure to touchscreen tablet PCs or computerized testing compared with participants in most lab studies. They provide direct, quantifiable measures of these important skills. Nonetheless, relating the units of these measures to outcomes with which population scientists are more directly concerned is not straightforward. In the next two sections, we investigate how these skills vary across subgroups within the Young Lives, Peru sample, in order to get further information on the internal validity of these measures, and to benchmark the novel units in which they are expressed.

### 4.2 Age Variation in Cognitive Skills

Since the four skills that we assessed are physiologically distinct, they are likely to follow different developmental trajectories. The Young Lives, Peru study is focused on longitudinal followup of a single birth cohort, which will eventually facilitate the use of repeated measures to trace these trajectories in the cohort as a whole, and for specific subgroups. In the meantime, we use the performance of younger children (aged 6-8 years) in the sibling sample to observe skill differences between older respondents and younger respondents. The younger respondents are all a next-younger coresident sibling of a main cohort respondent, and thus they are selected on birth order, birth spacing, and family living arrangements. In light of these selection dynamics, we are careful not to interpret any observed variation in this study as reflecting age variation at the population level in these communities. Nonetheless, since there is reason to expect almost all older children to have an advantage over almost all younger children in terms of these skills, comparing the two groups provides useful insights about the validity of the measures.

Table 3 reports age differences in performance on the task. On both challenge and baseline trials in every game, the younger children perform worse than the older children. This may reflect skill differences between the two groups, but it almost certainly also reflects differences in exposure to technology, ability to hear and follow instructions, fine motor skills, and other characteristics. In order to isolate skill differences specifically, we use differences in performance on baseline trials to account for these irrelevant characteristics. Specifically, we estimate regressions like (7) for each game, in which $S_i = 1$ for younger children and $S_i = 0$ for older children; we then report our estimate for $\delta$, which represents assessed skill differences between the age groups.

The top row of panel A indicates that in the implicit learning game, the younger children respond on average about 124 milliseconds slower on the challenge trials and 117 milliseconds slower on the baseline trials. These differences are substantial, representing about 15-18% of the response time of the older children. The average "muscle memory" effect is 9.6 milliseconds (28%) smaller for younger children than for older children who perform similarly in the baseline trials.
Panel B indicates that in the declarative memory game, even after having seen a pair three times, the younger respondents required 0.3 more guesses than the older children to correctly match the two members, and were 8 percentage points less likely to correctly complete the pair with their first guess. By contrast, the performance of both age groups is very similar at the first appearance of each pair. As a result, assessed declarative memory is 0.23 guesses, and 10.3 percentage points, better among the older children than the younger children. These differences represent 35-50% of the average among the older children.

The top row of panel C indicates that on the inhibitory control game, the challenge effect on response time is 42.4 milliseconds greater on younger children than it is on older children who perform similarly in the baseline trials. This pattern is consistent with the expectation that older children would have better inhibitory control than younger children. By contrast, the second row of panel C indicates that the challenge effect on accuracy is very similar for younger children and for older children who perform similarly on baseline trials.

Panel D indicates that in the working memory task, the challenge effects—whether the challenge is posed by a longer delay during which locations must be held in mind, or by more dot locations to hold in mind—are substantially larger for younger children compared with older children who perform similarly in the baseline trials.

Overall, these results indicate that our games are powerful for identifying skill differentials. Substantial evidence in psychology and child development indicates that these domain-general abilities improve as children grow and develop. This evidence provides reason to expect a priori that younger children would have lower levels of measured skill than older children. For the most part, our tasks quantify these skill differentials at a high level of precision, even with relatively small sample sizes (396 younger children). However, the results also indicate that using precise screen location in the inhibition game as a performance measure may not be appropriate for very young children, since even the baseline trials in those cases are sufficiently challenging to introduce consequential noise to the measures.

### 4.3 School Progress & Cognitive Skills

Cognitive skills like declarative memory or executive function are likely to be important for school success. For example, students with better declarative memory skills are likely to find it easier to learn basic rules of arithmetic or grammar, and students with better inhibitory control may be better able to take advantage of classroom opportunities, since doing so often requires overcoming counterproductive impulses or resisting distraction. Causal relationships are also likely run the other way; there is evidence from psychology, neuroscience, and child development research that skills like declarative memory or executive function improve
with practice, and a classroom environment may provide opportunities for that practice (Diamond et al., 2007; Jaeggi et al., 2008; Mackey et al., 2011, 2012). Finally, school success and these skills may jointly be influenced by family or background characteristics like parenting style or genetics.

