Policy Research Working Paper 9537

Do Large-Scale Student Assessments Really Capture Cognitive Skills?

Rafael de Hoyos
Ricardo Estrada
María José Vargas

WORLD BANK GROUP
Education Global Practice
February 2021

Policy Research Working Paper 9537

Abstract

This paper studies the relationship between test scores and cognitive skills using two longitudinal data sets that track student performance on a national standardized exam in grades 6, 9, and 12 and post-secondary school outcomes in Mexico. Using a large sample of twins, the analysis finds that primary school test scores are a strong predictor of secondary education outcomes and that this association is mainly driven by the relationship between test scores and cognitive skills, as opposed to family background and other general skills. Using a data set that links results in the national standardized test to later outcomes, the paper finds that secondary school test scores predict university enrollment and hourly wages. These results indicate that, despite their limitations, large-scale student assessments can capture the skills they are meant to measure and can therefore be used to monitor learning in education systems.

This paper is a product of the Education Global Practice. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://www.worldbank.org/prwp. The authors may be contacted at rdehoyos@worldbank.org.

Do Large-Scale Student Assessments Really Capture Cognitive Skills?*

Rafael de Hoyos† Ricardo Estrada‡ María José Vargas§

*Acknowledgments: The authors would not have been able to write this paper without the support of the Mexican Secretariat of Public Education (SEP), and we are particularly grateful to Ana María Aceves and Proceso Silva for granting us access to the data without compromising confidential information. The paper benefited from comments made by Lucila Berniell; José Luis Gaviria; Jérémie Gignoux; José Felipe Martínez, Liliana Sousa; Juan Vargas; Paula Villaseñor; and participants at seminars at CAF, the World Bank/IZA Conference at Universidad Javeriana, the RISE Conference at Oxford University, the RIDGE Poverty and Inequality Workshop at EAFIT Medellin, and the LACEA Meeting at BUAP Puebla. We are grateful to Elizabeth Monroy for putting together the data sets used in this paper. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. A previous version of this manuscript was circulated under the title "Predicting Individual Wellbeing Through Test Scores: Evidence from a National Assessment in Mexico."

†The World Bank, ITAM, Xaber, rdehoyos@worldbank.org

‡CAF-Development Bank of Latin America, restrada@caf.com

§The World Bank, mvargasmancera@worldbank.org

1 Introduction

There is increasing recognition that education brings about individual and society-wide benefits when students acquire a set of relevant skills during their formative years. Literacy and numeracy stand out among these skills, as they are seen as the foundation for the acquisition of other skills and a direct determinant of critical factors for personal and social well-being, such as labor market outcomes, health conditions, and democratic participation (World Bank, 2018; CAF, 2016). There is, however, no consensus about whether education systems can measure these foundational skills and track their progress at scale.

In recent years, many developing countries have implemented large-scale student assessments via standardized tests to monitor cognitive skills such as literacy and numeracy.[1] However, critics of standardized tests argue that this type of testing promotes a reductionist approach to education that emphasizes literacy and numeracy to the detriment of other equally important subject areas. Moreover, critics point out that the reliability of standardized tests is compromised by either the absence or the presence of incentives. On the one hand, students might not put enough effort into a test if no consequences are attached to the results (see, for example, Akyol et al., 2018 and Gneezy et al., 2019 and a review in Finn, 2015). On the other hand, high-stakes tests can create perverse incentives that lead to game the system or teach to the test, which may raise test scores but not truly improve students' learning.[2] The effects of these incentives on the reliability of standardized tests may be larger in developing countries, where weak implementation capacity is more prevalent. Hence, this paper seeks to answer the following question: Do large-scale student assessments based on standardized testing capture the cognitive skills they are designed to measure?

One way to address this question is to estimate the relationship between test scores and future education and labor market outcomes.[3] A positive correlation between test scores and

[1] Cheng and Gale (2014) survey 125 developing countries and find that 105 of them have implemented a national student assessment based on standardized testing.

[2] For a discussion on incentives and strategic behavior in testing, see Figlio and Loeb (2011), Koretz and Barron (1998), Koretz (2017), and Neal (2011).

[3] The other standard alternatives are to correlate test scores with other measures of student learning—notably GPA—or contemporaneous measures of labor market outcomes. However, the first strategy is subject to circularity and the second to reverse causality, as discussed in detail in Heckman and Kautz

future outcomes is indicative but does not prove that standardized tests capture cognitive skills, as other unobserved factors could drive this correlation. A recent strand of papers has documented the direct effects of school and neighborhood inputs on test scores (Salardi and Michaelsen, 2019; Chang and Padilla-Romo, 2019; Lavy et al., 2014). In addition, an important body of literature in psychology and economics has shown the significant influence of noncognitive skills on both test scores and labor market outcomes (Duckworth and Seligman, 2005; Heckman and Kautz, 2012; Heckman et al., 2006). In this paper, we deal with these potential confounders by implementing two strategies. First, we use twin fixed effects to control for all between-family differences in school and household inputs. Second, we use within-individual variation in test scores to control for the effect of general (noncognitive) skills on test scores.

We construct two longitudinal data sets to track students along their education trajectories and initial labor market outcomes in Mexico. Both data sets use information from the National Assessment of Academic Achievement in Schools (ENLACE, from the Spanish), a census-based standardized test given to primary and secondary school students between 2006 to 2014. The first data set tracks students who took the test in 2007 at the end of primary school (grade 6), 2010 at the end of lower secondary school (grade 9), and 2013 at the end of upper secondary school (grade 12). Given the large size of this student cohort (close to 2 million individuals) we are able to identify about 10,000 pairs of twins in this data set. The second data set merges the grade 12 test scores with a special module of the Mexican labor force survey (ENOE, from the Spanish) of 2010 that was given to secondary school graduates between the ages of 18 and 20.

The results of this paper show that large-scale standardized tests, despite their limitations, can meaningfully capture cognitive skills. We find a positive and significant relationship between test scores and future education outcomes that remains even after controlling for observed and unobserved family heterogeneity. In the twin fixed effects specification, a 1-standard-deviation (SD) higher score in grade 6 is correlated with a 3.3-percentage-point higher probability of on-time graduation from grade 9 and a 5.7-percentage point increase

(2012).

for grade 12, and with 0.49 and 0.53 SD higher test scores, respectively. These estimated coefficients can be interpreted as the composite effect of skills at grade 6 on future outcomes and the within-twin variation in factors with a direct effect both in grade 6 and later education outcomes (e.g., motivation to perform in the test). To explore the importance of noncognitive skills in this relationship, the main specification is expanded to include grade 6 test scores from a different subject area—grade 6 math (language) test scores are included as a control in the regressions for language (math) test scores at grades 9 and 12. The rational for this econometric specification is that cognitive abilities are captured by subject-specific correlations, say grade 6 math on grade 12 math, while the conditional cross-subject correlation captures general skills (which notably include general noncognitive skills). The results from this specification show that grade 6 test scores have a strong relationship to test scores in the same subject in grades 9 and 12, controlling for twin fixed effects and grade 6 test scores in the other subject area. These results suggest that large-scale student assessments measure mainly cognitive skills as opposed to capturing mainly socioeconomic status or noncognitive skills.

Finally, we analyze the relationship between grade 12 test scores and post-secondary school outcomes using the second panel data set. As this panel is less rich than the first one, the analysis follows a conventional strategy and substitutes the twin fixed effects with a set of covariates to control for family background. The results show that grade 12 test scores are a good predictor of university enrollment and hourly wages. A 1-SD increase in grade 12 test scores is correlated with a 10-percentage point increase in the likelihood of being enrolled in university, and, conditional on being employed, individuals with 1-SD-higher test scores in grade 12 have wages that are 6% higher 1 or 2 years after graduating from high school.

This paper joins the papers that study the relationship between test scores and future outcomes in developed countries (Chetty et al., 2011, Lin et al., 2018, Murnane et al., 1995, and Rose, 2006 for the United States; Currie and Thomas, 2001 for the United Kingdom; and Lindqvist and Vestman 2011 for Sweden) as the first paper to perform this analysis in a developing country. Conducting this study in a developing country is important because the relationship between test scores and future outcomes may differ from the one in developed countries, given the former's weaker implementation capacity and less mature labor markets. A second contribution of this paper is that it uses a census-based student assessment as opposed to the results of a controlled assessment conducted in the context of a research project—which is likely not subject to the same incentives and implementation challenges as large-scale student assessments. The third and most important contribution is that the main analysis is the first to be carried out in a large sample of twins, which allows us to study the relationship between test scores and future education outcomes among individuals with identical family background. Finally, the paper uses individual variation across two subject areas to implement an exploratory analysis to control for the direct effect of (general) noncognitive skills on test scores.[4] Overall, the evidence presented in this paper is of particular relevance to the large applied economics literature that uses large-scale student assessments to study the process of human capital formation or evaluate the effects of certain education policies.[5]

The results presented here have two important policy implications. First, although appropriate design and implementation matter for the quality and credibility of large-scale student assessments, the results presented here support the use of large-scale student assessments to measure and monitor the evolution of learning outcomes in complex education systems with relatively weak institutions and low implementation capacity. Second, the findings of this paper show that student test scores are a good predictor of future outcomes related to well-being. This is relevant for the design of effective policies to identify and address education disparities and, therefore, promote social mobility. Results from large-scale student assessments, especially those of early grades, could be used to target resources toward schools and students with the lowest learning levels—who also tend to be the poorest.

[4]The findings of this paper are related in a complementary way to the rich psychometrics literature that studies the validity of large-scale skills assessments, either by validating the internal consistency of the constructs, testing structure and sampling design, or by studying the use of results coming from such assessments by policy makers or researchers outside psychology (see, for example, Braun and von Davier 2017; Lin et al. 2014). This paper is also connected to the economics literature that uses structural models to identify the contributions of cognitive and noncognitive skills to test scores (Cunha et al., 2010; Heckman et al., 2006) and recent work by Laajaj and Macours (2019), who validate measures of human capital that are widely used in applied economics (survey-based measures of skills for rural contexts, in their case).

