Long-term Impacts of Alternative Approaches to Increase Schooling: Evidence from an Experimental Scholarship Program in Cambodia

Felipe Barrera-Osorio
Andreas de Barros
Deon Filmer*

Working Paper
This version: May 2018

Abstract. We use a randomized experiment to investigate the long-run effects of a primary school scholarship program in rural Cambodia. In 2008, fourth-grade students in 207 randomly assigned schools (103 treatment, 104 control) received scholarships based on the student's academic performance in math and language or on their level of poverty. Three years after the program's inception, an evaluation showed that both types of scholarship recipients progressed more in their education than non-recipients; however, only merit-based scholarships led to improvements on cognitive tests. This new study reports impacts after nine years of the program on educational attainment, cognitive skills, socioemotional skills, labor outcomes, socioeconomic perception and well-being of, on average, 21-year-old individuals. We find that both types of scholarships led to higher long-run educational attainment (about 0.21-0.27 grade levels), but only merit-based scholarships led to improvements in cognitive skills (0.11 standard deviations), greater well-being (0.18 standard deviations), and employment probability (3.4 percentage points). Neither type of scholarship increased socioemotional skills, suggesting that boosting schooling or learning (or even well-being) through demand-side programs is not sufficient to increase socioemotional outcomes.

JEL classification codes: I21; I24; I28; O10

Keywords: Cambodia; cognitive skills; education; labor outcomes; long term; merit-based targeting; poverty-based targeting; randomization; scholarships; socioemotional skills

* Barrera-Osorio: Graduate School of Education, Harvard University (felipe_barrera-osorio@gse.harvard.edu); de Barros: Graduate School of Education, Harvard University (adebarros@g.harvard.edu); Filmer: World Bank (dfilmer@worldbank.org). We are grateful for financial support from the World Bank's Strategic Research Program Trust Fund. Alice Danon provided outstanding research assistance. For helpful comments, we thank Maria Bertling, Theresa Betancourt, Monnica Chan, Olivia Chi, Mark Chin, Jishnu Das, David Deming, Alejandro Ganimian, Sibylla Leon Guerrero, Rema Hanna, Andrew Ho, Whitney Kozakowski, Guilherme Lichand, Sophie Litschwartz, Dana McCoy, Jonathan Mijs, Karthik Muralidharan, Charles Nelson, Gautam Rao, Margaret Sheridan, Abhijeet Singh, and Martin West. The usual disclaimers apply. The findings, interpretations, and conclusions expressed in this paper are those of the authors and do not necessarily represent the views of the World Bank, its Executive Directors, or the governments they represent. Harvard University IRB Protocol Title "From Schooling to Young Adulthood", Number IRB16-1518.

1. Introduction

How does additional schooling impact long-term life outcomes? According to the canonical human capital model, labor markets remunerate the skills acquired during the education process (Becker, 2009). According to a signaling model (Arrow, 1973; Spence, 1973), education provides the market with a signal of individuals' higher abilities; as a result, the market pays for these skills. Both models predict positive effects from investment in education. In contrast to this view, emerging research is proposing that several countries have significantly increased schooling of their populations, but students are learning little and the markets are not rewarding education (Pritchett, 2013; The World Bank, 2017). However, this debate for developing countries is informed by only few studies that can isolate the impacts of schooling and skills.[1] Our paper aims to contribute to this evidence by presenting the causal effect of an scholarship program —which induced more schooling— on cognitive, socioemotional, socioeconomic status and well-being[2], and labor outcomes in a group of 21-year-old individuals who received the scholarship nine years earlier, in Cambodia.

Our study setup is the following. In 2008, 207 schools in Cambodia were randomly allocated between two treatment arms (103 schools) and a control group (104 schools). In about half of the treatment schools, students in grade four received a scholarship based on merit —students were selected using a baseline test of math and language skills—and fourth-graders in the remaining treatment schools received a scholarship based on poverty—students were selected using a poverty index, based on household and family socioeconomic characteristics. Scholarships were given to recipients for three years (i.e. until the completion of primary school), conditional on continued school participation and basic performance standards. A first follow-up, three years after the inception of the program, showed two main effects: higher school progression for individuals in both types of scholarship recipients (compared with non-recipients), and effects on cognitive outcomes (as measured by a math test and a test of working memory) for the merit-based scholarship only (Barrera-Osorio & Filmer, 2016). In this paper we report results from data collected in 2016—nine years after the beginning of the program—from a subsample of the

[1] Few well-identified studies on the causal impacts of education exist for developing countries; important exceptions are Duflo, Dupas and Kremer (2017); Parker & Vogl (2018); Ozier (2016); Jakiela, Miguel, & te Velde (2015) and Friedman et al. (2011).

[2] From this point on, we will refer to both socioeconomic status and well-being as "well-being", for the sake of brevity.

original study participants to present evidence of the effects of the scholarship on several long-term outcomes. To this end, we collected data on a large array of measures in the domains of cognitive and socioemotional skills, well-being, labor market and other socioeconomic outcomes.

We present causal evidence to address three questions. First, what are the long-term effects of the scholarship on labor market and well-being outcomes? Based on the three-year follow-up, we know that both scholarships induced more schooling for treated individuals (Barrera-Osorio & Filmer, 2016). Theoretically, this additional education affects labor outcomes and well-being of individuals. Second, what are the long-term effects of the program on cognitive and socioemotional outcomes? Heckman and Kautz (2014) show that socioemotional skills (which are also sometimes referred to as "non-cognitive" skills) are important determinants of labor outcomes in the long-run. We can investigate the impacts of (exogenously induced) additional exposure to schooling on these cognitive and socioemotional outcomes. Third, what are the long-term effects of the scholarship on socioemotional skills? This question allows an investigation of whether socioemotional outcomes are co-produced (complement) with cognitive outcomes. We can pursue the answer to this question because only the merit-based scholarship induced changes in cognitive tests after the first three years of the intervention; therefore, we can test whether we observe effects on socioemotional skills for this group only, for both treatment groups, or for neither group.

Three other demand-side incentive programs are of particular relevance given their similar designs and scope. First, Friedman et al. (2011) and Jakiela, Miguel, & te Velde (2015) present experimental evidence on the effects of the a Kenyan merit-scholarship program for sixth-grade girls, nine years post-intervention. The studies find that short-term impacts on educational attainment and cognitive skill (initially reported in Kremer, Miguel, & Thornton, 2009) result in greater female empowerment, improved political knowledge and attitudes in young adulthood (finding weaker evidence for effects on political behaviors). Second, Barham et. al. (2014) evaluate the long run effects of a conditional cash transfer program targeted to poor families in Nicaragua. They find that boys who were 9-12 at the time of the program attained about half a year more schooling when they were 19-22 than boys in a comparison group, and subsequently had better labor market outcomes (the difference for girls was not statistically different from zero between the treatment and comparison groups). Third, Duflo, Dupas, & Kremer (2017) evaluate the long-term effects of a secondary school scholarship program, in Ghana. This randomized evaluation

finds that the program delayed fertility and marriage, improved educational attainment, cognitive skill, and reproductive and health behaviors, and led to heterogeneous effects on earnings. Our study builds on these: The Kenyan study does not report impacts on earnings, and the Ghana and Nicaragua studies investigate few impacts on socioemotional outcomes. Our paper includes indicators for both types of outcomes. Moreover, the Cambodian evaluation allows for a contrast of targeting approaches. Together, these evaluations inform the degree to which individual findings from specific contexts might have broader external validity (cf. Vivalt, 2017).

We estimate an intention-to-treat model. Our results show that, despite some catch-up from the control group between 2011 and 2016, scholarship recipients have on average 0.21-0.29 more years of schooling. This is in line with programs that attempt to reduce direct (for example scholarships; see Kremer et al., 2009, and Duflo et al., 2017) and indirect costs of education (as in conditional cash transfers; see Fiszbein and Schady, 2009).

We find positive effects on cognitive outcomes, but only for the merit-based approach to targeting. Impacts of the merit scholarships on a "family index"—that is an index that standardizes the cognitive skill measures and calculates a weighted average[3]—have an effect size of 0.11 standard deviations (significant at the 10% level); the effects for the poverty-based treatment are close to zero. This is consistent with the effects found after the initial three years, suggesting that there is no fade-out for this outcome.

We do not find any systematic impacts on two measures of socioemotional outcomes: emotional and behavioral difficulties (as measured by the Strengths and Difficulty Questionnaire, "SDQ") and the "Big 5" personality traits (openness, conscientiousness, extroversion, agreeableness and neuroticism).[4] For a "family index" of these outcomes, we find imprecisely estimated impacts of -0.01 (merit-based scholarships) and -0.10 (poverty-based scholarships) standard deviations. The findings therefore neither support the hypothesis that more schooling (necessarily) produces more socioemotional skills, nor the hypothesis that cognitive and socioemotional skills are (necessarily) co-produced.

[3] We calculate inverse covariance matrix-weighted averages. In the generation of these indices, we thus follow Anderson (2008).

[4] We collected information on other socioemotional outcomes such as grit and growth-mindset. However, the statistical properties of these measures in our context were weak. In a separate manuscript with Alice Danon, we provide a detailed discussion of these measures, and their properties (Danon et al., 2018).

We find that the probability of working increased by 3.4 percentage points for young adults who had received a merit-based scholarship (significant at the 10% level), but the impact for those who had received a poverty-based scholarship was close to zero (and statistically insignificant). The point estimates for earnings are both negative (but not statistically significant), perhaps because the scholarships induced individuals to delay entry into the labor market. At the same time, the population sample is situated in rural areas, with few labor opportunities outside the agricultural sector.

Finally, we find positive overall impacts on various measures of self-reported well-being, but again only for those who had received merit-based scholarships. For the "family index," the point estimate is 0.17 standard deviations (significant at the 1% level) for the merit-based treatment arm; for the poverty treatment, the point estimate is 0.04 standard deviations (and not statistically significant).

Overall, both types of scholarships led to more schooling attainment, but only the merit-based scholarships had positive impacts on cognitive, labor and well-being outcomes. Neither of the two types of scholarships induced greater socioemotional skills. These results need to be taken with caution: first, they are the marginal effect of increasing schooling by about four additional months—although they may be critical, inasmuch the program induced individuals to finish primary education. But it is possible that some of the key impacts of schooling on socioemotional skills happen early on (when both the control and treatment groups were still in school) or later on in adolescence (when, for this population, both groups would have left school). Second, while attrition is neither especially high nor systematically different across the three groups of students, our diminished sample size nonetheless may have reduced the precision of estimates. Our overall results present a complex picture, suggesting that demand-side interventions, such as scholarships, and their particular targeting approaches can have important long-term effects.

2. Related Literature

Our study builds on three strands of literature. First, we add to previous research on the effects of scholarship programs in the developing world, both in terms of their overall effect and with respect to varying targeting approaches. Second, we contribute to research on whether increased schooling produces outcomes that go beyond cognitive skills, in particular socioemotional skills. We furthermore investigate how these might be co-produced. Third, we contribute to the literature on

the long-run effects of increased school enrollment on outcomes such as employment status or well-being, as these impacts may only manifest themselves later in life.

There is a large empirical literature on the impact of scholarships in low- and middle-income countries, which are sometimes referred to as "conditional cash transfers" (Barham et. al., 2014; Baird, Ferreira, Özler, & Woolcock, 2014; García & Saavedra, 2017; Snilstveit et al., 2015).[5] Related analyses point to the importance of design characteristics, and particularly to whether scholarships are based on merit or students' economic need (Barrera-Osorio & Filmer, 2016). The majority of this literature is focused on schooling and cognitive skills outcomes, with some exceptions considering the impact of transfers on political and social factors, household consumption smoothing (Sparrow, 2007), labor market outcomes (Araujo, Bosch, & Schady, 2016; Filmer & Schady, 2014; Parker & Vogl, 2018; Silva & Sumarto, 2015), or health (Cruz, Moura, & Soares Neto, 2017). Few studies are able to analyze the impacts on various outcome dimensions simultaneously.