Table 4 illustrates that our assessment is sufficiently powerful to identify relationships between school progress and relevant cognitive skills. The Young Lives, Peru survey includes information on the date of birth, current school grade, and highest grade completed for all respondents. We used this information to compare each child’s educational progress against the school grade in which they would be enrolled at the time of the survey if they had started first grade in the school year after they reached the official starting age, and proceeded without interruption. We classified the 1857 respondents in the main cohort into three subgroups—behind normative grade for age (about 19% of the sample), ahead of normative grade (about 28%), and at normative grade (the remaining 53%). We estimated regressions like (7) for each of the games, and report our estimates of assessed skill differences across the subgroups.

Panel A indicates that respondents in the main cohort were similar in terms of implicit learning capability, independent of their schooling level. In the interest of succinctness, we only present results using the last two blocks; results using all blocks are substantively identical. This pattern is consistent with a priori expectation; implicit learning is a very basic skill, which is unlikely to be very responsive to environmental conditions. Respondents who have better “muscle memory” may have an advantage in athletic competition, for example, or in the performance of highly repetitive manual labor tasks; however, there is no reason to expect this skill to be important for school success, or to develop as a result of schooling experience.

By contrast, the rest of the table indicates that school progress is related to declarative memory and executive function skills. As we have discussed, these skills are likely to be critical for school success and to be more responsive than implicit learning to environmental inputs. Children who are behind their peers in school performed worse in the declarative memory game. The performance deficit of 6.2 percentage points is substantial; it corresponds to about 60% of the difference between the average 11-12 year old and the average 6-8 year old in the sample (see table 3). Results using the number of incorrect guesses as the performance measure are substantively identical. On the inhibitory control game, children who are ahead of their peers have substantial advantages over their peers; the 10.9 millisecond advantage corresponds to more than half the difference between the average 11-12 year old and the average 6-8 year old. Respondents who are behind in school perform about 31% (= 24.7/76.0) more poorly in the inhibitory control game when scoring is based on accuracy. The results in panel D indicate similar patterns for working memory skill. Respondents who are behind in terms school progress perform 15-20% more poorly in the working memory game compared with respondents who are progressing normatively.
Differences like these can also be used to benchmark magnitudes. For example, our assessment tool measures inhibitory control skills in milliseconds or screen pixels. These units are novel, but the results in table 4 provide a basis for relating them to outcomes with which population scientists are likely to be more familiar. A difference of one millisecond in the inhibitory control score corresponds to about 11% of the difference between children proceeding normatively through school, and those who are ahead of norm. Similarly, a deficit of 2 screen pixels would correspond to about 10% of the deficit among children who are behind in their schooling progress relative to those proceeding normatively.

5 Conclusion

Non-academic capabilities which are important for population health and socioeconomic well-being span a broad range, from intelligence to personality to emotional characteristics. In this study, we focus on a subset of cognitive skills that are involved with intelligence, but which are not specific to any circumscribed domain of knowledge like reading or mathematics. Measuring domain-general cognitive capabilities has been an object of increasing interest in population studies, and has motivated incorporating tools like Raven’s Progressive Matrices (RPM), standardized testing instruments, or Forward/Backward Digit Span into population surveys. We developed a novel application for tablet computers that administers assessments of four distinct skills, in the form of simple computer games. The assessments are based on paradigms that have been in wide use in the cognitive sciences for decades. Similar to tools like the RPM, our application is cheaply portable, and does not rely on any prior exposure to language, mathematics, or other topic-specific knowledge. Unlike those tools, it provides discrete measures for discrete skills, is non-proprietary, and provides analysts trained in statistics with all the underlying data necessary to score and analyze respondents’ performance judiciously given the specific study context and hypotheses.