[5]For examples using ENLACE see: Avitabile and de Hoyos, 2018; Dustan et al., 2017; Estrada, 2019; Estrada and Gignoux, 2017; Salardi and Michaelsen, 2019.

The rest of the paper is organized as follows. Section 2 presents the Mexican education system and the ENLACE test, Section 3 describes the panel data sets, Section 4 describes the analytical framework and empirical strategy, Section 5 presents and dicusses the results, and Section 6 concludes.

2 Context

2.1 The Mexican Education System

The basic education system in Mexico includes 3 years of preprimary, 6 years of primary, and 3 years of lower secondary (grades 7 to 9) school, and upper secondary education is for 3 years (grades 10 to 12). Both the basic and the upper secondary levels are mandatory. More than 30 million students are enrolled in mandatory education across 243,000 schools, that employ close to 1.5 million teachers. In the basic education system, public schools, which are decentralized at the state level, account for 90% of total school enrollment. In the upper secondary system, public schools are managed by the federal government, state governments, and public universities. Private sector education is a relatively small share of the education system, and it is heavily regulated by the Secretariat of Public Education (SEP, from the Spanish).

Most children ages 6 to 12 are enrolled in the education system and graduate from primary school. However, of every 100 students enrolled in lower secondary school, only 85 graduate on time, and this number falls to 65 in upper secondary. The weak performance of Mexican students in international assessments, notably the Program for International Student Assessment (PISA), gave the learning crisis a prominent place on the public agenda—Mexico ranked last among OECD (Organisation for Economic Co-operation and Development) countries in PISA 2000, the first wave of the assessment (OECD, 2001). These concerns contributed to the establishment of a national standardized assessment in 2006: ENLACE.[6]

[6]For more information on ENLACE see: http://www.enlace.sep.gob.mx/

2.2 The ENLACE Standardized Test

From 2006 to 2014, the SEP administered ENLACE, a census-based standardized test that gathered information on student achievement in math, literacy, and a rotating subject. Initially, ENLACE was given to students in grades 3 to 9 (primary and lower secondary), but starting in 2008, ENLACE was also given in grade 12 (the final year of secondary school).

ENLACE was designed as a low-stakes assessment and had no bearing for students on GPA, graduation, or admission to the next schooling level. The purpose of the assessment, as stated by SEP, was to increase parental and student participation in the learning process, improve lesson preparation, improve teacher and principal training, strengthen policy planning, and increase transparency and accountability.

School participation was mandatory for grades 3 to 9 and optional for grade 12. Nonetheless, the take-up of the test was consistently above 85%, even among 12th graders. The states of Michoacán and Oaxaca have consistently recorded the lowest participation rates. These states also have a heavy presence of teacher unions (Table 1). A total of 15.1 million students in 136,000 schools took the examination in 2013, the last year ENLACE was given in most grades.

By design, ENLACE had a national mean score of 500 and a SD of 100 for every subject area and grade in its first year of implementation. ENLACE's methodology followed item response theory, with comparability of the results over time. The test had 50 to 75 questions per subject and was applied in eight 45-minute sessions over 2 days. Several mechanisms were in place to prevent exam manipulation. An external coordinator was assigned to every school to oversee the test's administration, alongside the school principal. Teachers were not allowed to monitor their own classes, and parents were invited to act as exam monitors. Exams were centrally marked by SEP, and computer software identified abnormal response patterns to detect cheating.

SEP produced school report cards that were distributed among school principals and an online website where parents and students could check their individual results. Yet, few schools viewed ENLACE as a diagnostic tool, which limited its effectiveness as an improvement tool (de Hoyos et al., 2017). Results from ENLACE received wide attention from the

Mexican public. Every year, the results made the front page of most newspapers. NGOs produced and disseminated state and school rankings that made ENLACE a medium-stakes test from the point of view of the school directors and the school community. Despite the original objectives of the assessment, SEP used ENLACE scores to deliver monetary bonuses to teachers and principals participating in two different incentive programs—one starting in 2008 and a second one in 2009.[7] This decision made ENLACE a de facto high-stakes test for teachers and principals, encouraging strategic behavior and resulting in multiple concerns about teaching to the test and grade inflation.

Complaints about ENLACE—mainly related to grade inflation—and the creation of the national autonomous evaluation institute (INEE, from the Spanish), which was given the responsibility of regulating national student assessments, contributed to the cancellation of ENLACE. The test was administered for the last time in 2013 for grades 3–9 and in 2014 for grade 12. The INEE launched a new, survey-based test called *Plan Nacional para la Evaluación de los Aprendizajes* in 2015.

3 Data

Two longitudinal data sets are constructed for this paper. The first panel is formed from ENLACE test scores of students who completed primary school (grade 6) in 2007, lower secondary school (grade 9) in 2010, and upper secondary school (grade 12) in 2013. The second panel merges a special module of the Mexican labor survey (ENOE, from the Spanish) applied to individuals ages 18, 19, and 20 years during the third quarter of 2010, with students that sat the ENLACE test in grade 12 in May 2008, 2009, and 2010.[8]

[7] In 2008, SEP linked ENLACE to *Carrera Magisterial*, a national teacher incentive program that offered bonuses to primary and lower secondary teachers participating in the program. Students' ENLACE scores were given a weight of 20% in the program's total score; this weight was increased to 50% in 2011. In 2009, SEP launched the Program of Incentives for Teaching Quality, which delivered monetary bonuses to teachers and principals of classrooms and schools that performed highly on in specific categories on ENLACE.

[8] The data sets described in this paper can be requested from www.xaber.org.mx.

3.1 The ENLACE Panel

The ENLACE data set recorded the unique national population identifier (CURP, from the Spanish) of all test takers, enabling the construction of a panel of students with learning outcomes at different points in their education trajectory. In addition to these outcomes, the ENLACE data set includes a school identifier, and the full name, birthday, state of birth, and sex of the student.[9] Using the CURP, we merged the information from all grade 6 students who took the exam in 2007 with their exam results from 2010 (grade 9) and 2013 (grade 12). We begin the panel in 2007 because of the relatively low take-up in 2006 (the first year ENLACE was administered) and the lack of use of CURPs in some states during this first year of application. Among the 1,881,470 students who sat the test in 2007 and had complete information on their CURP (94.5% of the total), we were able to identify 72.9% three years later in grade 9 and 34.5% six years later in grade 12 (Figure 1).[10]

The large attrition observed in the panel is caused by (1) grade repetition, (2) school dropout, (3) exam take-up rates of less than 100%, and (4) imperfect matching due to administrative data errors. If a large share of the attrition in the panel is caused by low test take-up or imperfect matching, we might not be able to identify accurately the effects of grade 6 test scores on lower and upper secondary on-time graduation rates. To quantify the magnitude of the causes behind attrition, we use administrative data from the annual school census (known as *Formato 911*) to estimate the expected survival rates given state-level repetition and dropout rates in lower and upper secondary school. Given the school trajectories implied by administrative data, 77% of the student population that completed grade 6 in 2007 was expected to finish grade 9 in 2010 and 39% to graduate from upper secondary in 2013 (Figure 1). Therefore, the ENLACE panel has a survival rate that is 4 and 5 percentage points lower vis-à-vis the survival rate implied by administrative data,

[9] The CURPs as well as the students' names and dates of birth are confidential data protected by Mexico's personal information laws. For this study, we were able to use this information by closely collaborating with SEP.

[10] Some upper secondary schools—concentrated in the states of Nuevo León, Coahuila, and San Luis Potosí—follow a 2-year curriculum instead of the regular 3-year curriculum. Hence, we also merge the information from grade 6 students in 2007 with the results of those who took ENLACE at the end of their upper secondary education in 2012. From the 648,301 students observed at the end of upper secondary education, we find 4% in 2012 (2-year upper secondary students) and 96% in 2013 (3-year upper secondary students).

in lower and upper secondary, respectively. This difference is the result of both less than 100% test take-up and imperfect matching. The ENLACE take-up rate is high, around 93% of the student population (excluding the states of Michoacán and Oaxaca). The difference in graduation or survival rates between the ENLACE panel and administrative data is not trivial, but it is reassuring that most of the attrition observed in the panel is explained by grade repetition and school dropouts, which are dimensions we want to examine in this paper.[11]

ENLACE take-up rates in the years under analysis were close to 93%, with the exception of the states of Oaxaca and Michoacán, which we exclude from the panel (see Table 1).[12] ENLACE included a context questionnaire which was applied to a random sample of students and parents. We link the panel data set we constructed to results from the context questionnaire to retrieve information on parental education and occupation. The sample size for Grade 6 was 7,557 observations in 2007, with 5,677 individuals with full responses to the variables of interest.

We use student's personal information to identify 20,252 twins in the ENLACE panel. We define twins as students with identical last names (the surname in Mexico is composed of father's and mother's last names) and birthday who attend grade 6 at the same school in 2007. After the matching, the twins identified account for 1.08% of the ENLACE panel in 2007, a level close to the prevalence of multiple pregnancies in Mexico. Table A.1 in the Appendix presents summary statistics for the ENLACE panel data set. The main estimations in the table include the sample of twins, but the results are similar if we use only the main sample.[13]

[11]Even though it is possible to merge grade 6 ENLACE of 2007 with grade 12 ENLACE for 2014 to analyze, separately, the rate of those who drop out from those who repeat grades, as shown in Avitabile and de Hoyos (2018), most of the attrition between grades 9 and 12 in ENLACE is explained by dropout rates rather than grade repetition.

[12]A fierce opposition to standardized testing by a local teachers union (CNTE, from the Spanish) explains the low ENLACE take-up rates and a sample that was not statistically representative of the test in Oaxaca and Michoacán.

[13]Table A.1 in the Online Appendix reports differences in means between the sample of twins and non-twins. Although statistically significant at conventional levels, the differences are very small and support the idea that multiple (versus single) pregnancies are mostly a random event.