In high-income countries, socio-emotional skills have been found to be important predictors of success in school and life in general (see West et al., 2016 for an overview), and the importance of social skills has grown in the U.S. labor market between 1980 and 2012 (Deming, 2017). Research from the United States suggests that teachers can have large effects on socioemotional outcomes, although a teacher's productivity in terms of student cognitive achievement is only a weak predictor for her productivity in terms of students' performance socioemotional outcomes (Blazar, 2017; Blazar & Kraft, 2017; Jackson, 2016; Kraft, 2017; Santorella, 2017). At the same time, little is known—especially in low- and middle-income countries—about whether increased educational attainment leads to more socioemotional skills, and how this might interact with the formation of cognitive skills. Some analyses have tried to shed light on these relationships. For example, Kyllonen & Bertling (2013) report how participants' self-reported confidence in mathematics in the 2003 Programme for International Student Assessment (PISA) study was positively correlated with performance. Claro, Paunesku, & Dweck (2016) use a national data-set of all tenth-graders in Chile to show that a student's "growth mindset" can predict academic

[5] Scholarships may also be designed as incentive mechanisms, where payments are made based on future performance. See Fryer (2011) for related evidence from the US. See Berry (2015), Blimpo (2014), and Li et al. (2014) for examples of related research from India, Benin, and China, respectively. We study scholarships whose payout is not (or arguably, only weakly) incentivized. See Section 3, below, for a description of the scholarship program.

performance, alleviating socio-economic achievement gaps. However, these studies cannot identify exogeneous variation in schooling and cognitive skill, making causal inferences difficult.[6]

Research from the high-income countries suggests a common characteristic for effects of educational interventions is a lack of persistence (or "fade out"); i.e., initial positive effects that diminish in magnitude or disappear altogether over time (Bailey, Duncan, Odgers, & Yu, 2017; Protzko, 2015). But at the same time, other studies have shown positive effects on long-term outcomes, such as educational attainment, earnings, health outcomes, and (reduced) criminal behavior (K. H. Anderson, Foster, & Frisvold, 2009; Carneiro & Ginja, 2014; Chetty et al., 2011; Chetty, Friedman, & Rockoff, 2014; Currie & Thomas, 2000; Deming, 2009; Dynarski, Hyman, & Schanzenbach, 2013; Frisvold & Lumeng, 2011; Garces, Thomas, & Currie, 2002; Heckman, Moon, Pinto, Savelyev, & Yavitz, 2010; Ludwig & Miller, 2007). Yet, comparable evidence from low- and middle-income countries is scarcer; drawing lessons requires that educational interventions be defined more broadly. For example, Acevedo, Cruces, Gertler, & Martinez (2016) exploit a randomized controlled trial to assess the effect of a youth training and internship program in the Dominican Republic, approximately four years after its inception. The authors investigate personal skills (including grit and self-esteem), expectations, and labor market outcomes, finding that treatment effects differed substantially by gender. Further, Doyle et al. (2011) use a randomized experiment to evaluate the impact of a health education program in grades five to seven of Tanzanian primary schools (in combination with health services and community engagement). Six years after the program's implementation, the study documents improvements in sexual and reproductive health attitudes, knowledge, and behaviors. In addition, Walker et al. (2007) and Gertler et al. (2013) assess the long-term effects of a randomized early childhood stimulation program (in combination with food supplementation), for a small sample of adolescents in Jamaica. The authors find positive effects on anxiety, depressive symptoms, self-esteem, anti-social behavior, attention deficit, hyperactivity, and oppositional behavior, along with impacts on labor market outcomes.[7] Finally, both Ozier (2016) and Brudevold-Newman (2016)

[6] In a well-identified study, Fabregas (2017) investigates the effect of school quality and peer composition on students' academic performance, perseverance, aspirations, and time-management, in Mexico. Unfortunately, this and related research on peer effects (*ibid.*, for a review) does not shed light on the effects of educational attainment.
[7] See Krishnan & Krutikova (2013) for another, less well-identified study on the long-term effects of non-cognitive training in a small sample (n=154) of students in Mumbai, India. The authors (*ibid.*) find large impacts on self-esteem and self-efficacy, smaller impacts on life evaluation and aspirations, as well as positive impacts on

find positive effects of additional exposure to secondary education on labor market outcomes, in Kenya; Brudevold-Newman (*ibid.*) also demonstrates related delays in childbearing and marriage. A review of CCT programs in Latin America (Molina-Millan et. al, 2016) concludes that the literature is very mixed, with CCTs during the school years resulting in more cognitive, socioemotional skills and labor market outcomes in some settings, but not in others.

3. Intervention and Experimental Design

In 2008, the Government of Cambodia began implementing a new pilot scholarship program for Grade 4 students in 207 public schools. The program's stated goal was to reduce student drop-out rates and increase primary school completion, though the Government also implicitly sought to improve students' educational performance. At the time, the program's 207 schools represented all public schools in three of the country's 25 provinces[8] (Mondulkiri, Ratanakiri, and Preah Vihear); the three provinces had been selected for having the highest drop-out rates in the upper primary grades (grades four to six), according to Cambodia's Education Management Information System (EMIS).[9] The program was phased in as a pilot over two years, with a random set of 103 schools starting in 2008/09 and the remaining schools entering in the following year (random assignment was stratified by province).

The scholarship program targeted students entering Grade 4, using one of two selection approaches. In a randomly selected half of the scholarship schools (52 schools), students qualified given their combined performance on a test of Khmer and mathematics; in the remaining 51 schools, scholarships were disbursed based on a poverty-based test. Merit-based eligibility was determined through a centrally-scored test; the maximum possible score was 25. A student's "poverty score" was determined based on students' self-reported (but validated) household and socio-economic characteristics; the poverty index ranges from 0 (richest household) to 292 (poorest household). [10] Under both approaches, half of a given school's fourth-graders qualified

educational attainment and initial labor market outcomes (approximately eleven years after program participation started).

[8] Here, we count the capital as Cambodia's 25th "province". More precisely, Phnomh Penh is a special administrative district whose administrative characteristics partly resemble those of provinces.

[9] To limit the program's geographic scope, in Ratanakiri, only five of seven districts were selected, choosing those districts with the highest dropout rate. In the remaining two provinces, all districts were selected.

[10] The aptitude test was based on the 2005/06 Grade 3 National Learning Assessment. The poverty assessment asked respondents about household demographics and possession of a list of assets (as provided in Table 2). See Barrera-Osorio & Filmer (2016) for more details on the student assessment and the poverty score.

(i.e., the top half of performers, or the poorest half of students).[11] Crucially (for our study), students in all 207 schools completed both types of assessments, independent of their school's assignment status.

Scholarships were offered to beneficiaries for three years (i.e. through the end of primary school), conditional on their continued enrollment, passing grades, and regular attendance. These requirements were moderately enforced.[12] Scholarships disbursed a lump-sum payment of approximately USD20 in the first year, and two payments of approximately USD10 in each of the following two years. As reported by Barrera-Osorio & Filmer (2006), these amounts represent about 3.3 percent of the yearly per capita expenditure in the study sample. Nevertheless, these transfers are small in size, vis-à-vis transfers from similar programs such as conditional cash transfers (Fiszbein & Schady, 2009), making the program potentially quite cost-effective, even in the event of small impacts.

Our experimental design exploits the randomized roll-out of the program, over its two phases. In 2008/09, during phase one, fourth-graders in schools that were selected to disburse the program in the second phase did not receive any scholarship and did not become eligible in the years thereafter.[13] Note that a sub-set of these fourth-grade students would have been eligible under one of the two targeting schemes (merit-based or poverty-based), had their school been selected. In expectation, these two sub-samples are equal to their respective eligible peers from phase-one schools (below, we present supportive evidence that the two groups of students are in fact balanced, across phase-one and phase-two schools). Thus, we can identify the causal intent-to-treat effect of the scholarship program, under either of the two targeting approaches. As phase-one schools were moreover randomly assigned to either the poverty-based or merit-based targeting scheme, we can also compare the scholarship's effect across the two targeting schemes.

[11] Median students also qualified for the scholarship. The number of scholarships was determined using the previous year's official enrolment numbers.
[12] If a student lost her scholarship, its amount could not be re-allocated within the same school and the same year. Instead, the amount would be used for the subsequent cohort of fourth-graders.
[13] Recall that the program required students to maintain passing grades. Thus, a phase-one fourth-grader who attended a control group school could not become eligible in phase two by repeating the grade.

4. Estimation Framework, Internal Validity

We estimate a generic production function model:

$$Y_{t,i}^{j} = \beta_0 + \beta_1^{j} T_{0,i}^{j} + \mathrm{B}\mathbf{X}_{\mathbf{0},\boldsymbol{i}} + \mu_{t,i} \quad \text{for } j = merit \; or \; poverty \qquad (1)$$

where Y includes education outcomes (including educational attainment), cognitive skills, socioemotional skills, labor outcomes, and measures of well-being (which include socio-economic status, SES). Vector $\mathbf{X}_{\mathbf{0},\boldsymbol{i}}$ includes a rich set of baseline characteristics at the student's school-, village-, and individual-level (the next section describes these measures in greater detail). All estimations include district-level fixed effects and allow for the clustering of standard errors at the assignment level (i.e., within schools; cf. Abadie et al.,2017). Equation (1) estimates an intent-to-treat model, which captures the effect of offering the scholarship on outcomes Y.

Our default approach is to estimate Equation (1) as two separated OLS models, for the merit- and poverty-based sub-samples.[14] For yearly earnings we use a Tobit model with an inverse hyperbolic sine transformation of the outcome variable because its distribution shows a spike at zero (cf. Duflo et al., 2017).[15]

For each "family" of educational outcomes, cognitive outcomes, socioemotional outcomes, and well-being, we present the results from a test that the treatment coefficients are jointly zero (using seemingly unrelated regressions, SUR). Within these four sets of outcomes, we also use SUR to test whether the treatment coefficient for the poverty subsample is equal to that for the merit subsample.

Our sampling frame consists of 5,964 fourth-grade students (in the program's 207 schools), who participated in the "baseline" eligibility assessment, in December 2008 and January 2009. Of those, 2,996 respondents were randomly selected for the first three-year follow-up survey, in 2011. For this first follow-up, an additional 658 "replacement" students were randomly selected, in case students from the target group could not to be found. In the 2016 follow-up, we tracked all students

[14] Using simulations, we compared our strategy to others that would (a) use a regression-discontinuity approach (exploiting the continuous poverty- and merit-indices and their strict cut-off), (b) a difference-in-difference approach, and (c) a difference-in-discontinuity approach (not shown). All three alternative strategies make additional assumptions and do not lead to increased statistical power.

[15] For respondents' daily reservation wage, we also calculated a two-part regression or "Tobit II" model, where the second part of the model uses a log transformation of the outcome variable (see Belotti, Deb, Manning, & Norton, 2015). Results do not lead to substantial changes and are available upon request.

who had participated in the 2011 study, a random subset of 140 respondents who had previously been found to be attritors, and all replacement students who were interviewed in 2011. Our 2016 sample includes 2,252 respondents, of which 2,024 had been interviewed in 2011, 86 had not been reached previously, and 142 had served as replacements, in 2011.

Table 1 provides the control group means for key demographic characteristics, for the "merit" and "poverty" sub-samples (for the control group, these refer to respondents who would have qualified if their school had been assigned to one or the other scholarship approach). Our analysis sample consists of 890 and 825 respondents for the merit-based and poverty-based sub-samples respectively. Among those, about half (48% and 51%, respectively) are female. On average, respondents live with an additional six household members. Almost all of the respondents were already working at the time of the three-year follow-up survey.

[Table 1 about here]

The data support the fact that our experimental design is valid. First, we find that both sub-samples are balanced on observables. This holds true for the full set of respondents at baseline, as discussed by Barrera-Osorio and Filmer (2015), and for this paper's estimation samples (see Tables A1 and A2, in the Appendix). Second, overall attrition is 31 percent for either sub-sample, and we managed to track 88 percent of respondents who were included in the three-year follow-up study (i.e., six years after our last contact with study participants). As shown in Table 2, there are no systematic differences in attrition by treatment group. Columns (5) of the Table 2's "merit scholarship" and "poverty scholarship" panels present the difference-in-difference among attritors and non-attritors, across respondents in the treatment and control groups (computed by OLS regression, controlling for stratification fixed effects). Only two out of 16 indicators in the merit subsample and only three indicators in the poverty subsample show a statistically significant or marginally significant difference-in-differences; this result is not surprising given multiple comparisons. We also test for the individual coefficients being jointly equal to zero, using seemingly unrelated estimation (SUR); the resulting Chi-square statistics (and corresponding p-values) suggests that we should not reject that the two sub-samples are balanced.

[Table 2 about here]

5. Data and Measurement

Our analysis combines data from five main sources. First, we collect outcome data through in-person interviews at the respondents' residence, using handheld tablets. Second, to construct a variable reflecting intention-to-treat, we use the official government declaration ("Prakas") of scholarship recipients. Third, we match each respondent to application forms and baseline tests, as collected in December 2008 and January 2009. We can thus control for baseline test scores, and for students' initial household characteristics. Fourth, we construct a vector of control variables through administrative data on baseline school characteristics, as provided by the country's Educational Management Information System (EMIS).[16] Fifth, we take advantage of the fact that Cambodia's 2008 census was conducted just before the scholarship program started. Using geographic coordinates, we match each school to its closest village and include this village's demographic characteristics as additional controls.[17]

Data collection for the baseline and three-year follow-up occurred from December 2008 to January 2009, and from May to September 2011, respectively. Data collection for our latest round of follow-up took from December 2016 to May 2017. We guaranteed data-quality by following standard monitoring procedures, as described by Glennerster (2017). First, during the first week of field work, we conducted 30% of re-surveys ("back-checks", usually within three days) and then reduced this number, for an overall back-check rate of 15.7%. Second, we spot-checked approximately 20% of interviews, provided immediate feedback, and offered repeat-trainings to enumerators. These spot-checks were not only conducted by field supervisors but also through additional, independent field-monitoring. Third, we ran daily analytics on newly collected data to spot irregularities, and to identify training needs. Finally, we employed 15% of staff as dedicated quality-control officers, such that steps to improve data quality could be taken immediately, as part of the regular data flow.