This study introduces our measurement tool, and demonstrates its validity in a population-based sample from a developing country. We separately measured implicit learning, declarative memory, inhibitory control, and working memory in over 1800 Peruvian adolescents (aged 11-12) and about 400 of their younger siblings (aged 6-8). Despite limited prior experience with tablet computers or computer games within the sample, respondents’ performance in the assessments was consistent with a priori expectation based on evidence from the cognitive sciences. Measured skill gradients are also consistent with expectation. Younger children have lower levels of measured skills compared with adolescents. Among adolescents, those who were behind in school had lower measured skill levels specifically for those skills that were expected a priori to covary with academic achievement.
Our assessments generate skill measures that are recorded in units like milliseconds or screen pixels, which are novel in population science research. Documenting variation in measured skills at the population and subgroup level in multiple contexts is an important first step toward benchmarking these units of measure against standard outcomes. For example, our findings in this study suggest that, at least among adolescents and children in the Young Lives, Peru communities, a difference in inhibitory control score of 42 milliseconds might reasonably be described as “about 4 years’ worth of development.” Since children who are behind the norm in schooling progress score 11 milliseconds more poorly on the assessment, they might reasonably be described as “about a year’s worth of development behind their peers.” As more is learned about variation in these skills within and across populations, more nuanced and complete comparisons will be possible.

There is a large and growing body of research in the population sciences examining broad domain-general constructs related to cognitive ability, like IQ or fluid intelligence. Information and models from neuroscience and psychology indicate potential value for unpacking the black box of these broad constructs, by analyzing separate cognitive skills separately. For example, the physiology underlying declarative memory differs from that underlying inhibitory control. Explicitly accounting for these differences is likely to be important for generating a nuanced understanding of how social processes “get under the skin” to affect well-being. On the other hand, even state-of-the-art laboratory measures are of limited value if measurement is confined to observations and experiments involving small, non-representative samples of very specific subpopulations like college students in the United States, or if analysis and interpretation takes insufficient account of social or economic processes. Our objective in developing our tablet computer application was to facilitate the leveraging of state-of-the-art measurement in cognitive science with standard tools in population science like population representative sampling, long follow-up with managed attrition, broad interviews covering a range of relevant social and economic outcomes, comparable measurements collected on multiple family members, and analytical techniques that generate or exploit exogenous sources of variation. Our experience indicates that this can be done effectively even in samples with minimal prior exposure to tablet computers, at manageable cost.
### Table 1: Analytical sample

<table>
<thead>
<tr>
<th></th>
<th>Main cohort (aged 11-12 yrs at round 4 interview)</th>
<th>Younger siblings (aged 6-8 yrs at round 4 interview)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target respondents</td>
<td>2052</td>
<td>455</td>
</tr>
<tr>
<td>Successfully interviewed in round 4</td>
<td>1869</td>
<td>398</td>
</tr>
<tr>
<td>Analytical sample</td>
<td>1857</td>
<td>396</td>
</tr>
<tr>
<td>Age (in days) at round 4 interview (Median &amp; IQR)</td>
<td>4269 [4365, 4458]</td>
<td>2574 [2883, 3144]</td>
</tr>
</tbody>
</table>

**Notes for table 1:** Characteristics of the analytical sample. The left column describes children in the main cohort which is being followed as part of the Young Lives, Peru Study. At the time of the round 4 interview—when they completed our cognitive assessments—these respondents were aged 11-12 years. Age-eligible younger siblings, described in the right column, were 6-8 years old when they completed the cognitive assessments. Attrition represented between the top row and second row is accounted for by loss to follow up at some point between round 1 (which occurred in 2001-2002) and round 4 (2013). Attrition between the second row and third is accounted for by technical glitches in the administration of our cognitive skills assessment application. For more details on the Young Lives, Peru study, see the “Sample” section in the main text, and Sánchez et al. (2015).
Figure 1: Assessment of Implicit Learning Ability

(a) Task Design

(b) Average Performance

- Block A
  - Disordered
  - Block B1
  - Ordered
  - Block A
  - Disordered
  - Block B2
  - Ordered

Response times (ms):
- 695.9
- 741.3
- 786.6
- 832.0
- 877.4

(c) Cumulative Response Time Distributions

- Cumulative Distribution: Block A
- Cumulative Distribution: Block B

This involves 1857 kids. CDF of RTPrpt. See /Projects/YLperu/pgms/RTtruncCDFoverlay-IL.sts. It's now 15 Jul 2015 at 18:38:05.