3.2 The ENILEMS-ENLACE Panel

Every quarter of the year, the Mexican statistics office, INEGI (from the Spanish), collects labor market information through the national labor force survey (ENOE), a rotating household survey.[14] In some quarters, ENOE's core survey is complemented by a thematic module that is usually requested and financed by different secretariats and government agencies. During the 3rd quarter of 2010 (July to September), ENOE's special module was the *Encuesta Nacional de Inserción Laboral de los Egresados de Educación Media Superior* (ENILEMS), a survey targeting upper secondary school graduates ages 18, 19, and 20. The objective of ENILEMS was to provide information on the transition between the end of mandatory schooling (grade 12) and higher education or the labor market.[15]

The ENILEMS-ENLACE panel merges information from the respondents of the ENILEMS 2010 survey with their results on ENLACE grade 12 from May 2008, 2009, or 2010. Although ENILEMS 2010 did not capture the CURP, it included all the necessary information to create a pseudo-CURP formed by combining letters and numbers coming from the full name, sex, birth date, and state of birth. The difference between the pseudo-CURP and the CURP is that the former does not have the last three digits of the population identification code that the Mexican government generates. We created the pseudo-CURP for 7,105 observations included in ENILEMS 2010 using the official algorithm for generating the CURP.[16]

A simple merge of ENILEMS and ENLACE grade 12 using the pseudo-CURP and CURP, respectively, was able to match 2,820 observations (40% of the ENILEMS sample). This relatively low matching rate is almost entirely explained by the lack of the last three digits in the pseudo-CURP and measurement error in the variables used to produce the merge. An additional 18% of the sample was recovered manually by identifying coding or registration errors in the CURP generation process (i.e., errors in birth date or misspelled names). Overall, 58% of the individuals in the ENILEMS sample were matched to their ENLACE

[14] For more information on ENOE, see http://www.beta.inegi.org.mx/proyectos/enchogares/regulares/enoe/.

[15] For more information on ENILEMS, see http://www.beta.inegi.org.mx/proyectos/enchogares/modulos/enilems/.

[16] The CURP is an 18-digit unique personal identifier formed by a combination of letters and numbers taken from the individual's full name, date of birth, sex, and state of birth plus a three-digit code assigned by the Mexican population council. For more information, see https://renapo.gob.mx/swb/.

grade 12 test scores. After eliminating missing observations in the ENLACE score, the panel reaches a total of 3,781 observations. The ENILEMS-ENLACE panel also includes the information from the ENOE regular questionnaire for 3,718 matched observations. Table A.2 in the Online Appendix reports differences in means between the observations of ENILEMS that were matched with ENLACE scores and the ones that were not; the differences are small. Table A.2 in the Appendix presents summary statistics for the ENILEMS-ENLACE panel data set.

4 Analytical Framework and Empirical Strategy

4.1 Analytical Framework

The main challenge to study cognitive skills is that these are not directly observable. As many countries have implemented large-scale student assessments via standardized tests to measure cognitive skills, an important question is to what degree such test scores really capture cognitive skills. One way to validate skills assessments is to look at the predictive power of test scores over future education and labor market outcomes (Heckman and Kautz, 2012). To better understand this strategy consider the following structural equation:

$$test_{it} = \psi\left(C_{it}, P_{it}, I_{it}\right) \tag{1}$$

Equation 1 defines test scores of individual i at time t ($test_{it}$) as a function of contemporaneous cognitive skills (C_{it}), noncognitive skills (P_{it}) and other inputs (I_{it}). A test enjoys validity if higher cognitive skills produce higher test scores, in other words, $\partial test/\partial C > 0$. The higher the magnitude of $\partial test/\partial C$, the higher the validity of the test. But, as discussed in Section 1, test scores also depend directly on noncognitive skills (e.g., perseverance) and other elements such as school inputs (e.g., teaching to the test). Equation 1 makes explicit that the same test (i.e., set of questions) given to individuals of the same cognitive ability will produce different scores if individuals exhibit differences in, say, motivation to perform in the test.

Let future education or labor market outcomes of individual i in period *t+1*, y_{it+1}, be defined by the following structural equation:

$$y_{it+1} = \phi\left(C_{it+1}, P_{it+1}, I_{it+1}\right) \tag{2}$$

where C_{it+1} is the level of cognitive skills of individual i in *t+1*, P_{it+1} is a vector of noncognitive skills, and I_{it+1} is a vector of other inputs. In equation 2, $\partial y_{it+1}/\partial C_{it+1} > 0$ and $\partial y_{it+1}/\partial P_{it+1} > 0$ show that cognitive and noncognitive skills have a positive return in the labor market. Equation 2 recognizes that family, school, and neighborhood inputs, I_{it+1}, can have a direct effect on outcomes—net of their effect on skills formation.

As the following equations recognize, the process of skills formation is a dynamic process:

$$C_{it} = \chi\left(C_{it-1}, P_{it-1}, I_{it}\right) \tag{3}$$

$$P_{it} = \chi\left(C_{it-1}, P_{it-1}, I_{it}\right) \tag{4}$$

$$I_{it} = \chi\left(C_{it-1}, P_{it-1}, W_{i}\right) \tag{5}$$

Equations 3 and 4 state that skills are persistent over time but malleable. Cognitive and noncognitive skills follow a cumulative process whereby skills today depend on the amount of skills in the previous period plus the educational investment (including school, family, and other inputs). Ample evidence recognizes the cross-dependence of cognitive and noncognitive skills (Heckman and Kautz, 2012). Equation 5 shows that inputs at period t depend on skills on the previous period and parental endowment (W_i). Equations 3 to 5 also imply that skills and inputs are correlated across time. Notice that the framework distinguishes between the direct and indirect effects of skills on future outcomes; for instance, perseverance at time *t-1* can have a direct effect on outcomes at time t through perseverance at t, but also an indirect effect through its influence on cognitive skills formation between *t-1* and t.

The framework presented here details the causal path (time-persistent cognitive skills that produce both test scores and future outcomes) that link skills, test scores and future outcomes. However, it also outlines the existence of other causal paths that can potentially link test scores to future outcomes (mainly the direct effects of inputs and noncognitive skills).

4.2 Empirical Strategy

The starting point of the empirical strategy is an equation linking test scores at t with education and labor market outcomes at *t+1*:

$$y_{it+1} = \beta_0 + \beta_1 test_{it} + \beta_k X_{ik} + \epsilon_i \tag{6}$$

Where y_{it+1} is the education or labor market outcome of individual i in time period *t+1*, $test_{it}$ is the individual's simple average ENLACE score in math and language in period t, X_{ik} is a vector of controls for individual and family characteristics at t, and ϵ_i is a disturbance term. A $\hat{\beta}_1 > 0$ (or rather $\hat{\beta}_1 \neq 0$) is enough to conclude that test scores—from ENLACE in this study—predict future outcomes, but it is not enough evidence to conclude that test scores capture cognitive skills. As outlined in the previous section, components such as school inputs and noncognitive skills can have a direct effect on both test scores and education and labor market outcomes and, therefore, could induce a correlation between test scores and future outcomes.

Ideally, one would like to include I_{it} from equation 1, the full vector of inputs at time t, in the estimation of equation 6. However, this requires access to longitudinal data with information on present education or labor market outcomes, and past test scores, and family, school, and neighborhood inputs, which is rarely available in most data sets. In the absence of this information, researchers include a vector X_{ik} of controls for family background as a proxy for the flow of inputs—justified by the dependence of inputs on the availability of family resources. This strategy has strong limitations because of unobserved family heterogeneity that affects, in turn, school and household inputs. To improve on this strategy, we substitute X_{ik} for a vector of twin fixed effects (τ_f):

$$y_{it+1} = \beta_0 + \beta_1 test_{it} + \tau_f + \epsilon_i \tag{7}$$

This specification restricts the estimation to the sample of twins identified in the large ENLACE panel data set. Estimation of equation 7 exploits only the within-twin variation in grade 6 ENLACE test scores, identifying, therefore, the relationship between test scores and education outcomes net of all observed and unobserved family characteristics. As twins are individuals who share the same family and were born on the same day, we avoid differences in family circumstances or birth rank that could lead to differences in the availability of family resources devoted to children from the same family. Of course, there might still be differences in the actual inputs that each twin receives, but these are likely much smaller than those between individuals from different families. $\hat{\beta}_1$ can be interpreted as an estimate of the relationship between skills at grade 6, as captured by ENLACE, and future outcomes net of all between-family differences in household, school, and neighborhood inputs, which may arise because differences in parental endowments or preferences.

As discussed in Section 4.1, test scores could capture both cognitive skills as well as noncognitive skills that produce higher test scores, such as motivation or grit. Estimating the effects of test scores on future outcomes net of noncognitive skills is challenging, but we make an attempt to address this issue by dealing with the general component of noncognitive skills. To fix ideas, assume that the production function of test scores can be modeled in the following way:

$$test_{it}^{s} = \alpha_0 + \alpha_1 s_{it}(g_{it-1}) + \alpha_2 g_{it} + \tau_f + e_{it} \tag{8}$$

where $test_{it}^{s}$ is the score of student i in subject area s, s_{it} is the subject-specific skills of student i, g_{it} are general skills of student i, τ_f is a vector of twin fixed effects, and e_{it} is a random disturbance term. This model makes explicit that test scores are produced by both subject-specific and general skills. Notice that under equation 8, subject-specific skills are also a function of general skills, just the same way that cognitive and noncognitive skills are interrelated. While there is not a perfect matching between general and noncognitive skills, the cross-subject correlation in equation 8 is a proxy for the direct effects of noncognitive skills on future test scores. Since ENLACE measures skills in two different subjects, math and literacy, the following specification is estimated:

$$enlace_{it+1}^{math} = \alpha_0 + \alpha_1 enlace_{it}^{math} + \alpha_2 enlace_{it}^{literacy} + \tau_f + \epsilon_{it} \tag{9}$$

In this model, the test score of student i in a particular subject area at grade 9 or 12 is regressed on the score for the same subject area at grade 6 controlling for grade 6 score in the other subject area plus a vector of twin fixed-effects. This strategy controls for the direct effects of general skills such as perseverance and grit on both math and literacy test scores. Notice that equation 9 allows for a differentiated impact of general skills on literacy and numeracy—as we run separate specifications using math and literacy test scores as outcomes, and$\hat{\alpha_2}$ is a subject-specific parameter. Hence, $\hat{\alpha}_1$ in Equation 9 can be interpreted as an estimate of the relationship between *subject-specific* skills in grade 6 and future scores net of family inputs and the direct impact of general skills. Nevertheless, $\hat{\alpha}_1$ in Equation 9 can still be affected by subject-specific traits, such as the motivation to perform in a particular subject area in the test. While the strategy to control for noncognitive skills is not as robust as the one used to control for other inputs (twin fixed effects), it can provide suggestive evidence on the role played by noncognitive skills in determining test scores and future outcomes.