[16] We include a binary indicator of whether a school had access to drinking water, a binary indicator of whether the school counted with a toilet facility, the number of primary school classrooms, the number of newly enrolled fourth-graders, the number of teaching staff, and the school's income.

[17] Village-level data as published by the Cambodian National Institute of Statistics at the Ministry of Planning (2010). We control for the share of villagers who are literate in Khmer, the share of villagers with no schooling, the percentage of villagers engaged in crop or animal farming, the village's population size, and a continuous measure of villagers' household assets.

The following discusses our newly collected outcome measures in greater detail. As education outcomes, we measure educational attainment (highest grade completed), formal and informal training that lasted for at least one week (a binary variable), and whether the respondent received any formal education since the early three-year follow-up (a binary variable).

We also collected data on four measures of cognitive skill. First, we administered a computer-adaptive math-test, in which respondents answered ten questions from a larger pool of 23 items.[18] We used a three-parameter logistic (3PL) item response theory (IRT) model with a single guessing parameter (Birnbaum, 1968; Samejima, 1969) to analyze responses to math tests from an evaluation of a similar scholarship program in Cambodia that was targeted to secondary school students (Filmer & Schady, 2008). Participants in this assessment had been tested in two rounds, with overlapping items, and we follow the common Stocking-Lord (1983) methodology for IRT-based scale equating.[19] Our test begins with the item of median difficulty. As the test is administered and respondents answer correctly or incorrectly, our assessment picks the next item to be displayed based on maximum information, re-calculates a respondent's ability estimate using expected a posteriori, and continues thereafter until ten items are administered for each respondent (cf. Bock & Mislevy, 1982; van der Linden & Pashley, 2010). The second assessment is a test of shapes and puzzles loosely based on the Raven's Progressive Matrices. This test is a measure of fluid intelligence; respondents are asked to complete 15 sets of pattern recognition. Our third measure is a "Digit Span" test, which asks respondents to repeat sequences of single-digit numbers, of increasing length. This test is a common measure of respondents' working memory (cf. Hamoudi & Sheridan, 2015). Sequences are presented in sets of two and begin with two integers (asking respondents to repeat 2-1 and 1-3). No additional sequences are asked if a respondent fails to repeat both prompts; the last set of longest sequence presents two strings of eight integers (asking respondents to repeat 6-9-1-7-3-2-5-8 and 3-1-7-9-5-4-8-2). The fourth outcome is a vocabulary test based on picture recognition. This test asks respondents to identify the picture corresponding to a word which the enumerator reads out loud. For each word the respondent is asked to select from a choice of four pictures. The test is structured such that items become increasingly difficult (examples of easy items include, "citrus," and "garment"; items of highest

[18] To our best knowledge, this assessment constitutes the first computer-adaptive ability test as conducted during a household survey, in a developing country.
[19] We removed one item with low discrimination.

difficulty include "vitreous" and "lugubrious"). A maximum of 96 items is presented in sets of 12, and no additional item is displayed if a respondent fails to answer at least five items correctly, in a given set. The final ability estimate for each of the math, pattern recognition, and vocabulary recognition tests are calculated with a two-parameter logistic (2PL) IRT model. The Digit Span test score reflects the number of integer sequences a respondent repeated correctly. All four measures are standardized (mean zero and standard deviation of one).

We report on two sets of socioemotional outcomes: we screen for emotional and behavioral difficulties with the Strengths and Difficulty Questionnaire ("SDQ"), and measure the "Big 5" personality traits. The SDQ represents a common screening instrument; we use (the official Khmer translation of) its most frequently used version with 25 items on psychological attributes (Goodman, 1997). Following its scoring guidelines and official recommendations (*ibid.*), we report on three subscales, separated into 'internalizing problems' (emotional and peer symptoms, 10 items), 'externalizing problems' (conduct and hyperactivity symptoms, 10 items), and a scale of prosocial behavior (5 items). To capture respondents' personality traits, The Big Five Scale measures five core dimensions of personality. The five broad personality traits measured are extraversion, agreeableness, openness, conscientiousness, and neuroticism. Evidence of the Big Five as being relevant (and associated with life outcomes) has been growing, beginning with the research of Fiske (1949) and later expanded upon by other researchers including Norman (1967), Smith (1967), Goldberg (1981), and McCrae and Costa (1987). We use the short 15 item Big Five Inventory (BFI-S) (Lang, John, Lüdtke, Schupp, & Wagner, 2011), with three items per personality trait. Like the indicators of cognitive skill, all measures of socioemotional outcomes are standardized.[20]

Our household also survey collected information on five labor market outcomes. We ask whether a respondent is currently working (yes or no) and the age at which she or he first started working. We moreover construct a binary indicator of whether a respondent's main work activity is cognitively demanding. We categorize an occupation as such if it requires at least occasional use of reading, writing, mathematics, or a computer (according to the respondent). Our survey also

[20] Furthermore, we collected data on respondents' level of grit (Duckworth & Quinn, 2009) and their growth mindset (Dweck, 2000). We discard results for these measures given their psychometric properties (as detailed in a separate manuscript). The inclusion of neither of these measures would have changed our substantial results – we do not find effects on these outcomes. However, we caution against their measurement properties.

asked for respondents' income; our analysis reports on (the inverse hyperbolic sine of) yearly earnings and (the inverse hyperbolic sine of) a respondent's daily reservation wage, i.e., the minimum wage or payment for which a respondent is willing to accept work (both are reported in US dollars, a currency commonly used in Cambodia).

Our last set of outcomes includes six indicators of socio-economic status and well-being. We assess subjective social status using a "MacArthur community ladder".[21] Respondents were shown a picture of a ladder with ten rungs and were told that higher rungs correspond to higher socioeconomic status. They were then asked to place themselves on this ladder in relation to everyone in their community. As a second measure of socio-economic status, we construct an index of respondents' household assets, asking whether they possess items from a list similar to the one presented in Table 2. To calculate an individual's latent SES score, we borrow from the psychometric literature and estimate a two-parameter logistic (2PL) IRT model, placing responses from 2009, 2011 and 2016 on the same scale. We also asked respondents to rate their satisfaction with life at present, all things considered, on a scale from one ("completely dissatisfied") to ten ("completely satisfied") and to rate their quality of life and health, respectively, on a scale from one ("poor") to five ("excellent"). The fifth and last measure screens for (minor) mental health disorders, using the General Health Questionnaire ("GHQ"). We use the short form of the questionnaire (GHQ-12) with Likert scoring (D. Goldberg & Williams, 2006; Quek, Low, Razack, & Loh, 2001). All six measures are standardized (mean zero and standard deviation of one).[22]

Finally, for each set of educational outcomes, cognitive outcomes, socioemotional outcomes, and SES and subjective well-being, we also calculate an overall "family index," following Anderson (2008).[23] These indices have the benefit of reducing the number of statistical tests (and the temptation of cherry picking positive results). In constructing the indices, we made sure that the qualitative "direction" of the construct was preserved—higher values point to more desirable outcomes. However, our index construction is atheoretical and, as a result, they may lump together

[21] For a description and bibliography of papers that use MacArthur ladders, see the MacArthur Foundation's Network on SES and Health website: http://www.macses.ucsf.edu/Research/Psychosocial/subjective.php.
[22] We standardize by focusing on the endline measures for control group students (who would have qualified for at least one of the two types of scholarships, had they been in a treatment school instead).
[23] We also considered using an alternative index instead, following Kling, Liebman, & Katz (2007). Throughout, the two sets of results do not lead to qualitatively different conclusions.

measurements with different underlaying constructs. Nevertheless, we discuss results from both individual measurements and the family indices.

6. Results

Tables 3 to 7 present results on five main categories of outcomes: education; cognitive skill; well-being; socioemotional characteristics; and labor outcomes. These tables share a common structure. Each table has two panels; Panel A restricts the sample to the merit sample, whereas Panel B presents results for the poverty sample. Each panel presents separate regressions for a given dependent variable, as stated in the column headers. For the treatment variable (1 if assigned to treatment, 0 otherwise), the table presents regression coefficients and standard errors. All regressions control for covariates at baseline and district fixed effects; standard errors are clustered at the level of randomization (the school) (Abadie et al. 2017). Each of the two panels also presents the unconditional mean, as observed for the control group. Each panel moreover includes the results from a chi-square test on the null hypotheses that all treatment coefficients are jointly zero, using seemingly unrelated regression (SUR). Finally, across the two panels and for each of the outcome variables, we present results for a test of the null hypothesis that the two treatment coefficients (merit and poverty) are equal.

a. Education

A basic objective of the program was to increase school progression of low-income individuals. Early dropout from primary school is still a major obstacle in education in Cambodia, especially in rural areas. At inception of the program, only close to 40% of the poorest quintile of income completed 6th grade (Barrera-Osorio and Filmer, 2016). As such, the first set of outcomes that the program aimed to change was to induce greater school progression, at least until students successfully graduate from primary school (grade 6).

Table 3 presents results for school progression (highest grade attained), primary school graduation (a zero-or-one variable), an indicator of whether the respondent received any formal education since the three-year follow-up study (in 2011), and a "family index" of the previous three measures (measured in standard deviations, SDs). On average, students in the control group completed 5.45-5.57 grade levels. Both treatments increased educational attainment, with similar point estimates (0.213 and 0.291 for merit and poverty treatment, respectively, equivalent to about four additional

months of schooling). The effects on overall attainment, as reported by Barrera-Osorio and Filmer (2016) after three years of starting the program, were slightly higher for the poverty sample (0.34), and similar for the merit sample (0.225) (Barrera-Osorio and Filmer, 2016, Table 4, column 3), indicating some catch-up by the control group. These effects induced treated students to graduate from grade six and complete primary education (by 5.0 and 11.3 percentage points, respectively; statistically significant for the sample with poverty-based scholarships).

[Table 3 about here]

The point estimates for impacts on any formal education (in the 2011-2016 period) are positive and statistically significant for the poverty-based scholarship, suggesting some catch-up with the average in the merit-based scholarship group: 71% of individuals of the poverty-control-group enrolled in formal education, and the effect of the poverty scholarship is 10.0 additional percentage points (pp). In contrast, students who received a merit-based scholarship increased their enrollment by 4.4 pp (significant at the 0.1 level). The joint test of all coefficients being equal to zero is rejected for both treatments (a p-value of 0.06 and 0.04 for the merit and poverty treatments, respectively). Both point estimates of the regression with the "family index" as dependent variable are positive and statistically significant (at the 5% and 1% levels, respectively), with a point estimate of 0.131 standard deviations for the merit-based and of 0.264 standard deviations for the poverty-based scholarships. However, we cannot conclude that the two coefficients are in fact different, with a corresponding p-value above 0.1.

b. Cognitive skills

A second objective of the program was to induce an increase in students' cognitive skill, as measured through standardized assessments. The basic mechanism to induce learning was by exposing minors to additional schooling. The measures we use are proxies for cognitive outcomes, capturing skills that relate to an individual's knowledge, ability to tackle problems, and fluid intelligence. Note that, unsurprisingly, the control group for the merit sample has higher average test scores on these measures than the control group for the poverty sample.

Table 4 presents the results of treatment on these measures of cognitive skills. Across the different measures, we find suggestive evidence of positive effects for the merit-based treatment. All coefficients are positive, and two of them are statistically significant (Raven's and the overall

"family index.") The estimation suggests an overall effect of 0.113 standard deviations on these measures (significant at the 10% significance level). In contrast, the results for the poverty-based scholarship are either close to zero or even negative, in the case of the Forward Digit Span (a point estimate of -0.129 SDs, significant at the 10% level, and different from the effect for the merit-based transfer, significant at the 5% level). The "family index" is close to zero in the case of the poverty scholarship (0.01 standard deviations).

[Table 4 about here]

The findings here, nine years after program inceptions are similar to those documented in the previous three-year follow up study. In that study, merit-based scholarship recipients got statistically significant higher scores in mathematics and for the Digit Span test (Barrera-Osorio and Filmer, 2016), whereas poverty-based scholarship recipients did not.

c. Socioemotional skills

An important contribution of this study is the analysis of effects on socioemotional skills. We are not only interested in measuring the effects of the scholarship on these measures; in addition, we also consider the relationship between cognitive and socioemotional skills. The intuition of this analysis can be divided into two parts. First, if the scholarship induced more schooling for both types of scholarship recipients, then, under the assumption that schools also "produce" socioemotional skills, we should observe effects on these skills from both poverty- and merit-based scholarships. In contrast, if there is a complementary relationship between cognitive and socioemotional skills, we should observe effects on socioemotional skills for students with the merit-based scholarship, and not for students in the poverty-based scholarship. We formally present these relationship in the next paragraphs.