Cumulative Distribution:
- 0% responses faster
- 25% responses faster
- 50% responses faster
- 75% responses faster
- 100% responses faster
- 125% responses faster

Response time (ms): 0 250 500 750 1000 1250

Cumulative Distribution: Block A
Cumulative Distribution: Block B
Notes for figure 1: Implicit learning ability is the capacity to learn without conscious awareness. Panel 1a illustrates the design of the task we used to assess this capacity. Respondents were asked to touch the yellow dots as they appeared on the screen. Each dot disappeared after 1 second, or as soon as the respondent touched it–whichever came first. Baseline trials were in blocks A and A’ (35 trials each); in these blocks, the dots moved around the screen in random sequence. Challenge trials were in blocks B1 (70 dots) and B2 (35 dots); in these, the dots moved in a repeating cycle. A respondent would not be able to consciously note the repeating cycle, but as panel 1b illustrates, the average respondent was nonetheless able to learn it through unconscious “muscle memory,” which is a colloquial way to describe implicit learning. Each data point in panel 1b represents the average time taken for the respondent to touch the dot. Dotted lines and shaded regions represent block averages and 95% confidence intervals around those averages. The difference between a respondent’s reaction time in block A’ and block B2 represents a measure of their implicit learning ability. Cumulative distributions of response times in these two blocks are shown in panel 1c. For more details, see “Measuring Implicit Learning” in the main text.
Notes for figure 2: Declarative memory ability is the capacity to acquire, encode, and retrieve information that is not an object of continuous attention. Panel 2a is an example screen from the declarative memory game. Respondents were asked to identify the “partner” for the shape displayed at the top of the screen from among the shapes at the bottom. There were a total of 6 pairs in the task, and respondents saw each pair 4 times (twice at the beginning of the visit and then twice more at the end, about 20 minutes later). The first time they saw each pair, they could only guess which option was the correct “partner,” by touching the options one at a time. At subsequent appearances, however, respondents could rely on their declarative memory instead of just guessing. Panel 2b illustrates that by the fourth time they saw a pair, the average respondent had more than doubled the likelihood that they would identify the correct “partner” on the first try. Error bars in panel 2b indicate the 95% confidence interval around the difference from baseline (the maroon areas). These differences represent measures of declarative memory ability. For more details, see “Measuring Declarative Memory” in the main text.
Figure 3: Assessment of Inhibitory Control Ability

(a) Task Design:
- $t_0 = 0$
- $t_2 = t_1 + 0.1$
- $t_1 \leq 2.5$
- $t_3 \leq t_2 + 2.5$
- $t_4 = t_3 + 0.1$
- Dot 1 ON
- Dot 2 ON
- Dot 3 ON

(b) Overall Performance (Response Time):

<table>
<thead>
<tr>
<th>Density</th>
<th>450</th>
<th>900</th>
<th>1350</th>
<th>1800</th>
<th>2250</th>
</tr>
</thead>
<tbody>
<tr>
<td>SameSide (Baseline)</td>
<td>0.0006</td>
<td>0.0012</td>
<td>0.0017</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| OppSide | See /Projects/YLperu/pgms/RTnXl_dist

It's now 27 Nov 2014 at 13:22:04

(c) Overall Performance (Accuracy):

<table>
<thead>
<tr>
<th>Density</th>
<th>450</th>
<th>900</th>
<th>1350</th>
<th>1800</th>
<th>2250</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>0.0057</td>
<td>0.0113</td>
<td>0.0170</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| location | See /Projects/YLperu/pgms/RTnXl_dist
| Midline | See /Projects/YLperu/pgms/RTnXl_dist

It's now 27 Nov 2014 at 13:22:04

(d) Accuracy and response time:

<table>
<thead>
<tr>
<th>Correct Midline</th>
<th>Horizontal touch location (px)</th>
<th>250</th>
<th>750</th>
<th>1250</th>
<th>1750</th>
<th>2250</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short delay (Baseline)</td>
<td>0.0057</td>
<td>0.0113</td>
<td>0.0170</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| Long delay | Splines. See /Projects/YLperu/pgms/BIsvafig_CogMeas.sts.
| See /Projects/YLperu/pgms/RTnXl_dist
| See /Projects/YLperu/pgms/RTnXl_dist