5 Results

5.1 Grade 6 Test Scores and Secondary School Outcomes

Figure 2 plots local means of on-time graduation and test scores in grades 9 and 12 by percentiles of ENLACE grade 6 test scores. Higher test scores in grade 6 are associated with a higher probability of on-time graduation in grades 9 and 12, and, conditional on this, with higher test scores in grades 9 and 12. The differences in outcomes between students in the top and bottom of the grade 6 test score distribution are startling. For example, less than 20% of students in the bottom decile are enrolled in grade 12 six years later, compared to more than 50% of students in the top decile.

Table 2 (Columns 1–4) quantifies the relationships depicted in Figure 2 by regressing graduation and test scores in grades 9 and 12 on grade 6 test scores and on a dummy variable for whether the student is female, using the individual-level data.[17] The results confirm that ENLACE test scores at grade 6 are a strong predictor of lower and upper secondary education outcomes. A 1-SD increase in grade 6 test scores is associated with 8.1-percentage point and 11.6-percentage point increases in the probability of on-time graduation from lower and upper secondary school, respectively. A similar story goes for future test scores. A 1-SD increase in grade 6 test scores is correlated with a 0.63–0.61 SD increase in test scores in grades 9 and 12, conditional on taking the ENLACE exam. All results are statistically significant at the one-percent level.

5.2 Controlling for Family Background

Columns 5–8 in Table 2 report the results of estimating equation 7, which includes a vector of twin fixed effects. Qualitatively, the findings are the same as the ones from the previous specification: Test scores at grade 6 predict on-time graduation from grades 9 and 12, and conditional on this, they also predict test scores in grades 9 and 12. In the four cases, the results have a high statistical and economical significance. However, as expected, the coefficients from the twin fixed-effects specification are smaller than those estimated without taking into account differences in family background (Columns 1–4). A 1-SD higher test score in grade 6 is correlated with a 3.4- and 5.6-percentage point, respectively, higher probability of on-time graduation from grades 9 and 12 (Columns 5–6), and with 0.49- and 0.53-SD, respectively, higher test scores in grades 9 and 12. In other words, taking family background differences into account reduces the magnitude of the estimated correlation between grade 6 test scores and secondary school outcomes by 59% and 48% in the case of on-time graduation from grades 9 and 12, respectively, and by 22% and 14%, respectively, in

[17]For ease of comparison with results from other specifications, the results presented in Columns 1 to 4 in Table 2 are estimated in the sample of twins. In the robustness section, we show that these results are very similar to the ones obtained with the full sample of the ENLACE panel.

the case of test scores in the same grades. The larger reduction in the coefficient of interest in the dropout regressions suggests that family background plays a larger role in explaining on-time graduation from secondary school than future test scores.

The results show that there are significant gender gaps in outcomes. Conditional on the grade 6 score, girls are about 4 percentage points more likely than boys to follow an education trajectory that is free of age-grade distortions, a difference statistically significant at the 1% level (see Row 2 in Columns 5–6). When one looks at test scores, a surprising pattern emerges: The gender gap favors girls until it is reverses by grade 12. Conditional on initial test scores (and staying in school), girls do better on average than boys in grade 9 (by 0.11 SD) but worse by grade 12 (by 0.12 SD). The switch in the sign of the gender coefficient is explained by girls' lower performance in mathematics in grade 12. See results by ENLACE subject in Table A.4 in the Online Appendix and Avitabile and de Hoyos (2018) for a discussion on this issue.

5.3 Accounting for General Skills

Table 3 reports the results of estimating equation 9, the specification designed to control for general skills, in the sample of twins. Both math and literacy scores in grade 6 predict scores in those subject areas in grades 9 and 12, even when controlling for grade 6 scores in the other subject area. The coefficients of interest have a large statistical and economical significance. The magnitude of these point estimates is smaller than the magnitude of those presented in Table 2, as one would expect if test scores were driven by both subject-specific and general skills. Once the direct effects of general skills are removed, the association between 1-SD higher test scores in grade 6 is about 0.25--0.34 SD on test scores in grades 9 and 12 (as opposed to the 0.49--0.53 SD reported in Table 2). The implied estimates for general skills at grade 6 on future test scores are about 0.16--0.21 SD and are highly significant—see Row 2 in Columns 1 to 4 in Table 3. These results suggest that test scores at grade 6 are a strong predictor of future education outcomes, and that their predictive power is driven to a large extent by the link between subject-specific skills over time.

5.4 Grade 12 Test Scores and Post-Secondary School Outcomes

So far the analysis has focused on the effects of grade 6 test scores on secondary school outcomes. The information in the ENLACE panel allows us to implement a robust strategy to control for between-family differences in inputs and for general skills. However, the outcomes studied are limited to the education domain. In this section, we use the ENILEMS-ENLACE panel linking grade 12 test scores with post-secondary school outcomes to present evidence on the relationship test scores and access to university, employment, and wages. A limitation of the ENILEMS-ENLACE panel is that it does not allow for the identification of twins, so we have to rely on conventional and imperfect controls for differences in family background.

Figure 3 shows visual evidence of the simple correlation between post-secondary school outcomes and grade 12 test scores. Table 4 reports the results of regressing post-secondary school outcomes on grade 12 test scores, gender, status of being a graduate of a private high school (a proxy for high family income), and a dummy variable capturing rural/urban resident plus a vector of state and birth year fixed-effects. The results show a strong relationship between test scores and university enrollment. A 1-SD increase in the ENLACE score is associated with an 11-percentage point increase in the probability of university enrollment (statistically significant at the 1% level). Notably, conditional on ENLACE test scores, females are 4 percentage points less likely to be enrolled in university (statistical significance at the 10% level), and graduates of private secondary schools are 10 percentage points more likely to be enrolled in university (statistical significance at the 1% level). In other words, holding end-of-high school test scores constant, there is a gender and family background gap in university enrollment against women and public secondary school graduates.

Among individuals who are not enrolled in higher education, there is no observed association between test scores and employment status or being employed in the formal sector. Among those who are employed, grade 12 test scores have a large and significant relationship with future hourly wages, although not with the probability of being employed in a formal firm. A 1-SD increase in test scores is associated with an increase of about 6% in hourly wages (statistical significance at the 10% level). This hourly wage effect of a 1-SD increase in test scores is marginally smaller than those reported in by Lindqvist and Vestman 2011 using data for Sweden (between 6% and 15%) and significantly smaller than the effects found by Lin et al., 2018 using data for the United States (around 17%). These differences can be explained by a more imperfect labor market in Mexico—compared to the one in Sweden or the United States—or differences in the population subgroups analyzed.[18]

5.5 Robustness Checks

The results shown so far could be compromised by decisions we have made in various dimensions: (i) the sample used for analysis, (ii) the fact that in some specifications we are taking a simple average of math and language test scores, (iii) the estimation method, and (iv) the attrition process observed between the 6th and 12th grades. In this section, we address these issues and show that none of them are a threat to the results:

1. *Sample:* For ease of comparison, all regressions presented in Table 2 are estimated in the sample of twins of the ENLACE panel. That explains why the sample size in Columns 1 to 4 in Table 2 is the same as the sample size in Columns 5 to 8, although the latter includes the twin fixed effects. Using the full ENLACE data set, the sample size in Table A.3 in the Online Appendix varies between 600,000 and 1.8 million observations. However, despite the huge differences in sample size, the results are similar qualitatively and in magnitude.

2. *Average test scores*: For simplicity's sake, Tables 2 and 4 present results using aggregated ENLACE test scores as the simple average of math and literacy scores. To investigate if the reported results are driven by the scores in one of the two subject areas, Tables A.4 and A.5 in the Online Appendix present the results from math and literacy test scores in separate regressions. Although there are small differences between the estimates of each subject, the general findings are similar to the ones using the aggregated ENLACE test scores.

[18] The ENILEMS - ENLACE data set includes only adults from ages 18 to 20, while Lindqvist and Vestman 2011 uses data for men between 30 and 40 years old and Lin et al., 2018 include men and women ages 20 to 50.

3. *Binary outcomes:* The estimation of a linear probability model on binary outcomes (e.g., the probability of on-time graduation) could lead to potential biases in $\hat{\beta}_1$ because of the linear projection of test scores. To address this concern, a probit model is estimated for the binary outcomes included in Tables 2 and 4, using the same vector of independent variables as in the ordinary least squares estimation (except for the fixed effects). The results are available in Tables A.6 and A.7 in the Online Appendix and show similar finding to the ones presented in the previous sections.