Our approach is based on two different conceptual models of the relationships between years of education (E), cognitive skills (C), and socioemotional skills (S). As a starting point, based on the evaluation three years after the program's inception (Barrera-Osorio & Filmer, 2016), we know that treatment T_0 (at baseline, $t = 0$) increased years of education schooling for both merit- and poverty-based scholarships ($E_t = f(T_0;\ X_0, Z_0)$; $\frac{E_t}{\partial T_0} > 0$, for both types of scholarships). Furthermore, the evaluation showed a causal, positive effect of the intervention on cognitive

outcomes for the merit-based scholarship only ($C_t^M = f(T_0^M; X_0, Z_0)$, $\frac{\partial C_t^M}{\partial T_0^M} > 0$); and zero effects for the poverty-based scholarship ($C_t^P = f(T_0^P; X_0, Z_0)$, $\frac{\partial C_t^P}{\partial T_0^P} = 0$), where *M* denotes merit-based treatment and *P* denotes poverty-based treatment.

The first conceptual relationship we explore is between each type of skill—cognitive and socioemotional—and the years of education:

$$C_t = g(E_t; X_0, Z_0)$$

$$S_t = g(E_t; X_0, Z_0)$$

where X_0 are student characteristics and Z_0 are the inputs of the school (at baseline). These equations state that the effect on either set of skills is a function of the years of education; i.e., exposure to more schooling will induce higher cognitive and socioemotional skills. Therefore, the first set of relationships we are testing is:

$$\frac{\partial C}{\partial T} = \frac{\partial C}{\partial E} * \frac{\partial E}{\partial T} > 0 \quad (2)$$

and

$$\frac{\partial S}{\partial T} = \frac{\partial S}{\partial E} * \frac{\partial E}{\partial T} > 0 \quad (3)$$

Under the assumption that more schooling produces cognitive and socioemotional skills, both Equations (2) and (3) are positive, independently of the type of treatment (merit or poverty).

In contrast, the second conceptual relationship is based on one modification of this setup: for the merit-based scholarship we have an additional equation, relating cognitive skills and treatment:

$$C_t^M = f(T_0^M) \quad (4)$$

i.e., treatment induced higher cognitive skills only for the merit (M) treatment, as the third-year evaluation showed (Barrera-Osorio and Filmer, 2016). The basic relationship of interest is between socioemotional skills and cognitive skills:

$$S_t^M = g(C_t^M, E_t; X_0, Z_0)$$

The second relationship we test is therefore:

$$\frac{\partial S_t^M}{\partial T_0^M} = \frac{\partial S_t^M}{\partial C_t^M} * \frac{\partial C_t^M}{\partial T_0^M} + \frac{\partial S_t^M}{\partial E_t} * \frac{\partial E_t}{\partial T_0^M} > 0 \quad (5)$$

i.e., the effect of treatment on socioemotional skills is positive, and it depends on the effect of cognitive skills on socioemotional skills ($\frac{\partial S_t^M}{\partial C_t^M}$) and on the indirect effect of higher exposure to more schooling ($\frac{\partial S_t^M}{\partial E_t}$). If there is complementarity (or co-production) between cognitive and socioemotional skills (e.g., $\frac{\partial S_t^M}{\partial C_t^M} > 0$), then $\frac{\partial S_t^M}{\partial T_0^M} > 0$.

For the case of the poverty-based scholarship (P), we will have

$$\frac{\partial S_t^P}{\partial T_0^P} = \frac{\partial S_t^P}{\partial E_t} * \frac{\partial E_t}{\partial T_0^P} \quad (6)$$

since $\frac{\partial C_t^P}{\partial T_0^P} = 0$, as shown with the third-year evaluation.

There are three main relevant cases for Equations (5) and (6). If exposure to school produces socioemotional skills, both equation (5) and (6) are positive. If exposure to schooling does not produce socioemotional skills, Equation (6) is equal to zero. Finally, under complementarities between cognitive and socioemotional skills (e.g. if cognitive skills help in the acquisition of socioemotional skills, or if they are co-produced), then Equation (5) is positive, independent of the relationship between socioemotional skills and exposure to school.

Table 5 presents results on the Strengths and Difficulties Questionnaire (SDQ)—separating out the three attributes: prosocial, internalizing, and externalizing—and on the Big 5—separating by its five traits: Openness, Conscientiousness, Extraversion, Agreeableness and Neuroticism (OCEAN).

Overall, we reject the hypothesis of impacts on any subcomponent of these two groups of socioemotional outcomes. All point estimates are close to zero, with the exception of Neuroticism for the poverty treatment, with a coefficient of 0.186 standard deviations (statistically significant at the 1% level). Nevertheless, the coefficient for the impact on "family indices" for both treatments are close to zero (-0.005 and -0.099 standard deviations for the merit- and poverty-based treatment, respectively), and neither coefficient is statistically significant.

[Table 5 about here]

The broad pattern of the table suggests that the program did not produce effects on socioemotional skills, despite the observed impact on school progression and, for the merit sample, on cognitive outcomes. We cannot rule out competing hypotheses such as low marginal exposure to schooling (i.e., the treatment groups increased their educational attainment by only about four months, on average)—as result, we may observe no effects on socioemotional outcomes. We note, however, that this amount of additional schooling was sufficient to produce improved cognitive performance among the recipients of merit-based scholarships.

d. Labor outcomes

Table 6 presents the effects of the program on labor status (coded as one if the respondent is "currently working", and zero otherwise); the respondent's age when they started working (which captures child labor); whether the recipient participated in work-related training that lasted for at least one week (formal or informal, a zero-or-one variable); the cognitive demands of the respondent's main work activity; and two measures of income: yearly earnings and the daily reservation wage (both transformed using an inverse hyperbolic sine).

We find a positive impact for recipients of the merit-based scholarships on the probability of working (3.4 percentage points, statistically significant, at the 10% level); the point estimate for the poverty arm is lower (1.2 percentage points), and not statistically significant. These effects are from a high baseline level of people who report to be currently working (the means of both control groups are around 92%).[24]

[Table 6 about here]

Respondents in our two samples started to work very early in life, when they were between 12 and 13 years old. Respondents who were offered a scholarship delayed entering the labor market by 0.074 and by 0.339 years for the merit- and poverty-based program, respectively, in line with the results for school progression. However, these estimates have large standard errors and are not significantly different from zero. In addition, while about 58% of control group respondents report having received formal or informal training since 2011 (which could have improved their work

[24] Of those, approximately 84% report agriculture, fishery, or forestry as their main field of work. Approximately 90% report to engage in agriculture, fishery, or forestry among their overall labor activities. We follow the International Standard Classification of Occupations (ISCO); these individuals' occupation falls into ISCO Major Group 6 ("skilled agricultural and fishery workers").

prospects), we do not see any effects on this outcome. This also relates to our finding that only a small share of respondents (less than 18%) engage in economic activities that are cognitively demanding. There is no evidence of impact on the cognitive demands of the main work activity, for either of the two treatments.

The point estimates on yearly earnings are negative for both treatments arms (but not statistically significant); one potential explanation is that the scholarship program delayed entry into the market for recipients and, as result, they have lower experience than non-recipients. We observe a positive impact on the daily reservation wage for both groups; however, the estimates are very imprecise.

e. **Well-being outcomes**

Table 7 presents effects on various measures of well-being. These include both self-assessed measures as well as more readily observed indicators such as measures of household asset ownership. All these measures are standardized to have a mean of zero and a standard deviation of one. As with the previous tables, we also present results for a standardized "family index" of measures of respondent well-being.

[Table 7 about here]

Both treatments caused a positive impact on perceived status as measured by the SES ladder, with point estimates of 0.173 and 0.208 standard deviations for merit-based and poverty-based scholarships respectively. In addition, merit-based scholarships resulted in statistically significant positive impacts on respondents' ownership of household assets (0.186 standard deviations), quality of health (0.129 standard deviations) and on the "family index" (0.174 standard deviations). None of the other impacts for the poverty-based scholarships are statistically significantly different from zero (that it, other than the SES ladder) although most of the point estimates are positive (with the exception of asset ownership). We reject the null hypothesis, for both targeting approaches, of all estimators being equal to zero (at the 1% level of significance). Impacts on the overall "family index" weakly suggest that the overall impact on well-being was larger for the merit-based scholarship than the poverty-based one (p-value = 0.168).

f. Heterogeneity

We investigate two types of heterogeneous effects. Both sets of analyses use the "family indexes" for education, cognition, socioemotional outcomes, and well-being outcomes; as an indicator of labor outcomes, we use a respondent's daily reservation wage. Table 8 presents heterogeneous effects by treatment label. Our first analysis of heterogeneous effects compares the impact of scholarships for respondents who would have qualified for a transfer under either of the two targeting schemes. For those individuals, the scholarship only differs in terms of its name or "label". Accordingly, we estimate a regression that includes the treatment dummy, an indicator for whether a respondent would *not* have qualified under the *other* scheme, and the interaction between the treatment and this indicator. Of interest is a comparison of the two treatment coefficients; a difference of point estimates would indicate a labelling effect, as was shown to be the case in the analysis of the three-year follow up (Barrera-Osorio & Filmer, 2016).

Table 8 provides suggestive evidence for heterogeneous effects by treatment label, favoring the merit-based presentation of scholarships over their poverty-based presentation. The table confirms that there is no difference in effects on educational attainment and no effect on respondents' socioemotional outcomes, for either of the two treatment labels (compare to Tables 3 and 5, above). If at all, poverty-based scholarships led to greater impacts on educational attainment but lower effects on socio-emotional outcomes (differences in effects are not statistically significant). Moreover, the previously reported difference in impacts on cognition and well-being (compare to Tables 4 and 7, above) remains meaningful in size; the effect sizes differ by 0.155 and 0.144 standard deviations, respectively. However, this comparison is imprecise. Finally, column (5) of Table 8 suggests substantial labelling effects on respondents' reservation wage (p-value = 0.018): "merit-based" recipients demand higher wages than "poverty-based" recipients—even when their underlying characteristics are similar.[25]

[Table 8 about here]

Table 9 investigates heterogeneous effects by gender. We present results from regressions of the dependent variables on a treatment indicator, a gender indicator (female = 1, and zero otherwise),

[25] It is worth pointing out that effects of the merit-based scholarships on cognition, socioemotional outcomes, and on respondents' reservation wage appear to be largely driven by poor individuals (as indicated by the coefficient for the interaction term).

and their interaction. The table also assesses a potential difference in effects across the two treatments, for boys only (as indicated by a Chi-square test and its corresponding p-value). Results are mixed and are interpreted with caution, as some of the point estimates suffer from large error bands. We do not find differential effects on a beneficiary's educational attainment, socioemotional outcomes, or well-being (whether within or across the two programs and samples). In contrast, we document large, substantial differences in the effect of poverty-based transfers on cognition; the treatment effect for females is negative and substantially smaller, in comparison to the effect for males (a statistically significant difference of 0.309 standard deviations). For cognitive outcomes, we now also observe a positive treatment effect of the poverty-based transfer, for male recipients; the effect is slightly larger than the merit scholarship's point estimate (statistically significant, at the 10% level). Finally, Column (6) suggests that the (imprecisely estimated) impact on a recipient's daily reservation wage favors males; for females, the point estimate is close to zero, for either one of the two programs.

[Table 9 about here]

7. Conclusions

This study has investigated the long-term impacts of increased schooling, with a particular focus on potential complementarities across schooling, the development of cognitive skill, and socioemotional and labor market outcomes later in life. To this end, we evaluated the long-run effects of a primary school scholarship program in rural Cambodia, nine years after the program's inception, tracking study participants when they were, on average, 21-years-old. Overall, we find that targeting approach matters for the impact on cognitive skills, socio-economic status, and well-being. The merit-based and poverty-based targeting schemes both led to increased schooling, but only the merit-based scholarship led to improvements in cognitive skill and to greater well-being.

Our study points to important new avenues for researchers and policy makers. Prior work argues that more schooling does not necessarily imply more learning (World Bank, 2017); in turn, our work highlights that more schooling, even if it enhances learning, may not necessarily translate to noticeable changes in the labor market outcomes and may not lead to measurable improvements in socioemotional skills. To better understand this puzzle, additional research is needed, in at least three areas. First, our analysis of heterogeneous effects provides suggestive evidence that labor market effects may be concentrated among poorer beneficiaries who are male. This result echoes

the findings of Duflo, Dupas and Kremer (2017), who find labor market effects for a subset of male students, only. This renders the questions of how to effectively combine merit-based targeting with poverty-based targeting in a multi-step approach, and how to affect female beneficiaries such that improved learning can "translate" into labor market outcomes. Second, our findings resonate with prior research by Jackson (2016), which suggests that the school-based production of cognitive skill may not necessarily go hand-in-hand with improvements in socioemotional outcomes. However, research on how to purposeful foster socioemotional skill in school settings is only in its infancy, in particular in developing countries (*cf.* West et al., 2016). Third, we acknowledge that our reported lack of impacts on socioemotional skills may be at least partially driven by a lack of precision; we encourage other researchers to improve upon our study through continued work on the measurement of socioemotional skill in developing countries (such as Laajaj & Macours, 2017) and through similar, long-run evaluations with even larger samples.