It's now 16 Jul 2015 at 16:08:21
Notes for figure 3: Inhibitory control ability is the capacity to control attention and behavior and override a counterproductive impulse. Panel 3a illustrates the design of the inhibitory control game. Respondents were shown 60 dots in quick succession, half of which were solid (yellow) and the other half striped (pink and grey). They were asked to touch inside the solid dots, and across the screen from the striped dots; the solid dots represent the baseline trials. On average, respondents reacted 37.3 ms faster to the solid dots than to the striped (panel 3b), because the latter required them to recruit cognitive resources to override the impulse to touch the dot itself, and instead redirect their movement toward the opposite side of the screen. In addition, their touches in response to the striped dots were on average 60.1 pixels closer to the middle of the screen compared with the solid dots (panel 3c). This task offers two measures of an individual’s inhibitory control ability. Each measure is determined jointly with the other. Panel 3d illustrates how average accuracy covaries with speed; it is a graphical representation of a regression of accuracy against a spline of response time with knots at the 10th, 25th, 50th, 75th, and 90th percentiles. For more details, see “Measuring Inhibitory Control” in the main text.
Figure 4: Assessment of Working Memory Ability

(a) Task Design

(b) Overall Performance (effect of delay)

(c) Overall Performance (effect of memory load)

Density

Deviation from correct location (pixels) [log scale]

See /Projects/YLperu/pgms/WMdev_CogMeas.sts. It's now 12 Dec 2014 at 23:25:19 and this is the one that includes the fills.
Notes for figure 4: Working memory ability is the capacity to hold in mind and manipulate information that is an object of continuous attention. Panel 4a is a schematic representation of our working memory game. In each trial, respondents were shown a configuration of 1, 2, or 3 dots. They had 1 second to note the locations of the dots on the screen, before the screen went to a blank grey either for 0.1 seconds or 3 seconds. During this “delay period,” respondents had to hold in mind the locations of the dots. Finally, the screen went to a blank white and respondents were asked to touch the screen in one place at a time, indicating where the dots used to be. There are two scoring approaches for this game. One treats the shorter delay trials as the baseline, and the longer delay as the challenge. On the longer delay rounds, respondents had to rehearse the dot locations in their minds for a longer time; as a result, they were on average less accurate in indicating where the dots used to be (panel 4b). Another approach restricts comparisons to the longer delay trials only, and treats the rounds with only 1 dot as the baseline, and those with 2 or 3 dots as the challenge. On the rounds with multiple dots, respondents had more information to rehearse to themselves; as a result, they were on average less accurate in indicating where the dots used to be (panel 4c). For more details, see “Measuring Working Memory” in the main text.
Inhibitory Control Assessment

Notes for figure 5: Each cognitive skill is assessed based on the difference in performance between baseline and challenge trials; in the specific example of inhibitory control, shown here, one performance measure is response time (for more details, see Measurement of Knowledge-Domain-General Cognitive Skills in the main text).
When comparing subgroups on assessed cognitive skill, we account for non-overlap in the support of the baseline distributions of different subgroups. For example, as illustrated in panel 5a, baseline distributions do not perfectly overlap in the younger sibling sample compared with the main cohort. Using regression equation (7), we compare the challenge effect within each bin, between the two age groups. As shown in panel 5c, the non-overlap is important, since challenge effects tend to be smaller for those who perform more poorly, on average, on baseline trials. We address this by computing each respondent’s average response time across the 30 baseline trials that they completed, and then assigning each respondent to one of 30 bins, as shown in panel 5b. Failure to account for this pattern will result in underestimation of the difference between the main cohort and younger sibling samples. For more details, see “Subgroup Comparisons” in the main text.
Table 2: Domain general cognitive skills, Main cohort (aged 11-12 years)