4. *Selective attrition*: The estimates of grade 6 test scores on scores in grades 9 and 12 are subject to selective attrition. As is shown in Figure 1, there is a considerable proportion of the population that finished primary school (grade 6) but did not finish grades 9 or 12 on time, and therefore, we cannot observe their test scores. Is this driving part of the results? As we show in the results, attrition is not random, since students with low initial (grade 6) test scores have a lower probability of graduating on time. Despite this, as shown in Panel (b) of Graph 2, there are observations on grade 9 and grade 12 test scores along the full support of the grade 6 test scores distribution. In other words, although it is more likely for students with low test scores at grade 6 to drop out or repeat a grade, it is not a deterministic process, since many students with low scores in grade 6 still graduate on time from grades 9 and 12. Furthermore, conditional on grade 6 test scores, a priori, it is more likely for students in the lower part of the test score distribution—which also corresponds to poorer households—to have unobservable characteristics that make them more prone to drop out or repeat a grade. If this is the case, then the relationship between test scores from grade 6 and grades 9 and 12 becomes steeper than the one presented in Panel (b) of Graph 2. Therefore, the results presented above can be seen as a conservative estimation of the true relationship between scores from grade 6 to grades 9 or 12. A second argument supporting the robustness of the results to selective attrition is that the relationship between grade 6 test scores and scores at grade 9 is very similar to the effects on grade 12 scores, despite the fact that attrition is much larger at grade 12. Finally, we also estimated the relationship between grade 9 and grade 12—where there is considerably

less attrition vis-à-vis the level observed between grades 6 and 12—and found very similar results (see results in Table A.8 in the Online Appendix)

6 Conclusions

The purpose of this paper was to test if a census-based student assessment implemented in a large developing country captures cognitive skills. To do so, we use the Mexican census-based standardized test ENLACE to construct a longitudinal data set tracking students' education trajectories along with test scores through grades 6, 9, and 12. The analysis shows that higher test scores in grade 6 have a large and significant relationship with the student's likelihood of finishing lower and upper secondary school on time; among those who finish, grade 6 test scores are a strong predictor of secondary school test scores. Using a sample of twins to deal with differences by family background, we find that a reduction of 1 SD in test scores in sixth grade reduces by 5.5 percentage points the probability of graduating from secondary school, and, among those who graduated, it reduces their grade 12 test scores by 0.53 SD. Using variation in test scores between the two subject areas included in the test (math and literacy) to control for general (including noncognitive) skills, we present suggestive evidence that ENLACE captures the cognitive skills that it is designed to measure.

The paper also performs a second test of the validity of ENLACE by identifying the short-term relationship between grade 12 test scores and post-secondary outcomes such as university enrollment and labor market outcomes. The results show that grade 12 test scores are a strong predictor of university enrollment and hourly wages. A positive change of 1 SD in test scores at the end of upper secondary is associated with a 11-percentage point increase in the likelihood of enrolling in university, and, conditional on being employed, with a 6% increase in hourly wages.

These findings indicate that, despite their limitations, large-scale standardized tests like ENLACE capture relevant life skills. That said, appropriate design and implementation matter for the quality of large-scale student assessments. Also, as Neal (2011) warns, assigning two competing goals to one test can backfire. Notably, the objective of assessing the evolution of learning in a school system over time is in conflict with the objective of producing student learning measures to reward teachers and principals performance. The ability to compare test scores over time requires keeping similar exam questions on the test, but doing so facilitates teaching to the test, a practice that is more likely to emerge when test scores are linked to financial incentives. Setting and keeping non-competing policy objectives and coupling it with careful monitoring of test implementation are necessary conditions for reaping the benefits of standardized testing in education systems.

The results of this study also shed light on how disadvantages at early stages of education can have important and persistent implications for future education and labor market outcomes. The concept that learning begets learning is corroborated by evidence presented here that suggests that a low learning outcome in sixth grade can have a negative consequence in labor incomes 10 years later. The lower performance in sixth grade works its way forward in time, signaling lower chances of completing upper secondary school. If the student graduates, the individual's reduced learning outcomes diminish the probability of starting university and, among those who work, reduce the individual's wages. This should make a failing mark at the end of primary school—or earlier—a trigger to provide strong attention and support to the student. By identifying students who need more support early in their education trajectories, large-scale student assessments can help create an education system that promotes equal opportunities and social mobility rather than one that simply replicates or even exacerbates existing inequalities.

References

Akyol, P., Krishna, K., and Wang, J. (2018). Taking pisa seriously: How accurate are low stakes exams? NBER Working Papers 24930, National Bureau of Economic Research, Inc.

Avitabile, C. and de Hoyos, R. (2018). The heterogeneous effect of information on student performance: Evidence from a randomized control trial in mexico. *Journal of Development Economics*, 135:318 – 348.

Braun, H. and von Davier, M. (2017). The use of test scores from large-scale assessment surveys: psychometric and statistical considerations. *Journal of Human Resources*.

CAF (2016). *RED 2016. Más habilidades para el trabajo y la vida: los aportes de la familia, la escuela, el entorno y el mundo laboral.* CAF, Bogotá.

Chang, E. and Padilla-Romo, M. (2019). The effects of local violent crime on high-stakes tests. Working Papers 2019-03, University of Tennessee, Department of Economics.

Cheng, X. and Gale, C. (2014). List of national learning assessments. retrieved at https://www.epdc.org/education-data-research/list-national-learning-assessments.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., and Yagan, D. (2011). How does your kindergarten classroom affect your earnings? evidence from project star *. *The Quarterly Journal of Economics*, 126(4):1593–1660.

Cunha, F., Heckman, J. J., and Schennach, S. M. (2010). Estimating the technology of cognitive and noncognitive skill formation. *Econometrica*, 78(3):883–931.

Currie, J. and Thomas, D. (2001). Early test scores, school quality and ses: Longrun effects on wage and employment outcomes. *Research in Labor Economics*, 20.

de Hoyos, R., Garcia-Moreno, V. A., and Patrinos, H. A. (2017). The impact of an accountability intervention with diagnostic feedback: Evidence from mexico. *Economics of Education Review*, 58:123–140.

Duckworth, A. L. and Seligman, M. E. (2005). Self-discipline outdoes iq in predicting academic performance of adolescents. *Psychological Science*, 16(12):939–944. PMID: 16313657.

Dustan, A., de Janvry, A., and Sadoulet, E. (2017). Flourish or fail?: The risky reward of elite high school admission in mexico city. *Journal of Human Resources*, 52(3):756–799.

Estrada, R. (2019). Rules versus discretion in public service: Teacher hiring in mexico. *Journal of Labor Economics*, 37(2):545–579.

Estrada, R. and Gignoux, J. (2017). Benefits to elite schools and the expected returns to education: Evidence from mexico city. *European Economic Review*, 95:168 – 194.

Figlio, D. and Loeb, S. (2011). School accountability. In Hanushek, E., Machin, S., and Woessmann, L., editors, *Handbook of The Economics of Education*, volume 3, chapter 08, pages 383–421. Elsevier, 1 edition.

Finn, B. (2015). Measuring motivation in low-stakes assessments. *ETS Research Report Series*, 2015(2):1–17.

Gneezy, U., List, J. A., Livingston, J. A., Qin, X., Sadoff, S., and Xu, Y. (2019). Measuring success in education: The role of effort on the test itself. *American Economic Review: Insights*, 1(3):291–308.

Heckman, J. J. and Kautz, T. (2012). Hard evidence on soft skills. *Labour Economics*, 19(4):451 – 464. European Association of Labour Economists 23rd annual conference, Paphos, Cyprus, 22-24th September 2011.

Heckman, J. J., Stixrud, J., and Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics*, 24(3):411–482.

Koretz, D. (2017). *The Testing Charade: Pretending to Make Schools Better*. University of Chicago Press.

Koretz, D. M. and Barron, S. I. (1998). *The Validity of Gains in Scores on the Kentucky Instructional Results Information System (KIRIS).* ERIC.

Laajaj, R. and Macours, K. (2019). Measuring skills in developing countries. *Journal of Human Resources.*

Lavy, V., Ebenstein, A., and Roth, S. (2014). The Impact of Short Term Exposure to Ambient Air Pollution on Cognitive Performance and Human Capital Formation. NBER Working Papers 20648, National Bureau of Economic Research, Inc.

Lin, D., Lutter, R., and Ruhm, C. J. (2018). Cognitive performance and labour market outcomes. *Labour Economics*, 51:121 – 135.

Lin, M., Bumgarner, E., and Chatterji, M. (2014). Understanding validity issues in international large scale assessments. *Quality Assurance in Education.*

Lindqvist, E. and Vestman, R. (2011). The Labor Market Returns to Cognitive and Noncognitive Ability: Evidence from the Swedish Enlistment. *American Economic Journal: Applied Economics*, 3(1):101–128.

Murnane, R. J., Willett, J. B., and Levy, F. (1995). The growing importance of cognitive skills in wage determination. *The Review of Economics and Statistics*, 77(2):251–266.

Neal, D. (2011). The design of performance pay ineducation. In Hanushek, E., Machin, S., and Woessmann, L., editors, *Handbook of The Economics of Education*, volume 4, chapter 06, pages 495–550. Elsevier, 1 edition.

OECD (2001). *Knowledge and Skills for Life: First Results from PISA 2000.* OECD Publishing, Paris.

Rose, H. (2006). Do gains in test scores explain labor market outcomes? *Economics of Education Review*, 25(4):430–446.

Salardi, P. and Michaelsen, M. (2019). Violence, psychological stress and educational performance during the war on drugs in mexico. *Journal of Development Economics*, page 102387.

World Bank (2018). *World Development Report 2018: Learning to Realize Education's Promise*. World Bank, Washington, DC.

Tables

Table 1: ENLACE: Student Take-Up (percentage)

	(1) Primary 2007	(2) Lower secondary 2010	(3) Upper secondary 2013
National	89.6	86.5	89.4
Michoacán	48.7	33.9	26.0
Oaxaca	65.1	0.7	92.9
National without Michoacán and Oaxaca	92.6	93.3	93.0

Notes: Primary school take-up includes grades 3 to 6, lower secondary grades 7 to 9, and upper secondary grade 12. Source: SEP.