Tables and Figures

Table 1: Sample Characteristics at Baseline and Three-year Follow-up

	Merit Scholarship					Poverty Scholarship				
	n	All	Treatment	Control	Difference	n	All	Treatment	Control	Difference
	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
Panel A. Baseline Characteristics										
Age	822	11.882	12.034	11.732	0.358	776	11.927	12.048	11.799	0.353
		(2.094)	(2.066)	(2.114)	(0.24)		(2.043)	(2.028)	(2.054)	(0.237)
Female	828	0.504	0.507	0.5	0.005	786	0.587	0.64	0.53	0.098***
		(0.5)	(0.501)	(0.501)	(0.035)		(0.493)	(0.481)	(0.5)	(0.035)
Number of minors in HH	809	1.686	1.631	1.742	-0.123	770	1.821	1.856	1.783	0.081
		(1.103)	(1.092)	(1.112)	(0.123)		(1.086)	(1.052)	(1.122)	(0.12)
Poverty Index (0-292)	858	210.88	204.417	217.557	-10.614	790	244.295	242.447	246.239	-1.2
		(60.471)	(66.586)	(52.676)	(8.141)		(32.826)	(34.06)	(31.404)	(4.482)
Test score (0-25)	858	19.751	19.798	19.701	0.194	790	18.229	18.686	17.748	1.097
		(3.098)	(2.869)	(3.321)	(0.481)		(4.744)	(4.729)	(4.717)	(0.672)
Panel B. Follow-up Characteristics										
Age	797	15.238	15.101	15.382	-0.208	718	15.355	15.082	15.643	-0.501*
		(2.202)	(2.269)	(2.123)	(0.244)		(3.063)	(2.092)	(3.809)	(0.287)
Female	797	0.497	0.506	0.487	0.019	718	0.568	0.633	0.5	0.124***
		(0.5)	(0.501)	(0.5)	(0.034)		(0.496)	(0.483)	(0.501)	(0.034)
Number of minors in HH	797	2.592	2.595	2.59	0.058	718	2.687	2.557	2.823	-0.213
		(1.645)	(1.547)	(1.743)	(0.148)		(1.659)	(1.539)	(1.77)	(0.148)
HH size	797	7.12	7.209	7.028	0.284	718	7.088	6.897	7.289	-0.294
		(2.545)	(2.501)	(2.591)	(0.227)		(2.401)	(2.22)	(2.566)	(0.218)
Married	503	0.074	0.076	0.071	0.017	454	0.07	0.068	0.073	0.001
		(0.261)	(0.266)	(0.257)	(0.031)		(0.256)	(0.252)	(0.261)	(0.031)
Currently Working	826	0.903	0.895	0.911	-0.012	760	0.872	0.857	0.889	-0.032
		(0.296)	(0.307)	(0.285)	(0.027)		(0.334)	(0.35)	(0.315)	(0.032)

Notes: Minors refers to respondents age 14 and under; this may include the respondent. HH size refers to the number of people living in the respondent's household, including the respondent. Married is a dummy equal to 1 if the respondent is currently married and 0 if never married, divorced or separated. This variable is missing for minors. Currently working is a dummy equal to 1 if the respondent worked during the last week or has a job at the moment and 0 otherwise; respondents may work and also be a student. Column (1) presents the number of observations in the analysis sample. Colums (2) to (4) display the means for the full sample, the treatment group, and the control group, respectively. Standard deviations in parentheses. Column (5) is the difference between the treatment group mean and the control group mean. Differences in means are computed by OLS regression, controlling for province fixed effects. Standard errors in parentheses are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1.

Table 2: Analysis of Differential Attrition

	Merit Scholarship					Poverty Scholarship				
	Attritor C	Attritor T	Non-attritor C	Non-attritor T	Diff-in-Diffs	Attritor C	Attritor T	Non-attritor C	Non-attritor T	Diff-in-Diffs
	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
Female	0.4	0.38	0.5	0.51	-0.03	0.46	0.4	0.53	0.65	-.16**
	(.49)	(.49)	(.5)	(.5)	(.06)	(.5)	(.49)	(.5)	(.48)	(.07)
Number of minors	1.7	1.52	1.74	1.65	-0.09	1.88	1.89	1.78	1.85	-0.07
	(1.15)	(1.08)	(1.12)	(1.1)	(.16)	(1.12)	(1.16)	(1.13)	(1.06)	(.18)
Own motorcycle	0.35	0.37	0.43	0.42	0.02	0.27	0.29	0.27	0.23	0.04
	(.48)	(.48)	(.5)	(.49)	(.07)	(.45)	(.46)	(.45)	(.42)	(.06)
Own car/truck	0.13	0.1	0.13	0.18	-0.06	0.05	0.04	0.03	0.05	-0.02
	(.33)	(.3)	(.33)	(.38)	(.05)	(.23)	(.19)	(.17)	(.21)	(.03)
Own oxen/buffalo	0.45	0.44	0.54	0.58	-0.03	0.38	0.4	0.37	0.51	-0.07
	(.5)	(.5)	(.5)	(.49)	(.08)	(.49)	(.49)	(.48)	(.5)	(.08)
Own pig	0.49	0.42	0.56	0.6	-.12*	0.45	0.48	0.42	0.55	-0.06
	(.5)	(.5)	(.5)	(.49)	(.07)	(.5)	(.5)	(.49)	(.5)	(.08)
Own ox or buffalo cart	0.22	0.22	0.29	0.31	0.01	0.18	0.23	0.17	0.26	-0.01
	(.41)	(.41)	(.46)	(.46)	(.06)	(.39)	(.42)	(.38)	(.44)	(.08)
Hard roof	0.34	0.43	0.5	0.58	0.01	0.24	0.4	0.36	0.36	.17**
	(.48)	(.5)	(.5)	(.49)	(.07)	(.43)	(.49)	(.48)	(.48)	(.08)
Hard wall	0.47	0.59	0.57	0.56	.12*	0.35	0.4	0.4	0.42	0.05
	(.5)	(.49)	(.5)	(.5)	(.07)	(.48)	(.49)	(.49)	(.49)	(.08)
Hard floor	0.85	0.82	0.83	0.92	-.11**	0.85	0.86	0.78	0.82	-0.01
	(.36)	(.39)	(.37)	(.27)	(.05)	(.36)	(.35)	(.41)	(.38)	(.05)
Have automatic toilet	0.05	0.06	0.06	0.05	0.02	0.02	0.03	0.01	0	0.02
	(.22)	(.24)	(.23)	(.22)	(.03)	(.14)	(.17)	(.12)	(.05)	(.02)
Have pit toilet	0.1	0.11	0.14	0.14	0.02	0.08	0.19	0.12	0.12	.1*
	(.3)	(.32)	(.35)	(.34)	(.05)	(.28)	(.4)	(.32)	(.32)	(.06)
Electricity	0.23	0.23	0.25	0.22	0.02	0.18	0.18	0.17	0.15	0.02
	(.42)	(.42)	(.43)	(.42)	(.07)	(.39)	(.38)	(.38)	(.35)	(.05)
Piped water	0.05	0.06	0.05	0.05	0.01	0.04	0.03	0.02	0.02	0
	(.22)	(.24)	(.22)	(.21)	(.04)	(.19)	(.17)	(.15)	(.12)	(.02)
Poverty Index (0-292)	221.83	220.91	217.84	205.24	9.93	245.18	243.56	246.17	242.13	-0.51
	(53.02)	(52.46)	(51.9)	(66.18)	(9.95)	(32.17)	(34.38)	(31.89)	(34.08)	(5.99)
Test score (0-25)	19.72	19.86	19.79	19.83	0.01	17.46	18.44	17.87	18.63	0
	(3.24)	(3.3)	(3.23)	(2.85)	(.49)	(4.87)	(5.08)	(4.69)	(4.78)	(.83)
Observations	201	153	378	417	1149	190	135	341	390	1056
Attrition rate					0.31					0.31
Joint significance: Ho: all coef. =0										
Chi-square					19.93					18.59
p-value					0.22					0.29

Notes: All variables measured at baseline. Columns 1 to 4 display the means for the control group attritors, the treatment group attritors, the control group surveyed and the treatment group surveyed. Standard deviations in parentheses. Column (5) is the difference between the treatment group mean and the control group mean among attritors minus the difference between the treatment group mean and the control group mean among respondents. Differences in means are computed by OLS regression, controlling for province fixed effects. Standard errors in parentheses are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1. The Chi-square (and corresponding p-value below) is the result of a test testing for the individual coefficients being jointly equal to 0 using seemingly unrelated estimation.

Table 3: Average Treatment Effects on Educational Outcomes

	Highest grade completed	Completed primary	Received any formal education in 2011-2017	Family index
	(1)	(2)	(3)	(4)
Panel A. Merit				
Treatment	0.213*	0.0500	0.044*	0.131**
	(0.117)	(0.036)	(0.026)	(0.065)
Observations	814	814	814	814
R-squared	0.160	0.148	0.129	0.167
F-statistic	3.240	3.610	3.420	5.040
Covariates	Yes	Yes	Yes	Yes
Control mean	5.570	0.610	0.770	-0.0200
Joint significance: Ho: all coef.				
Chi-square			3.548	
p-value			0.0600	
Panel B. Poverty				
Treatment	0.291*	0.113***	0.100**	0.264***
	(0.149)	(0.040)	(0.039)	(0.088)
Observations	753	753	753	753
R-squared	0.169	0.173	0.124	0.174
F-statistic	3.250	4.650	2.890	3.720
Covariates	Yes	Yes	Yes	Yes
Control mean	5.450	0.570	0.710	-0.160
Joint significance: Ho: all coef. =0				
Chi-square			4.107	
p-value			0.0430	
Joint test: Poverty vs. Merit				
Chi-square	0.253	2.049	2.184	2.287
p-value	0.615	0.152	0.139	0.131

Notes: Estimated treatment effects. The dependent variable in column (1) is the highest grade the individual completed and is equal to -1 if the individual received no education, 0 if he only went to kindergarten and then ranges from 1 to 11 for Grade 1 to Grade 11. In column (2), the dependent variable is a dummy equal to 1 if the individual completed primary education. In column (3), the dependent variable is equal to 1 if the individual was enrolled in the formal education system during any of the years 2011 to 2016. In column (4), the family index is the inverse covariance matrix-weighted mean of the standardized dependent variables from the three previous columns following Anderson (2008). All regressions control for district fixed effects, baseline test score, baseline poverty score, individual-level socio-economic variables from baseline, 6 school-level (EMIS) variables and 5 census village-level variables, measured at baseline. Panel A includes respondents who were eligible for the merit scholarship (Treatment=1, 0 otherwise) and Panel B respondents who were eligible for the poverty scholarship (Treatment=1, 0 otherwise). Robust standard errors are in parentheses (clustered at the school level). *** p<0.01, ** p<0.05, * p<0.1. The joint significance Chi-square (and corresponding p-value below) is a result of testing for the coefficients of individuals regressions being jointly equal to 0, using seemingly unrelated estimation. The poverty vs. merit Chi-square (and corresponding p-value below) is a result of testing for the coefficient of the merit sample and the coefficient of the poverty sample being equal.

Table 4: Average Treatment Effects on Cognition

	Math	Raven's	Forward Digit Span	Picture Recognition Vocabulary Test	Family index
	(1)	(2)	(3)	(4)	(5)
Panel A. Merit					
Treatment	0.0670	0.155**	0.0700	0.0200	0.113*
	(0.075)	(0.066)	(0.065)	(0.073)	(0.068)
Observations	814	814	813	814	813
R-squared	0.189	0.180	0.0930	0.293	0.225
F-statistic	6.130	6.670	2.350	11.02	8.710
Covariates	Yes	Yes	Yes	Yes	Yes
Control mean	0.0800	-0.0500	0	0.100	0.0700
Joint significance: Ho: all coef. =0					
Chi-square				0.861	
p-value				0.353	
Panel B. Poverty					
Treatment	0.0860	0.0520	-0.129*	0.0600	0.00500
	(0.064)	(0.080)	(0.070)	(0.081)	(0.072)
Observations	753	753	752	753	752
R-squared	0.150	0.156	0.114	0.279	0.196
F-statistic	6.610	7.370	3.450	4.780	4.190
Covariates	Yes	Yes	Yes	Yes	Yes
Control mean	-0.0600	-0.160	0.0400	-0.0500	-0.0300
Joint significance: Ho: all coef. =0					
Chi-square				1.938	
p-value				0.164	
Joint test: Poverty vs. Merit					
Chi-square	0.0530	1.304	6.455	0.195	1.812
p-value	0.818	0.254	0.0110	0.659	0.178

Notes: Estimated treatment effects. The dependent variable in column (1) is the score on the mathematics computer adaptive test, computed using Item Response Theory (IRT) with a two parameter logistic (2PL) model, standardized. In column (2), the dependent variable is the score on the Raven's matrices test computed using IRT with a 2PL model, standardized. In column (3), the dependent variable is the standardized score on the digit span test using forward items only, standardized. In column (4), the dependent variable is the score on a Picture Recognition Vocabulary Test computed using IRT with a 2PL model, standardized. In column (5), the family index is the inverse covariance matrix-weighted mean of the standardized dependent variables from the four previous columns following Anderson (2008). All regressions control for district fixed effects, baseline test score, baseline poverty score, individual-level socio-economic variables from baseline, 6 school-level (EMIS) variables and 5 census village-level variables, measured at baseline. Panel A includes respondents who were eligible for the merit scholarship (Treatment=1, 0 otherwise) and Panel B respondents who were eligible for the poverty scholarship (Treatment=1, 0 otherwise). Robust standard errors are in parentheses (clustered at the school level). *** p<0.01, ** p<0.05, * p<0.1. The joint significance Chi-square (and corresponding p-value below) is a result of testing for the coefficients of individuals regressions being jointly equal to 0, using seemingly unrelated estimation. The poverty vs. merit Chi-square (and corresponding p-value below) is a result of testing for the coefficient of the merit sample and the coefficient of the poverty sample being equal.