<table>
<thead>
<tr>
<th></th>
<th>Challenge trials</th>
<th>Baseline trials</th>
<th>Difference (Adjusted(^a))</th>
<th>Regression equation(^b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1. Implicit learning</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Challenge: both ordered blocks; Baseline: both disordered blocks)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Response time, in milliseconds</td>
<td>740.6</td>
<td>779.3</td>
<td>-34.2</td>
<td>(1)</td>
</tr>
<tr>
<td>A2. Implicit learning</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Challenge: last ordered block; Baseline: last disordered block)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Response time, in milliseconds</td>
<td>699.4</td>
<td>747.0</td>
<td>-36.3</td>
<td>(1)</td>
</tr>
<tr>
<td>B. Declarative memory</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Challenge: last appearance of pair; Baseline: first appearance)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number incorrect guesses</td>
<td>1.00</td>
<td>1.65</td>
<td>-0.65</td>
<td>(3)</td>
</tr>
<tr>
<td>Percent first touches correct</td>
<td>46.9</td>
<td>22.0</td>
<td>24.9</td>
<td>(2)</td>
</tr>
<tr>
<td>C. Inhibitory control</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Challenge: touch opposite side; Baseline: touch the dot)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Response time, in milliseconds</td>
<td>1112.5</td>
<td>1074.9</td>
<td>53.1</td>
<td>(5)</td>
</tr>
<tr>
<td>Horizontal error, in pixels</td>
<td>91.2</td>
<td>31.1</td>
<td>64.2</td>
<td>(4)</td>
</tr>
<tr>
<td>D1. Working Memory</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Challenge: 3 second delay; Baseline: 0.1 second delay)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Location error, in ln(pixels)</td>
<td>4.1</td>
<td>3.9</td>
<td>0.28</td>
<td>(6)</td>
</tr>
<tr>
<td>D2. Working Memory</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Challenge: Multiple dots; Baseline: 1 dot)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Location error, in ln(pixels)</td>
<td>4.2</td>
<td>3.6</td>
<td>0.63</td>
<td>(6)</td>
</tr>
</tbody>
</table>

**Notes for table 2:** Sample comprises 1857 children from the main cohort of the Young Lives, Peru study (aged 11-12 years at the time of interview). Each skill is assessed in its own game, consisting of baseline and challenge trials. The skill is assessed as the difference in performance between these two types of trials. For more details, see “Measurement Approaches” in the main text, and figures 1-4.

\(^a\) Standard errors for all differences (reported in square brackets) are computed using the “clustered sandwich” estimator of the variance/covariance matrix, to account for the fact that each game involves multiple observations of each respondent. Differences are computed in a regression context including a respondent fixed effect and other controls specific to each game. All differences are significantly different from 0 at up to 99% confidence.

\(^b\) Reported adjusted differences are estimates of \(\beta_1\) from the regression equation identified in this column. Regression equations are specified and discussed in the “Measurement Approaches” section of the main text.
Table 3: Age gradients in domain general cognitive skills

<table>
<thead>
<tr>
<th>Δ: (Younger siblings - Main cohort)</th>
<th>Challenge trials</th>
<th>Baseline trials</th>
<th>Difference (Adjusted&lt;sup&gt;a&lt;/sup&gt;)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1. Implicit learning</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Challenge: both ordered blocks; Baseline: both disordered blocks)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Response time, in milliseconds</td>
<td>124.4</td>
<td>117.0</td>
<td>9.6 [3.0]</td>
</tr>
<tr>
<td>A2. Implicit learning</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Challenge: last ordered block; Baseline: last disordered block)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Response time, in milliseconds</td>
<td>131.0</td>
<td>124.2</td>
<td>7.72&lt;sup&gt;b&lt;/sup&gt; [4.38]</td>
</tr>
<tr>
<td>B. Declarative memory</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Challenge: Last appearance of pair; Baseline: First appearance)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number incorrect guesses</td>
<td>0.31</td>
<td>0.08</td>
<td>0.23 [0.05]</td>
</tr>
<tr>
<td>Percent first touches correct</td>
<td>-8.13</td>
<td>2.21</td>
<td>-10.3 [1.55]</td>
</tr>
<tr>
<td>C. Inhibitory control</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Challenge: touch opposite side; Baseline: touch the dot)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Response time, in milliseconds</td>
<td>204.6</td>
<td>180.7</td>
<td>42.4 [5.94]</td>
</tr>
<tr>
<td>Horizontal error, in pixels</td>
<td>10.1</td>
<td>24.4</td>
<td>2.2&lt;sup&gt;b&lt;/sup&gt; [6.8]</td>
</tr>
<tr>
<td>D1. Working memory</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Challenge: 2 second delay; Baseline: 0.1 second delay)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Location error, in ln(pixels)</td>
<td>0.24</td>
<td>0.25</td>
<td>0.038 [0.010]</td>
</tr>
<tr>
<td>D2. Working memory</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Challenge: Multiple dots; Baseline: 1 dot)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Location error, in ln(pixels)</td>
<td>0.24</td>
<td>0.24</td>
<td>0.154 [0.034]</td>
</tr>
</tbody>
</table>
Notes for table 3: Sample comprises 1857 children from the main cohort of the Young Lives, Peru study (aged 11-12 years at the time of interview), and 396 of their younger siblings aged 6-8. Each skill is assessed in its own game, consisting of baseline and challenge trials. The skill is assessed as the difference in performance between these two types of trials. For more details, see “Measurement Approaches” in the text and figures 1-4.