Table 2: OLS – Grade 6 Test Scores and Secondary School Outcomes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Graduation		Score		Graduation		Score	
VARIABLES	Grade 9	Grade 12	Grade 9	Grade 12	Grade 9	Grade 12	Grade 9	Grade 12
Grade 6 score	0.0811***	0.116***	0.631***	0.611***	0.0335***	0.0567***	0.489***	0.530***
	(0.00298)	(0.00329)	(0.00687)	(0.0101)	(0.00515)	(0.00620)	(0.0148)	(0.0236)
Girl	0.0273***	0.0388***	0.0554***	-0.144***	0.0396***	0.0456***	0.113***	-0.119***
	(0.00588)	(0.00673)	(0.0128)	(0.0184)	(0.00796)	(0.00950)	(0.0208)	(0.0361)
Observations	20,252	20,252	15,494	8,108	20,252	20,252	15,494	8,108
R-squared	0.038	0.059	0.381	0.331	0.805	0.810	0.861	0.885
Twins FE	No	No	No	No	Yes	Yes	Yes	Yes
Mean Dep. Var.	0.764	0.400	0.0101	-0.0380	0.764	0.400	0.0101	-0.0380

Notes: (1) The table displays results of the estimation of equations 6 and 7. (2) Dependent variable is on-time graduation and ENLACE test scores in grades 9 and 12. Graduation is measured as ENLACE take-up. ENLACE test score is the mean of the mathematics and literacy test scores. (3) Sample: Twins (students in the same school in grade 6, with identical last names and birth date) who took the ENLACE exam in grade 6 in 2007. (4) Data: ENLACE panel. (4) Robust standard errors are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table 3: OLS – Grade 6 Test Scores and Secondary School Test Scores by Subject

	(1)	(2)	(3)	(4)
	Mathematics		Spanish	
VARIABLES	Grade 9	Grade 12	Grade 9	Grade 12
Grade 6 same subject score	0.254***	0.339***	0.309***	0.310***
	(0.0175)	(0.0241)	(0.0155)	(0.0271)
Grade 6 other subject score	0.199***	0.157***	0.196***	0.208***
	(0.0166)	(0.0256)	(0.0154)	(0.0261)
Girl	0.0208	-0.307***	0.187***	0.152***
	(0.0235)	(0.0390)	(0.0217)	(0.0392)
Observations	15,494	8,108	15,494	8,110
R-squared	0.819	0.873	0.849	0.850
Twins FE	Yes	Yes	Yes	Yes
Mean Dep. Var.	-0.00162	-0.0558	0.0208	-0.00699

Notes: (1) The table displays the estimation of equation 9. (2) Dependent variable is mathematics and literacy ENLACE test scores in grades 9 and 12. (3) Sample: Twins (students in the same school in grade 6, with identical last names and birth date) who took the ENLACE exam in grade 6 in 2007. (3) Data: ENLACE panel. (4) Robust standard errors are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

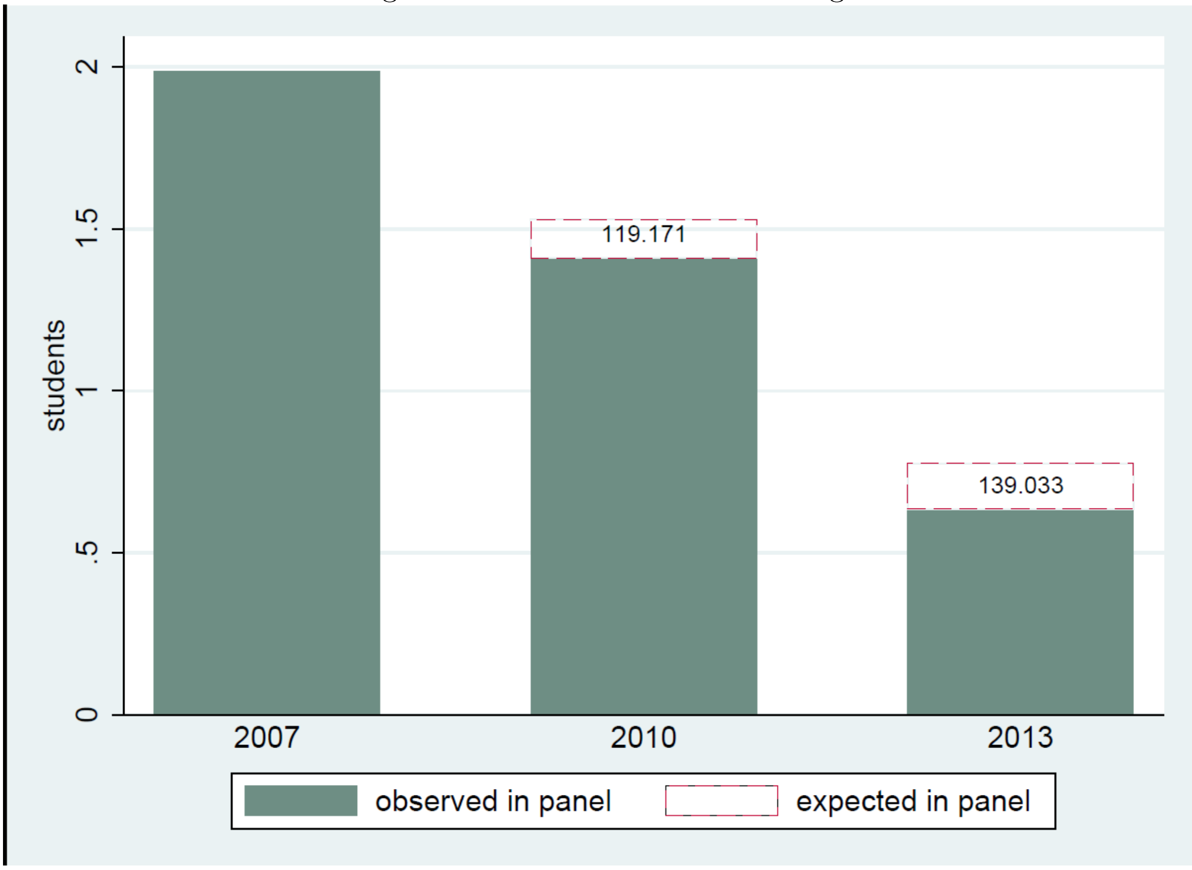
Table 4: OLS – Grade 12 Test Scores and Post-Secondary School Outcomes

	(1)	(2)	(3)	(4)
VARIABLES	University Student	Employed	ln hourly wage	Formal firm
Grade 12 score	0.111***	-0.0132	0.0582**	-0.00764
	(0.00887)	(0.0192)	(0.0244)	(0.0197)
Girl	-0.0372***	-0.266***	-0.0633	0.00402
	(0.0144)	(0.0283)	(0.0394)	(0.0321)
Private secondary school	0.101***	-0.103**	0.139**	-0.0697
	(0.0181)	(0.0425)	(0.0688)	(0.0439)
Urban resident	0.265***	0.0122	0.108*	0.112**
	(0.0270)	(0.0342)	(0.0560)	(0.0468)
Observations	3,705	1,162	1,020	1,020
R-squared	0.173	0.122	0.106	0.072
Sample	All	Out of school	Employed	Employed
Birth Year Dummies	Yes	Yes	Yes	Yes
Birth State Dummies	Yes	Yes	Yes	Yes
Clusters	1778	822	706	706
Mean Dep. Var.	0.630	0.578	2.821	0.396

Notes: (1) The table displays the results of the estimation of Equation 6. (2) The dependent variables are post-secondary school outcomes: a dummy indicating enrollment in university (column 1), a dummy indicating if employed (column 2), ln of hourly wage (column 3), a dummy for being employed in a formal firm (column 4). (3) Outcomes are measured in the ENILEMS survey at ages 18 to 20 in the third quarter of 2010. ENLACE test scores come from the years 2008, 2009 and 2010. (4) Data: ENILEMS-ENLACE panel. (5) Robust standard errors are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