Table 5: Average Treatment Effects on Socioemotional Outcomes

	SDQ			Big 5					Family index
	Prosocial	Internalizing	Externalizing	Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A. Merit									
Treatment	-0.0390	-0.0770	0.00200	0.0300	-0.0910	0.00500	-0.0720	0.0580	-0.00500
	(0.066)	(0.059)	(0.070)	(0.067)	(0.062)	(0.077)	(0.066)	(0.075)	(0.065)
Observations	813	812	812	813	812	813	811	814	807
R-squared	0.0790	0.130	0.0950	0.0900	0.0600	0.0960	0.0950	0.0700	0.102
F-statistic	2.180	6.140	2.430	3.540	2	2.820	2.530	2.110	3.550
Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Control mean	-0.0300	0.0100	0.0200	0.0200	-0.0100	0	0.0100	-0.0400	0
Joint significance: Ho: all coef.									
Chi-square								0.370	
p-value								0.543	
Panel B. Poverty									
Treatment	-0.00400	-0.0420	0.0530	0.0110	-0.00200	-0.0600	-0.0780	0.186***	-0.0990
	(0.078)	(0.087)	(0.072)	(0.074)	(0.077)	(0.084)	(0.078)	(0.071)	(0.074)
Observations	751	753	753	753	752	752	752	753	749
R-squared	0.0960	0.129	0.0960	0.0720	0.0980	0.0910	0.0640	0.0840	0.107
F-statistic	2.520	5.250	2.960	2.350	6.250	2.200	1.510	2.950	2.890
Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Control mean	-0.0800	0	-0.0200	-0.0400	-0.0100	0.0100	-0.0200	-0.0200	-0.0200
Joint significance: Ho: all coef.									
Chi-square								0.003	
p-value								0.957	
Joint test: Poverty vs. Merit									
Chi-square	0.175	0.161	0.339	0.0490	1.103	0.437	0.00400	2.269	1.190
p-value	0.676	0.688	0.560	0.824	0.294	0.509	0.948	0.132	0.275

Notes: Estimated treatment effects. The dependent variable in column (1) is the score from 0 to 10 on the pro-social facet (the higher the score, the more pro-social) of the Strength and Difficulty Questionnaire (SDQ), standardized. In column (2), the dependent variables is the score from 0 to 20 on the internalizing behavior facet (the higher the score, the more externalizing behavior problems) of the SDQ, standardized. In column (3), the dependent variables is the score from 0 to 20 on the externalizing behavior facet (the higher the score, the more externalizing behavior problems) of the SDQ, standardized. In columns (4) to (8), the dependent variables are the scores from 3 to 15 on the Openness, Conscientiousness, Extroversion, Agreeableness and Neuroticism facets of the Big Five scale, standardized. In column (9), the family index is the inverse covariance matrix-weighted mean of the standardized dependent variables from the eight first columns following Anderson (2008) (scores from columns (2), (3), and (8) have been flipped beforehand). In column (10), the family index represents the first factor from an exploratory factor analysis (EFA) with quartimin rotation, on the same set of variables as in (9). All regressions control for district fixed effects, baseline test score, baseline poverty score, individual-level socio-economic variables from baseline, 6 school-level (EMIS) variables and 5 census village-level variables, measured at baseline. Panel A includes respondents who were eligible for the merit scholarship (Treatment=1, 0 otherwise) and Panel B respondents who were eligible for the poverty scholarship (Treatment=1, 0 otherwise). Robust standard errors are in parentheses (clustered at the school level). *** p<0.01, ** p<0.05, * p<0.1. The joint significance Chi-square (and corresponding p-value below) is a result of testing for the coefficients of individuals regressions being jointly equal to 0, using seemingly unrelated estimation. The poverty vs. merit Chi-square (and corresponding p-value below) is a result of testing for the coefficient of the merit sample and the coefficient of the poverty sample being equal.

Table 6: Average Treatment Effects on Labor Market Outcomes

	Currently working	Age started working	Any training since 2011	Cog. demands of main work (1/0)	Yearly earnings (inv. hyperbolic sine, USD)	Daily res. wage (inv. hyperbolic sine, USD)
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Merit						
Treatment	0.034*	0.0740	-0.0140	-0.0400	-0.294	0.0640
	(0.020)	(0.225)	(0.037)	(0.027)	(0.188)	(0.048)
Observations	772	794	814	775	791	805
R-squared	0.106	0.0720	0.156	0.0990		0.102
F-statistic	1.940	1.910	4.830	2.570	3.710	3.830
Covariates	Yes	Yes	Yes	Yes	Yes	Yes
Control mean	0.918	12.72	0.581	0.179	7.579	2.263
Panel B. Poverty						
Treatment	0.0120	0.339	-0.0310	0.0270	-0.382	0.0370
	(0.019)	(0.235)	(0.036)	(0.030)	(0.239)	(0.054)
Observations	712	726	753	713	732	746
R-squared	0.0870	0.0910	0.165	0.0900		0.118
F-statistic	2.850	1.980	7.100	2.080	959.6	2.170
Covariates	Yes	Yes	Yes	Yes	Yes	Yes
Control mean	0.924	12.48	0.577	0.153	7.614	2.222

Notes: Estimated treatment effects. The dependent variable in column (1) is a dummy equal to 1 if the individual is currently working, i.e. she worked for at least 1 hour during the last week or has a job at the moment but did not work during the last week. In column (2), the dependent variable is the age at which the individual started to work. In column (3), the dependent variable is a dummy equal to 1 if the individual participated in any formal or informal training that lasted at least one week, since 2011. In column (4), the dependent variable is a dummy equal to 1 if the main work activity demands cognitive ability (read, write, calculate, or use a computer) and 0 otherwise. In column (5), the dependent variable is the yearly earning expressed in US dollars and transformed using an inverse hyperbolic sine. In column (6), the dependent variable is the daily reservation wage in US dollars and transformed using an inverse hyperbolic sine. In column (4), values for respondents who did not work have been imputed with 0, except if they were students. In columns (1) and (4), the sample is restricted to respondents who are not currently students. Column (2) includes everyone who ever worked. Column (4) includes only people who worked over the past 12 months. Columns (3) and (6) include the entire sample. Column (1), (2), (3), (4), and (6) are estimated using OLS regression; Column (5) is estimated using Tobit regression. All regressions control for district fixed effects, baseline test score, baseline poverty score, individual-level socio-economic variables from baseline, 6 school-level (EMIS) variables and 5 census village-level variables, measured at baseline. Panel A includes respondents who were eligible for the merit scholarship (Treatment=1, 0 otherwise) and Panel B respondents who were eligible for the poverty scholarship (Treatment=1, 0 otherwise). Robust standard errors are in parentheses (clustered at the school level). *** p<0.01, ** p<0.05, * p<0.1.

Table 7: Average Treatment Effects on Socio-Economic Status and Well-being

	SES Ladder (village)	SES Index (IRT)	Life Satisfaction	Quality of Health	Quality of Life	Health Issue Index (GHQ)	Family index
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A. Merit							
Treatment	0.173**	0.186**	0.0570	0.129**	0.0410	-0.0260	0.174***
	(0.068)	(0.073)	(0.067)	(0.058)	(0.058)	(0.070)	(0.065)
Observations	814	814	814	814	814	804	804
R-squared	0.122	0.275	0.111	0.104	0.0970	0.0950	0.182
F-statistic	3.880	13.06	2.740	3.760	2.450	3.200	6.740
Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Control mean	0.0100	0.0300	-0.0400	-0.0100	-0.0200	0.0100	-0.280
Joint significance: Ho: all coef. =0							
Chi-square						7.041	
p-value						0.00800	
Panel B. Poverty							
Treatment	0.208***	-0.0640	0.0380	0.0620	0.0250	0.0540	0.0410
	(0.073)	(0.084)	(0.071)	(0.078)	(0.073)	(0.081)	(0.085)
Observations	753	753	752	753	753	744	743
R-squared	0.120	0.177	0.104	0.0730	0.0920	0.114	0.120
F-statistic	7.400	4.950	2.450	4.180	4.110	4.280	4.620
Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Control mean	-0.0800	-0.0700	-0.0200	0.0300	0	-0.0100	-0.310
Joint significance: Ho: all coef. =0							
Chi-square						8.864	
p-value						0.00300	
Joint test: Poverty vs. Merit							
Chi-square	0.179	6.899	0.0480	0.621	0.0400	0.719	1.903
p-value	0.673	0.00900	0.827	0.431	0.841	0.397	0.168

Notes: Estimated treatment effects. The dependent variable in column (1) is the score from 1 to 10 on an economic ladder as compared to people of the same age in the village, standardized. In column (2), the dependent variables is a socio-economic index constructed based on asset ownership computed using Item Response Theory with a two parameter logistic model, standardized. In column (3), the dependent variable is the score from 1 to 10 on a life satisfaction question, standardized. In column (4), the dependent variable is the score from 1 to 5 on a health quality question, standardized. In column (5), the dependent variable is the score from 1 to 5 on a life quality question, standardized. In column (6), the dependent variable is the standardized score on the General Health Questionnaire. In column (7), the family index is the inverse covariance matrix-weighted mean of the standardized dependent variables from all the previous columns following Anderson (2008). All regressions control for district fixed effects, baseline test score, baseline poverty score, individual-level socio-economic variables from baseline, 6 school-level (EMIS) variables and 5 census village-level variables, measured at baseline. Panel A includes respondents who were eligible for the merit scholarships (Treatment=1, 0 otherwise) and Panel B respondents who were eligible for the poverty scholarship (Treatment=1, 0 otherwise). Robust standard errors are in parentheses (clustered at the school level). *** p<0.01, ** p<0.05, * p<0.1. The joint significance Chi-square (and corresponding p-value below) is a result of testing for the coefficients of individuals regressions being jointly equal to 0, using seemingly unrelated estimation. The poverty vs. merit Chi-square (and corresponding p-value below) is a result of testing for the coefficient of the merit sample and the coefficient of the poverty sample being equal.

Table 8: Heterogeneous Treatment Effects by Intervention Label

	Family index: Education	Family index: Cognition	Family index: Socioemotional	Family index: SES/Well-being	Daily res. wage (inv. hyperbolic sine, USD)
	(1)	(2)	(3)	(4)	(5)
Panel A. Merit					
Treatment	0.130	0.211**	0.0770	0.139	0.173***
	(0.094)	(0.094)	(0.099)	(0.084)	(0.065)
Below the median poverty score	0.0850	0.0340	0.105	0.0420	0.0350
	(0.111)	(0.118)	(0.126)	(0.114)	(0.076)
Below the median poverty score and treatment	0.00900	-0.214	-0.175	0.0810	-0.240**
	(0.128)	(0.139)	(0.157)	(0.120)	(0.092)
Observations	814	813	807	804	805
R-squared	0.168	0.228	0.104	0.183	0.110
F-statistic	5.200	8.120	3.530	6.880	3.960
Covariates	Yes	Yes	Yes	Yes	Yes
Control mean	-0.0200	0.0700	0	-0.280	2.260
Panel B. Poverty					
Treatment	0.199*	0.0560	-0.0180	-0.00500	0.0150
	(0.103)	(0.090)	(0.105)	(0.097)	(0.062)
Below the median test score	-0.285***	-0.0720	0.0480	0.0430	0.0740
	(0.093)	(0.103)	(0.146)	(0.136)	(0.084)
Below the median test score and treatment	0.159	-0.140	-0.211	0.128	0.0630
	(0.114)	(0.131)	(0.179)	(0.165)	(0.110)
Observations	753	752	749	743	746
R-squared	0.183	0.200	0.110	0.122	0.122
F-statistic	4.040	4.400	2.830	4.370	2.330
Covariates	Yes	Yes	Yes	Yes	Yes
Control mean	-0.160	-0.0300	-0.0200	-0.310	2.220
Joint test: Poverty vs. Merit					
Chi-square	1.070	0.213	0.0310	0.0630	5.563
p-value	0.301	0.644	0.859	0.802	0.0180

Notes: Estimated treatment effects. The dependent variables in columns (1) to (3) and (5) are the family indices from Tables 3 to 5 and 7. Column (4) is the same variable as in Table 6. Treatment captures effects for students who would have qualified for a scholarship under either scheme. Below the median poverty score are individuals who qualify under the merit-based scheme but would not have received a poverty-based scholarship. Below the median test score are individuals who qualify under the poverty-based scheme but would not have received a merit-based scholarship. All regressions control for district fixed effects, baseline test score, baseline poverty score, individual-level socio-economic variables from baseline, 6 school-level (EMIS) variables and 5 census village-level variables, measured at baseline. Panel A includes respondents who were eligible for the merit scholarship (Treatment=1, 0 otherwise) and Panel B respondents who were eligible for the poverty scholarship (Treatment=1, 0 otherwise). Robust standard errors are in parentheses (clustered at the school level). *** p<0.01, ** p<0.05, * p<0.1. The Chi-square (and corresponding p-value below) is the result of testing the equality between the interaction term from the merit sample and interaction term from the poverty sample.