a - Standard errors for all differences (reported in square brackets) are computed using the “clustered sandwich” estimator of the variance/covariance matrix, to account for the fact that each game involves multiple observations of each respondent. Differences are computed in a regression context including a respondent fixed effects, challenge condition by baseline performance group fixed effects, and other controls specific to each game. For more details, see “Subgroup comparisons” in the main text.

b - Not different from zero even at 95% confidence. All other differences in the table are different from zero at up to 99% confidence.
Table 4: Domain general cognitive skills & school progress

<table>
<thead>
<tr>
<th></th>
<th>At normative grade for age</th>
<th>Behind norm</th>
<th>Ahead of norm</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A. Implicit learning</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adjusted\textsuperscript{a} Difference: \textit{(last} ordered block) - \textit{(last} disordered block)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Response time, in milliseconds</td>
<td>-40.5\textsuperscript{b}</td>
<td>-1.05</td>
<td>2.70</td>
</tr>
<tr>
<td></td>
<td>[2.18]</td>
<td>[4.48]</td>
<td>[3.89]</td>
</tr>
<tr>
<td><strong>B. Declarative memory</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adjusted\textsuperscript{a} Difference: \textit{(last} appearance of pair) - \textit{(first} appearance)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percent first touches correct</td>
<td>25.3\textsuperscript{b}</td>
<td>-6.2\textsuperscript{b}</td>
<td>2.0</td>
</tr>
<tr>
<td></td>
<td>[1.0]</td>
<td>[1.8]</td>
<td>[1.6]</td>
</tr>
<tr>
<td><strong>C. Inhibitory control</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adjusted\textsuperscript{a} Difference: \textit{(touch} opposite dot) - \textit{(touch} on dot)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Response time, in milliseconds</td>
<td>69.8\textsuperscript{b}</td>
<td>6.48</td>
<td>-10.9\textsuperscript{c}</td>
</tr>
<tr>
<td></td>
<td>[6.2]</td>
<td>[5.40]</td>
<td>[4.4]</td>
</tr>
<tr>
<td>Horizontal error, in pixels</td>
<td>76.0\textsuperscript{b}</td>
<td>24.7\textsuperscript{b}</td>
<td>-5.98</td>
</tr>
<tr>
<td></td>
<td>[5.3]</td>
<td>[7.0]</td>
<td>[5.57]</td>
</tr>
<tr>
<td><strong>D1. Working memory</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adjusted\textsuperscript{a} Difference: \textit{(3 second} delay) - \textit{(0.1 second} delay)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Location error, in ln(pixels)</td>
<td>0.26\textsuperscript{b}</td>
<td>0.051\textsuperscript{b}</td>
<td>-0.022\textsuperscript{c}</td>
</tr>
<tr>
<td></td>
<td>[0.01]</td>
<td>[0.011]</td>
<td>[0.010]</td>
</tr>
<tr>
<td><strong>D2. Working memory</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adjusted\textsuperscript{a} Difference: \textit{(multiple} dots) - \textit{(one} dot)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Location error, in ln(pixels)</td>
<td>0.57\textsuperscript{b}</td>
<td>0.081\textsuperscript{b}</td>
<td>-0.023</td>
</tr>
<tr>
<td></td>
<td>[0.02]</td>
<td>[0.017]</td>
<td>[0.013]</td>
</tr>
<tr>
<td>Number of 11-12 year old children</td>
<td>986</td>
<td>344</td>
<td>527</td>
</tr>
</tbody>
</table>
Notes for table 4: Sample comprises 1857 children from the main cohort of the Young Lives, Peru study (aged 11-12 years at the time of interview). Each skill is assessed in its own game, consisting of baseline and challenge trials. The skill is assessed as the difference in performance between these two types of trials. For more details, see “Measurement Approaches” in the text and figures 1-4.

a - Standard errors for all differences (reported in square brackets) are computed using the “clustered sandwich” estimator of the variance/covariance matrix, to account for the fact that each game involves multiple observations of each respondent. Differences are computed in a regression context including a respondent fixed effect, a challenge condition by baseline performance group fixed effect, and other controls specific to each game. For more details, see “Subgroup Comparisons” in the text.

b - Different from 0 at up to 99% confidence.
References