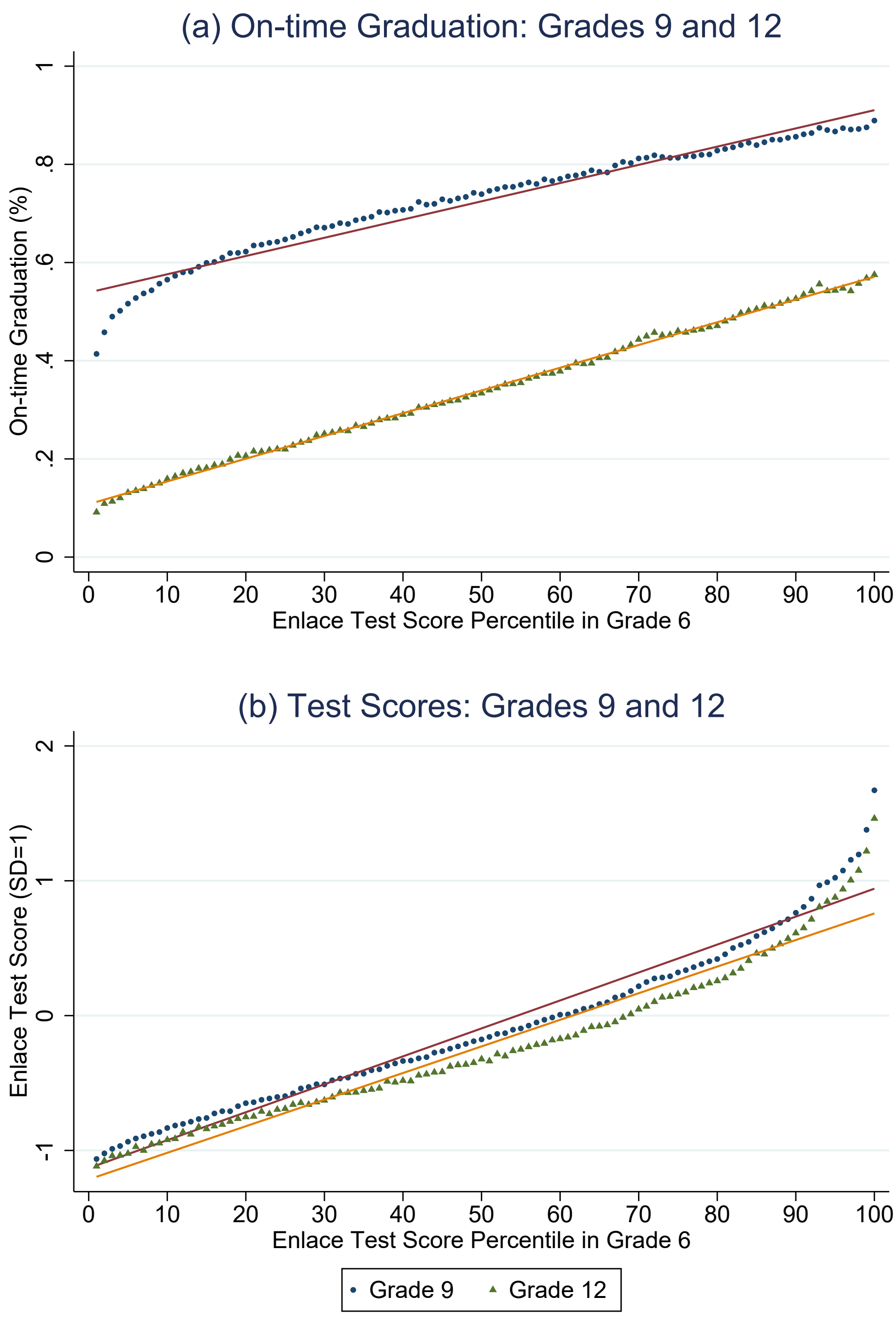
Figures

Figure 1: ENLACE Panel Matching



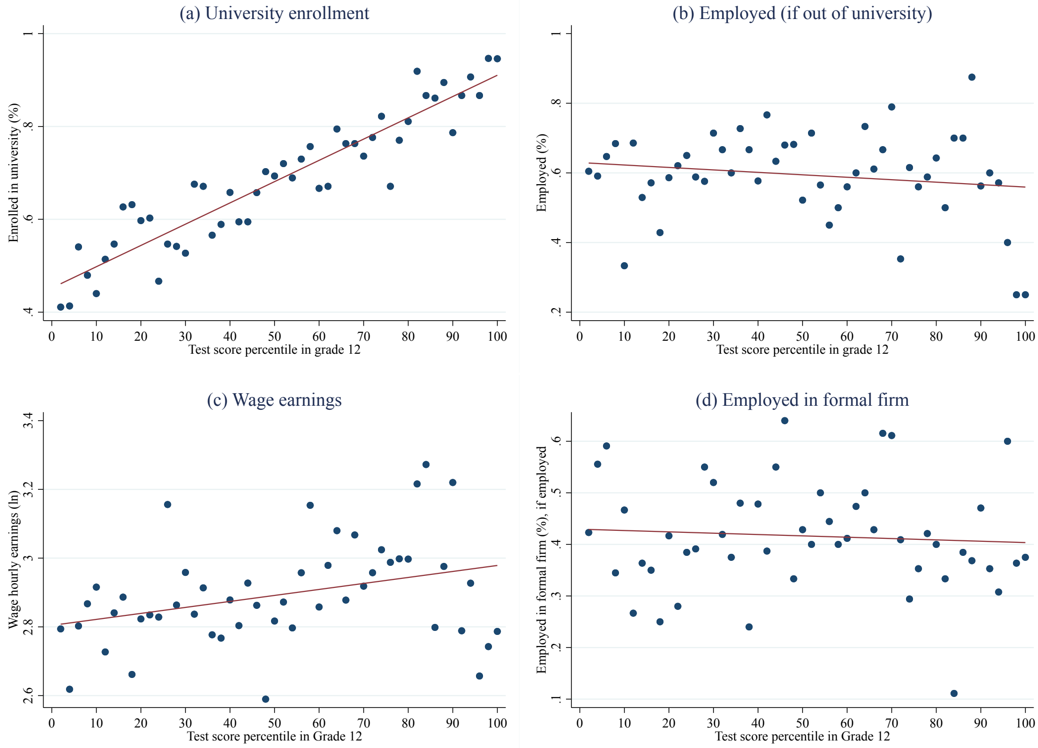
Notes: (1) The graph presents the number of students in the ENLACE panel in 2007, 2010 and 2013 and the expected number of observations given the estimated school trajectories in secondary school using the Formato 911 data. (2) Sample: Students who took the ENLACE exam in grade 6 in 2007. (3) Data: ENLACE panel and Formato 911 data.

Figure 2: Grade 6 Test Scores and Secondary School Outcomes



Notes: (1) The graph plots local means of secondary school outcomes by ENLACE test score percentile in grade 6 in 2007. The solid line shows a linear fit estimated using the grouped data. Panel (a) reports the probability of on-time graduation from grades 9 and 12, proxied by siting in the Enlace exam in those grades. Panel (b) reports Enlace test scores in grades 9 and 12 (normalised with mean 0 and SD 1) conditional on taking the Enlace exam in 2010 and 2013. (2) Sample: Students who took the ENLACE exam in grade 6 in 2007. (3) Data: ENLACE panel.

Figure 3: Grade 12 Test Scores and Post-Secondary School Outcomes



Notes: (1) The graph plots local means of post-secondary school outcomes by ENLACE test score ventile in grade 12. The solid line shows a linear fit estimated using the grouped data. Panel (a) reports the probability of university enrollment; (b) reports the probability of being employed conditional on not being enrolled in college; and (c) and (d) report, respectively, the logarithm of the hourly wage and the probability of working in a formal firm conditional in both cases on being employed. (2) Outcomes are measured in the ENILEMS survey at ages 18 to 20 in the third quarter of 2010. ENLACE test scores come from the years 2008, 2009 and 2010. (3) Data: ENILEMS-ENLACE panel.

Appendix

Table A.1: ENLACE Panel – Means and Standard Deviations

VARIABLES	(1) Grade 6	(2) Grade 9	(3) Grade 12	(4) Twins Grade 6	(5) Survey Grade 6
Grade 6 score	0.0209	0.162	0.398	0.0407	0.213
	(0.991)	(0.975)	(0.942)	(0.994)	(1.038)
Grade 9 enrollment	0.730	1	1	0.765	0.779
	(0.444)	(0)	(0)	(0.424)	(0.415)
Grade 12 enrollment	0.345	0.473	1	0.400	0.372
	(0.475)	(0.499)	(0)	(0.490)	(0.483)
Girl	0.494	0.515	0.545	0.536	0.492
	(0.500)	(0.500)	(0.498)	(0.499)	(0.500)
Private primary school	0.0848	0.0988	0.124	0.0959	0.109
	(0.279)	(0.298)	(0.330)	(0.295)	(0.311)
Mother has lower secondary					0.587
					(0.492)
Father has lower secondary					0.628
					(0.483)
Mother is white collar					0.135
					(0.341)
Father is white collar					0.245
					(0.430)
Observations	1,878,931	1,371,437	648,018	20,252	5,677

Notes: (1) The table shows the mean and standard deviations of all students in the ENLACE panel observed in 2007 (Column 1), 2010 (Column 2) and 2013 (Column 3). Column 4 reports statistics for the sample of students identified as twins in grade 6. Column 5 displays additional variables from student and parents surveys that were applied to a sample of ENLACE takers. (2) Sample: Students who took the ENLACE exam in grade 6 in 2007. (3) Data: ENLACE panel.

Table A.2: ENILEMS-ENLACE panel – Means and Standard Deviations

VARIABLES	(1) All	(2) University	(3) Out of university	(4) Employed
Grade 12 score	0.212	0.374	-0.0634	0.113
	(0.854)	(0.856)	(0.778)	(0.808)
University sudent	0.630	1	0	0.436
	(0.483)	(0)	(0)	(0.496)
Employed	0.379	0.262	0.578	1
	(0.485)	(0.440)	(0.494)	(0)
Girl	0.565	0.546	0.596	0.503
	(0.496)	(0.498)	(0.491)	(0.500)
Private secondary school	0.175	0.203	0.128	0.133
	(0.380)	(0.402)	(0.335)	(0.340)
Urban resident	0.848	0.912	0.739	0.814
	(0.359)	(0.283)	(0.439)	(0.390)
Age	19.18	19.16	19.21	19.23
	(0.701)	(0.688)	(0.723)	(0.697)
Observations	3,718	2,551	1,167	1,385

Notes: (1) The table displays the mean and standard deviations of several characteristics of all students matched in the ENILEMS-ENLACE (Column 1), students that reported to be in college (Column 2), out of college (Column 3) and employed (Column 4). (2) Sample: Respondents to the ENILEMS survey who were matched to their ENLACE results in 2008, 2009, or 2010. (3) Data: ENILEMS-ENLACE panel.

Online Appendix

Table A.1: Twins and Singletons in ENLACE PANEL: Means and Standard Deviations

	(1)		(2)		T-test
	0		1		Difference
Variable	N	Mean/SE	N	Mean/SE	(1)-(2)
Enlace Score Spanish Grade 6	1942306	512.842	20982	515.377	-2.536***
		(0.075)		(0.719)	
Enlace Score Math Grade 6	1942246	514.237	20982	516.474	-2.238***
		(0.079)		(0.762)	
Enlace taker Grade 9	1943583	0.702	20995	0.742	-0.040***
		(0.000)		(0.003)	
Enlace taker Grade 12	1943583	0.317	20995	0.368	-0.052***
		(0.000)		(0.003)	
Girl	1943583	0.492	20995	0.536	-0.043***
		(0.000)		(0.003)	
Private School Grade 6	1943583	0.083	20995	0.095	-0.012***
		(0.000)		(0.002)	

Notes: (1) This table presents sample size, mean and standard deviation of twins (Column 1) and singletons in the ENLACE panel (Column 2). (2) Column 3 reports the t-tests differences in the means across the groups, and ***, **, and * indicate significance at the 1, 5, and 10 percent critical level. (3) Sample: Students who took the ENLACE exam in grade 6 in 2007. (4) Data: ENLACE panel.