Table 9: Heterogeneous Treatment Effects by Gender

	Family index: Education	Family index: Cognition	Family index: Socioemotional	Family index: SES/Well-being	Daily res. wage (inv. hyperbolic sine, USD)
	(1)	(2)	(3)	(4)	(5)
Panel A. Merit					
Treatment	0.142*	0.133	-0.0180	0.188**	0.104
	(0.085)	(0.096)	(0.096)	(0.084)	(0.077)
Female	-0.00700	-0.337***	-0.195	-0.199*	-0.0200
	(0.114)	(0.110)	(0.127)	(0.118)	(0.091)
Female and treatment	-0.0220	-0.0410	0.0260	-0.0270	-0.0820
	(0.115)	(0.131)	(0.146)	(0.127)	(0.106)
Observations	814	813	807	804	805
R-squared	0.167	0.225	0.102	0.182	0.103
F-statistic	4.910	8.520	3.480	6.710	3.690
Covariates	Yes	Yes	Yes	Yes	Yes
Control mean	-0.0200	0.0700	0	-0.280	2.260
Panel B. Poverty					
Treatment	0.257***	0.168*	-0.0860	0.0420	0.118
	(0.096)	(0.098)	(0.097)	(0.119)	(0.081)
Female	-0.146	-0.254**	-0.289**	-0.247	-0.118
	(0.132)	(0.117)	(0.119)	(0.163)	(0.089)
Female and treatment	0.0120	-0.309**	-0.0250	-0.00100	-0.154
	(0.133)	(0.141)	(0.136)	(0.168)	(0.101)
Observations	753	752	749	743	746
R-squared	0.174	0.202	0.107	0.120	0.121
F-statistic	3.690	4.380	2.820	4.510	2.060
Covariates	Yes	Yes	Yes	Yes	Yes
Control mean	-0.160	-0.0300	-0.0200	-0.310	2.220
Joint test: Poverty vs. Merit					
Chi-square	0.0670	2.891	0.103	0.0220	0.345
p-value	0.796	0.0890	0.748	0.882	0.557

Notes: Estimated treatment effects. The dependent variables in columns (1) to (4) are the family indices from Tables 3 to 5 and 7. Column (5) is the same variable as in Table 6. All regressions control for district fixed effects, baseline test score, baseline poverty score, individual-level socio-economic variables from baseline, 6 school-level (EMIS) variables and 5 census village-level variables, measured at baseline. Panel A includes respondents who were eligible for the merit scholarship (Treatment=1, 0 otherwise) and Panel B respondents who were eligible for the poverty scholarship (Treatment=1, 0 otherwise). Robust standard errors are in parentheses (clustered at the school level). *** p<0.01, ** p<0.05, * p<0.1. The Chi-square (and corresponding p-value below) is the result of testing the equality between the interaction term from the merit sample and interaction term from the poverty sample.

References

Abadie, A., Athey, S., Imbens, G., & Wooldridge, J. (2017). *When Should You Adjust Standard Errors for Clustering?* (No. w24003). Cambridge, MA: National Bureau of Economic Research. https://doi.org/10.3386/w24003

Acevedo, P., Cruces, G., Gertler, P., & Martínez, S. (2016, May 25). *Soft Skills and Hard Skills in Youth Training Programs. Long Term Experimental Evidence from the Dominican Republic*. Working Paper, Washington, D.C. Retrieved from http://jobsanddevelopmentconference.org/wp-content/uploads/2016/10/CRUCES_Soft-Skills-and-Hard-Skills-in-Youth-Training-Programs-1.pdf

Anderson, K. H., Foster, J. E., & Frisvold, D. E. (2009). Investing in Health: The Long-Term Impact of Head Start on Smoking. *Economic Inquiry*, *48*(3), 587–602. https://doi.org/10.1111/j.1465-7295.2008.00202.x

Anderson, M. L. (2008). Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects. *Journal of the American Statistical Association*, *103*(484), 1481–1495. https://doi.org/10.1198/016214508000000841

Araujo, M. C., Bosch, M., & Schady, N. (2016). *Can Cash Transfers Help Households Escape an Inter-Generational Poverty Trap?* (Working Paper No. 22670). National Bureau of Economic Research. https://doi.org/10.3386/w22670

Arrow, K. J. (1973). Higher Education as a Filter. *Journal of Public Economics*, *2*(3), 193–216. https://doi.org/10.1016/0047-2727(73)90013-3

Bailey, D., Duncan, G. J., Odgers, C. L., & Yu, W. (2017). Persistence and Fadeout in the Impacts of Child and Adolescent Interventions. *Journal of Research on Educational Effectiveness*, *10*(1), 7–39. https://doi.org/10.1080/19345747.2016.1232459

Baird, S., Ferreira, F. H. G., Özler, B., & Woolcock, M. (2014). Conditional, Unconditional and Everything in Between: A Systematic Review of the Effects of Cash Transfer Programmes on

Schooling Outcomes. *Journal of Development Effectiveness*, *6*(1), 1–43. https://doi.org/10.1080/19439342.2014.890362

Barrera-Osorio, F., & Filmer, D. (2015). Incentivizing schooling for learning: Evidence on the impact of alternative targeting approaches. *Journal of Human Resources*. https://doi.org/10.3368/jhr.51.2.0114-6118R1

Barrera-Osorio, F., & Filmer, D. (2016). Incentivizing Schooling for Learning: Evidence on the Impact of Alternative Targeting Approaches. *Journal of Human Resources*, *51*(2), 461–499. https://doi.org/10.3368/jhr.51.2.0114-6118R1

Becker, G. S. (2009). *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education* (3rd ed.). Chicago: University of Chicago Press.

Belotti, F., Deb, P., Manning, W. G., & Norton, E. C. (2015). twopm: Two-Part Models. *Stata Journal*, *15*(1), 3–20.

Berry, J. (2015). Child Control in Education Decisions An Evaluation of Targeted Incentives to Learn in India. *Journal of Human Resources*, *50*(4), 1051–1080. https://doi.org/10.3368/jhr.50.4.1051

Birnbaum, A. (1968). Some Latent Trait Models and Their Use in Inferring an Examinee's Ability. *Statistical Theories of Mental Test Scores*.

Blazar, D. (2017). Validating Teacher Effects on Students' Attitudes and Behaviors: Evidence From Random Assignment of Teachers to Students. *Education Finance and Policy*, 1–52. https://doi.org/10.1162/edfp_a_00251

Blazar, D., & Kraft, M. A. (2017). Teacher and Teaching Effects on Students' Attitudes and Behaviors. *Educational Evaluation and Policy Analysis*, *39*(1), 146–170. https://doi.org/10.3102/0162373716670260

Blimpo, M. P. (2014). Team Incentives for Education in Developing Countries: A Randomized Field Experiment in Benin. *American Economic Journal: Applied Economics*, *6*(4), 90–109.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP Estimation of Ability in a Microcomputer Environment. *Applied Psychological Measurement*, *6*(4), 431–444.

Brudevold-Newman, A. (2016, November 14). *The Impacts of Free Secondary Education: Evidence from Kenya*. Job Market Paper, College Park, MD. Retrieved from econ.andrewbrudevold.com/KenyaFSE.pdf

Carneiro, P., & Ginja, R. (2014). Long-Term Impacts of Compensatory Preschool on Health and Behavior: Evidence from Head Start. *American Economic Journal: Economic Policy*, *6*(4), 135–173.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star. *The Quarterly Journal of Economics*, *126*(4), 1593–1660. https://doi.org/10.1093/qje/qjr041

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *American Economic Review*, *104*(9), 2633–2679. https://doi.org/10.1257/aer.104.9.2633

Claro, S., Paunesku, D., & Dweck, C. S. (2016). Growth mindset tempers the effects of poverty on academic achievement. *Proceedings of the National Academy of Sciences*, *113*(31), 8664–8668. https://doi.org/10.1073/pnas.1608207113

Cruz, R. C. de S., Moura, L. B. A. de, & Soares Neto, J. J. (2017). Conditional cash transfers and the creation of equal opportunities of health for children in low and middle-income countries: a literature review. *International Journal for Equity in Health*, *16*, 161. https://doi.org/10.1186/s12939-017-0647-2

Currie, J., & Thomas, D. (2000). School Quality and the Longer-Term Effects of Head Start. *Journal of Human Resources*, *35*(4), 755–774.

Deming, D. (2009). Early Childhood Intervention and Life-Cycle Skill Development: Evidence from Head Start. *American Economic Journal: Applied Economics*, *1*(3), 111–134. https://doi.org/10.1257/app.1.3.111

Deming, D. (2017). The Growing Importance of Social Skills in the Labor Market. *The Quarterly Journal of Economics*, *132*(4), 1593–1640. https://doi.org/10.1093/qje/qjx022

Doyle, A. M., Weiss, H. A., Maganja, K., Kapiga, S., McCormack, S., Watson-Jones, D., … Ross, D. A. (2011). The Long-Term Impact of the MEMA kwa Vijana Adolescent Sexual and Reproductive Health Intervention: Effect of Dose and Time since Intervention Exposure. *PLOS ONE*, *6*(9), e24866. https://doi.org/10.1371/journal.pone.0024866

Duflo, E., Dupas, P., & Kremer, M. (2017, April 13). *The Impact of Free Secondary Education: Experimental Evidence from Ghana*. Working Paper, Stanford. Retrieved from http://web.stanford.edu/~pdupas/DDK_GhanaScholarships.pdf

Dynarski, S., Hyman, J., & Schanzenbach, D. W. (2013). Experimental Evidence on the Effect of Childhood Investments on Postsecondary Attainment and Degree Completion. *Journal of Policy Analysis and Management*, *32*(4), 692–717. https://doi.org/10.1002/pam.21715

Fabregas, R. (2017). A Better School but a Worse Position? The Effects of Marginal School Admissions in Mexico City.

Filmer, D., & Schady, N. (2008). Getting Girls into School: Evidence from a Scholarship Program in Cambodia. *Economic Development and Cultural Change*, *56*(3), 581–617. https://doi.org/10.1086/533548

Filmer, D., & Schady, N. (2014). The Medium-Term Effects of Scholarships in a Low-Income Country. *Journal of Human Resources*, *49*(3), 663–694. https://doi.org/10.1353/jhr.2014.0022

Fiske, D. W. (1949). Consistency of the factorial structures of personality ratings from different sources. *The Journal of Abnormal and Social Psychology*, *44*(3), 329–344. https://doi.org/10.1037/h0057198

Fiszbein, A., & Schady, N. R. (2009). *Conditional Cash Transfers: Reducing Present and Future Poverty*. World Bank Publications.

Friedman, W., Kremer, M., Miguel, E., & Thornton, R. (2011). *Education as Liberation?* (No. w16939). Cambridge, MA: National Bureau of Economic Research. https://doi.org/10.3386/w16939

Frisvold, D. E., & Lumeng, J. C. (2011). Expanding Exposure Can Increasing the Daily Duration of Head Start Reduce Childhood Obesity? *Journal of Human Resources*, *46*(2), 373–402. https://doi.org/10.3368/jhr.46.2.373

Fryer Jr, R. G. (2011). Financial incentives and student achievement: Evidence from randomized trials. *The Quarterly Journal of Economics*, *126*(4), 1755–1798.

Fryer, R. G. J. (2017). The Production of Human Capital in Developed Countries. In A. V. Banerjee & E. Duflo (Eds.), *Handbook of Economic Field Experiments* (Vol. 2, pp. 95–322). Elsevier. https://doi.org/10.1016/bs.hefe.2016.10.002

Garces, E., Thomas, D., & Currie, J. (2002). Longer-Term Effects of Head Start. *The American Economic Review*, *92*(4), 999–1012.

García, S., & Saavedra, J. E. (2017). Educational Impacts and Cost-Effectiveness of Conditional Cash Transfer Programs in Developing Countries: A Meta-Analysis. *Review of Educational Research*, *87*(5), 921–965. https://doi.org/10.3102/0034654317723008

Gertler, P., Heckman, J., Pinto, R., Zanolini, A., Vermeersch, C., Walker, S., … Grantham-McGregor, S. (2013). *Labor Market Returns to Early Childhood Stimulation: A 20-year Followup to an Experimental Intervention in Jamaica* (No. w19185). Cambridge, MA: National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w19185.pdf

Goldberg, D., & Williams, P. (2006). *A user's guide to the General Health Questionnaire*. GL assessment.