Table A.2: ENILEMS-ENLACE: Matched and Not Matched Samples

Variable	N	(1) 0 Mean/SE	N	(2) 1 Mean/SE	T-test Difference (1)-(2)
Girl	3323	0.542 (0.009)	3781	0.565 (0.008)	-0.024**
Age	3323	19.266 (0.013)	3781	19.188 (0.011)	0.079***
Urban resident	3323	0.889 (0.005)	3781	0.882 (0.005)	0.007
Schooling	3323	3.506 (0.025)	3781	3.505 (0.026)	0.001
University student	3323	0.648 (0.008)	3781	0.684 (0.008)	-0.036***
Private secondary school	3323	0.247 (0.007)	3781	0.176 (0.006)	0.070***
Upper secondary school GPA	3319	83.668 (0.118)	3777	84.207 (0.109)	-0.539***
Employed	3244	0.373 (0.008)	3718	0.373 (0.008)	0.000
Unemployed	3244	0.056 (0.004)	3718	0.048 (0.004)	0.007
No study nor work	3244	0.156 (0.006)	3718	0.124 (0.005)	0.032***
Wage earner	1210	0.772 (0.012)	1385	0.771 (0.011)	0.001
Hourly wage	900	24.511 (0.664)	1021	22.390 (0.683)	2.121**

Notes: (1) The table presents sample size, mean and standard deviation of individuals not matched in the sample (Column 1) and individuals matched in the sample (Column 2). (2) Column 3 reports the t-tests differences in the means across the groups, and ***, **, and * indicate significance at the 1, 5, and 10 percent critical level. (3) Sample: Respondents to the ENILEMS survey who were matched to their ENLACE results in 2008, 2009, or 2010. (4) Data: ENILEMS-ENLACE panel.

Table A.3: OLS – Grade 6 Test Scores and Secondary School Outcomes: Full Sample

	(1)	(2)	(3)	(4)
	Graduation		Score	
VARIABLES	Grade 9	Grade 12	Grade 9	Grade 12
Grade 6 score	0.103***	0.130***	0.617***	0.623***
	(0.000313)	(0.000330)	(0.000737)	(0.00111)
Girl	0.0357***	0.0384***	0.0783***	-0.139***
	(0.000633)	(0.000673)	(0.00137)	(0.00203)
Observations	1,878,931	1,878,931	1,371,437	647,583
R-squared	0.057	0.078	0.371	0.344
Mean Dep. Var.	0.729	0.345	0.0112	0.0103

Notes: (1) The table displays results of the estimation of Equation 6. (2) Dependent variable is on-time graduation and ENLACE test score in grades 9 and 12. Graduation is measured as ENLACE take-up. ENLACE test score is the mean of the mathematics and literacy test scores. (3) Sample: Students who took the ENLACE exam in grade 6 in 2007. (4) Data: ENLACE panel. (5) Robust standard errors are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table A.4: OLS – Grade 6 Test Scores in Mathematics and Literacy and Secondary School Outcomes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Graduation		Score		Graduation		Score	
VARIABLES	Grade 9	Grade 12	Grade 9	Grade 12	Grade 9	Grade 12	Grade 9	Grade 12
Grade 6 literacy score	0.0350***	0.0512***	0.414***	0.417***				
	(0.00473)	(0.00563)	(0.0141)	(0.0245)				
Grade 6 math score					0.0200***	0.0414***	0.358***	0.418***
					(0.00474)	(0.00565)	(0.0154)	(0.0219)
Girl	0.0377***	0.0441***	0.181***	0.143***	0.0430***	0.0507***	0.0527**	-0.282***
	(0.00799)	(0.00951)	(0.0221)	(0.0399)	(0.00796)	(0.00951)	(0.0236)	(0.0392)
Observations	20,252	20,252	15,494	8,110	20,252	20,252	15,494	8,108
R-squared	0.805	0.809	0.845	0.846	0.804	0.809	0.814	0.871
Twins FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean Dep. Var.	0.764	0.400	0.0208	-0.00699	0.764	0.400	-0.00162	-0.0558

Notes: (1) The table displays results from OLS regressions of graduation and literacy (mathematics) ENLACE test scores in grades 9 and 12 on literacy (mathematics) test scores in grade 6. (2) Sample: Twins (students in the same school in grade 6, with identical last names and birth date) who took the ENLACE exam in grade 6 in 2007. (4) Data: ENLACE panel. (5) Robust standard errors are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table A.5: OLS – Grade 12 Test Scores in Mathematics and Literacy and Post-Secondary School Outcomes

VARIABLES	(1) College Student	(2) Employed	(3) ln hourly wage	(4) Formal firm	(5) College Student	(6) Employed	(7) ln hourly wage	(8) Formal firm
Literacy score	0.122***	-0.0122	0.0622***	-0.0114				
	(0.00762)	(0.0172)	(0.0217)	(0.0176)				
Mathematic score					0.106***	-0.000756	0.0418**	-0.00583
					(0.00685)	(0.0158)	(0.0193)	(0.0157)
Girl	-0.0428***	-0.257***	-0.0721*	0.00369	0.0104	-0.258***	-0.0482	-0.000187
	(0.0140)	(0.0279)	(0.0384)	(0.0315)	(0.0144)	(0.0286)	(0.0384)	(0.0320)
Private secondary school	0.0992***	-0.103**	0.134*	-0.0688	0.106***	-0.103**	0.144**	-0.0705
	(0.0183)	(0.0424)	(0.0690)	(0.0439)	(0.0180)	(0.0426)	(0.0689)	(0.0439)
Urban resident	0.274***	0.00918	0.108*	0.112**	0.272***	0.00831	0.110*	0.112**
	(0.0275)	(0.0338)	(0.0559)	(0.0468)	(0.0276)	(0.0340)	(0.0564)	(0.0467)
Observations	3,709	1,165	1,020	1,020	3,709	1,165	1,020	1,020
R-squared	0.147	0.121	0.106	0.072	0.143	0.120	0.103	0.072
Birth Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Birth State Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Clusters	1779	822	706	706	1779	822	706	706
Mean Dep. Var.	0.630	0.578	2.821	0.396	0.630	0.578	2.821	0.396

Notes: (1) The table displays results from OLS regressions of post-secondary school outcomes on literacy and mathematics ENLACE test scores in grade 12. (2) Sample: Respondents to the ENILEMS survey who were matched to their ENLACE results in 2008, 2009, or 2010. (3) Data: ENILEMS-ENLACE panel. (4) Robust standard errors are reported in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Table A.6: Probit – Grade 6 Test Scores and Secondary School Graduation

	(1)	(2)	(3)	(4)
	Graduation			
	Coefficients		Marginal Effects	
VARIABLES	Grade 9	Grade 12	Grade 9	Grade 12
Grade 6 score	0.273***	0.313***	0.0851***	0.118***
	(0.0172)	(0.0121)	(0.00540)	(0.00464)
Girl	0.0854***	0.106***		
	(0.0230)	(0.0212)		
Observations	20,252	20,252	20,252	20,252
Clusters	8597	8597		
Mean	0.764	0.400		

Notes: (1) The table displays results from Probit regressions of the probability of being enrolled in grades 9 and 12 given grade 6 tests scores and a set of socioconomic variables. (2) Sample: ENLACE Contexto survey respondants to parental questionnaire. (3) Data: ENLACE panel. (4) Robust standard errors are reported in parentheses *** p<0.01, ** p<0.05, * p<0.1.

Table A.7: Probit – Grade 12 Test Scores and Post-Secondary School Outcomes

	(1)	(2)	(3)	(4)	(5)	(6)
	Coefficients			Marginal effects		
VARIABLES	College Student	Employed	Formal firm	College Student	Employed	Formal firm
Grade 12 score	0.434***	-0.0245	-0.0276	0.171***	-0.00763	-0.00973
	(0.0281)	(0.0500)	(0.0481)	(0.0114)	(0.0156)	(0.0169)
Girl	-0.0431	-0.715***	-0.00899	-0.0170	-0.223***	-0.00317
	(0.0451)	(0.0795)	(0.0795)	(0.0179)	(0.0222)	(0.0281)
Private secondary	0.224***	-0.260**	-0.109	0.0882***	-0.0810**	-0.0383
	(0.0632)	(0.110)	(0.114)	(0.0249)	(0.0340)	(0.0406)
Urban resident	0.743***	0.0434	0.336***	0.293***	0.0135	0.119***
	(0.0757)	(0.0943)	(0.125)	(0.0271)	(0.0301)	(0.0380)
Observations	3,718	1,167	1,021	3,718	1,167	1,021
Clusters	1781	824	706			
Mean Dep. Var.	0.630	0.578	0.396			

Notes: (1) The table displays results from Probit regressions of the probability of being enrolled in university, being employed and being employed in a formal firm on grade 12 test scores and a set of socioconomic variables. (2) Outcomes are measured in the ENILEMS survey at ages 18 to 20 in the third quarter of 2010. ENLACE test scores come from the years 2008, 2009 and 2010. (3) Data: EMILEMS-ENLACE panel. (4) Robust standard errors are reported in parentheses *** p<0.01, ** p<0.05, * p<0.1.

Table A.8: OLS – Grade 9 Test Scores and Grade 12 Outcomes

	(1)	(2)	(3)	(4)
VARIABLES	Graduation	Score	Graduation	Score
Grade 9 score	0.141***	0.683***	0.0790***	0.510***
	(0.00375)	(0.00958)	(0.00678)	(0.0207)
Girl	0.0305***	-0.134***	0.0379***	-0.124***
	(0.00777)	(0.0175)	(0.0119)	(0.0353)
Observations	15,494	8,108	15,494	8,108
R-squared	0.082	0.411	0.826	0.890
Twins FE	No	No	Yes	Yes
Mean Dep. Var.	0.523	-0.0380	0.523	-0.0380

Notes: (1) The table displays results from OLS regressions of enrollemnt and ENLACE test scores in grade 12 on ENLACE test scores in grade 9. (2) Sample: Twins (students in the same school in grade 6, with identical last names and birth date) who took the ENLACE exam in grade 9 in 2010. (4) Data: ENLACE panel. (5) Robust standard errors are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1.