Goldberg, L. R. (1981). Language and individual differences: The search for universals in personality lexicons. *Review of Personality and Social Psychology*, *2*(1), 141–165.

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A Research Note. *Journal of Child Psychology and Psychiatry*, *38*(5), 581–586. https://doi.org/10.1111/j.1469-7610.1997.tb01545.x

Hamoudi, A., & Sheridan, M. (2015, November). *Unpacking the Black Box of Cognitive Ability. A novel tool for assessment in a population based survey*. Manuscript, Durham, NC. Retrieved from http://thew.web.unc.edu/files/2013/08/hamoudi.pdf

Heckman, J. J., & Kautz, T. (2014). Fostering and Measuring Skills: Interventions that Improve Character and Cognition. In J. J. Heckman, J. E. Humphries, & T. Kautz (Eds.), *The Myth of Achievement Tests: The GED and the Role of Character in American Life* (pp. 341–430). Chicago: University of Chicago Press.

Heckman, J. J., Moon, S. H., Pinto, R., Savelyev, P., & Yavitz, A. (2010). A New Cost-Benefit and Rate of Return Analysis for the Perry Preschool Program: A Summary. In A. J. Reynolds, A. J. Rolnick, M. M. Englund, & J. A. Temple (Eds.), *Childhood Programs and Practices in the First Decade of Life* (pp. 366–380). Cambridge: Cambridge University Press. Retrieved from http://dx.doi.org/10.1017/CBO9780511762666.020

Jackson, C. K. (2016). *What Do Test Scores Miss? The Importance of Teacher Effects on Non-Test Score Outcomes* (Working Paper No. 22226). Cambridge, Mass.: National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w22226

Jakiela, P., Miguel, E., & te Velde, V. L. (2015). You've earned it: estimating the impact of human capital on social preferences. *Experimental Economics*, *18*(3), 385–407. https://doi.org/10.1007/s10683-014-9409-9

Kling, J. R., Liebman, J. B., & Katz, L. F. (2007). Experimental Analysis of Neighborhood Effects. *Econometrica*, *75*(1), 83–119. https://doi.org/10.1111/j.1468-0262.2007.00733.x

Kraft, M. A. (2017). Teacher Effects on Complex Cognitive Skills and Social-Emotional Competencies. *Journal of Human Resources*, 0916–8265R3.

Kremer, M., Miguel, E., & Thornton, R. (2009). Incentives to Learn. *The Review of Economics and Statistics*, *91*(3), 437–456. https://doi.org/10.1162/rest.91.3.437

Krishnan, P., & Krutikova, S. (2013). Non-cognitive skill formation in poor neighbourhoods of urban India. *Labour Economics*, *24*, 68–85. https://doi.org/10.1016/j.labeco.2013.06.004

Kyllonen, P. C., & Bertling, J. P. (2013). Innovative questionnaire assessment methods to increase cross-country comparability. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of*

international large-scale assessment: Background, technical issues, and methods of data analysis (pp. 277–285). London: Chapman & Hall.

Laajaj, R., & Macours, K. (2017). *Measuring skills in developing countries* (Working Paper No. WPS8000) (pp. 1–82). Washington, D.C.: The World Bank. Retrieved from http://documents.worldbank.org/curated/en/775311488980295780/Measuring-skills-in-developing-countries

Lang, F. R., John, D., Lüdtke, O., Schupp, J., & Wagner, G. G. (2011). Short assessment of the Big Five: robust across survey methods except telephone interviewing. *Behavior Research Methods*, *43*(2), 548–567. https://doi.org/10.3758/s13428-011-0066-z

Li, T., Han, L., Zhang, L., & Rozelle, S. (2014). Encouraging classroom peer interactions: Evidence from Chinese migrant schools. *Journal of Public Economics*, *111*(Supplement C), 29–45. https://doi.org/10.1016/j.jpubeco.2013.12.014

Ludwig, J., & Miller, D. L. (2007). Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design. *The Quarterly Journal of Economics*, *122*(1), 159–208. https://doi.org/10.1162/qjec.122.1.159

McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, *52*(1), 81–90. https://doi.org/10.1037/0022-3514.52.1.81

National Institute of Statistics, Ministry of Planning. (2010). *2008 Census CambodiaRedatam+SP*. Phnom Penh: National Institute of Statistics, Ministry of Planning.

Norman, W. T. (1967). *2800 personality trait descriptors - normative operating characteristics for a university population* (No. UM-00310-1-T). Ann Arbor: Michigan University. Retrieved from https://eric.ed.gov/?q=ED014738&id=ED014738

Parker, S. W., & Vogl, T. (2018). *Do Conditional Cash Transfers Improve Economic Outcomes in the Next Generation? Evidence from Mexico* (Working Paper No. 24303). National Bureau of Economic Research. https://doi.org/10.3386/w24303

Pritchett, L. (2013). *The Rebirth of Education: Schooling Ain't Learning*. Washington, D.C.: Brookings Institution Press for Center for Global Development.

Protzko, J. (2015). The Environment in Raising Early Intelligence: A Meta-Analysis of the Fadeout Effect. *Intelligence*, *53*, 202–210. https://doi.org/10.1016/j.intell.2015.10.006

Quek, K. F., Low, W. Y., Razack, A. H., & Loh, C. S. (2001). Reliability and validity of the General Health Questionnaire (GHQ-12) among urological patients: A Malaysian study. *Psychiatry and Clinical Neurosciences*, *55*(5), 509–513. https://doi.org/10.1046/j.1440-1819.2001.00897.x

Samejima, F. (1969). Estimation of Latent Ability Using a Response Pattern of Graded Scores. *Psychometrika*, *34*(4), Part 2.

Santorella, E. (2017, November 21). *Multi-Dimensional Teacher Effects*. Unpublished manuscript, Cambridge, MA. Retrieved from http://esantorella.com/public/full_chapter1.pdf

Silva, I. D., & Sumarto, S. (2015). How do Educational Transfers Affect Child Labour Supply and Expenditures? Evidence from Indonesia of Impact and Flypaper Effects. *Oxford Development Studies*, *43*(4), 483–507. https://doi.org/10.1080/13600818.2015.1032232

Smith, G. M. (1967). Personality correlates of cigarette smoking in students of college age. *Annals of the New York Academy of Sciences*, *142*(1), 308–321. https://doi.org/10.1111/j.1749-6632.1967.tb13733.x

Snilstveit, B., Stevenson, J., Phillips, D., Vojtkova, M., Gallagher, E., Schmidt, T., … Eyers, J. (2015). *Interventions for improving learning outcomes and access to education in low-and middle-income countries: a systematic review*. London: International Initiative for Impact Evaluation.

Sparrow, R. (2007). Protecting Education for the Poor in Times of Crisis: An Evaluation of a Scholarship Programme in Indonesia*. *Oxford Bulletin of Economics and Statistics*, *69*(1), 99–122. https://doi.org/10.1111/j.1468-0084.2006.00438.x

Spence, M. (1973). Job Market Signaling. *The Quarterly Journal of Economics*, *87*(3), 355–374. https://doi.org/10.2307/1882010

Stocking, M. L., & Lord, F. M. (1983). Developing a Common Metric in Item Response Theory. *Applied Psychological Measurement*, *7*(2), 201–210. https://doi.org/10.1177/014662168300700208

The World Bank. (2017). *Learning to Realize Education's Promise. World Development Report 2018*. Washington, D.C.: The World Bank.

van der Linden, W. J., & Pashley, P. J. (2010). Item Selection and Ability Estimation in Adaptive Testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 3–30). New York: Springer.

Vivalt, E. (2017, September 23). *How Much Can We Generalize From Impact Evaluations?* (Job market paper). Australian National University, Canberra, Australia. Retrieved from http://evavivalt.com/wp-content/uploads/How-Much-Can-We-Generalize.pdf

Walker, S. P., Wachs, T. D., Meeks Gardner, J., Lozoff, B., Wasserman, G. A., Pollitt, E., & Carter, J. A. (2007). Child development: risk factors for adverse outcomes in developing countries. *The Lancet*, *369*(9556), 145–157. https://doi.org/10.1016/S0140-6736(07)60076-2

West, M. R., Kraft, M. A., Finn, A. S., Martin, R. E., Duckworth, A. L., Gabrieli, C. F. O., & Gabrieli, J. D. E. (2016). Promise and Paradox: Measuring Students Non-Cognitive Skills and the Impact of Schooling. *Educational Evaluation and Policy Analysis*, *38*(1), 148–170. https://doi.org/10.3102/0162373715597298

Appendix

Appendix A1: Additional checks of validity and robustness of findings

Table A1: Balance of Estimation Sample at Baseline

	Merit Scholarship					Poverty Scholarship				
	n	All	Treatment	Control	Difference	n	All	Treatment	Control	Difference
	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
Female	767	0.502	0.505	0.499	0.006	728	0.593	0.646	0.533	0.098***
		(0.5)	(0.501)	(0.501)	(0.036)		(0.492)	(0.479)	(0.5)	(0.035)
Number of minors	749	1.689	1.645	1.736	-0.105	712	1.815	1.845	1.779	0.075
		(1.111)	(1.102)	(1.12)	(0.128)		(1.092)	(1.059)	(1.129)	(0.127)
Own motorcycle	756	0.425	0.423	0.426	0.003	723	0.249	0.229	0.271	-0.022
		(0.495)	(0.495)	(0.495)	(0.055)		(0.433)	(0.421)	(0.445)	(0.042)
Own car/truck	743	0.152	0.177	0.125	0.035	726	0.039	0.047	0.029	0.008
		(0.359)	(0.382)	(0.332)	(0.036)		(0.193)	(0.211)	(0.169)	(0.022)
Own oxen/buffalo	762	0.56	0.58	0.539	0.017	723	0.448	0.514	0.372	0.087*
		(0.497)	(0.494)	(0.499)	(0.058)		(0.498)	(0.5)	(0.484)	(0.052)
Own pig	764	0.581	0.604	0.557	0.034	724	0.493	0.554	0.423	0.097*
		(0.494)	(0.49)	(0.497)	(0.05)		(0.5)	(0.498)	(0.495)	(0.056)
Own ox or buffalo cart	753	0.3	0.307	0.293	-0.011	715	0.218	0.26	0.171	0.054
		(0.459)	(0.462)	(0.456)	(0.047)		(0.413)	(0.439)	(0.377)	(0.047)
Hard roof	752	0.54	0.581	0.496	0.078	714	0.36	0.36	0.36	-0.013
		(0.499)	(0.494)	(0.501)	(0.052)		(0.48)	(0.481)	(0.481)	(0.052)
Hard wall	764	0.565	0.563	0.568	-0.015	723	0.408	0.416	0.399	0.002
		(0.496)	(0.497)	(0.496)	(0.053)		(0.492)	(0.494)	(0.49)	(0.062)
Hard floor	753	0.878	0.921	0.831	0.075**	720	0.804	0.825	0.78	0.012
		(0.328)	(0.271)	(0.375)	(0.035)		(0.397)	(0.381)	(0.415)	(0.046)
Have automatic toilet	748	0.053	0.049	0.058	-0.01	720	0.008	0.003	0.015	-0.012
		(0.225)	(0.217)	(0.234)	(0.023)		(0.091)	(0.051)	(0.122)	(0.01)
Have pit toilet	748	0.136	0.135	0.138	0.005	720	0.118	0.119	0.117	0.013
		(0.343)	(0.342)	(0.345)	(0.042)		(0.323)	(0.324)	(0.322)	(0.036)
Electricity	766	0.238	0.225	0.251	-0.026	725	0.157	0.145	0.171	-0.024
		(0.426)	(0.418)	(0.434)	(0.047)		(0.364)	(0.353)	(0.377)	(0.041)
Piped water	760	0.047	0.046	0.049	-0.001	719	0.019	0.016	0.024	-0.01
		(0.213)	(0.21)	(0.216)	(0.022)		(0.138)	(0.125)	(0.152)	(0.013)
Poverty Index (0-292)	795	211.234	205.245	217.841	-9.864	731	244.015	242.133	246.167	-1.11
		(60.108)	(66.176)	(51.899)	(8.175)		(33.115)	(34.078)	(31.892)	(4.479)
Test score (0-25)	795	19.814	19.832	19.794	0.134	731	18.275	18.633	17.865	0.951
		(3.036)	(2.851)	(3.232)	(0.467)		(4.753)	(4.784)	(4.691)	(0.684)
Joint significance: Ho: all coef. =0										
Chi-square					18.62					15.99
p-value					0.29					0.45

Notes: Column (1) presents the number of observations in the analysis sample (excluding observations with imputed baseline information). Colums (2) to (4) display the means for the full sample, the treatment group, and the control group, respectively. Standard deviations in parentheses. Column (5) is the difference between the treatment group mean and the control group mean. Differences in means are computed by OLS regression, controlling for province fixed effects. Standard errors in parentheses are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1. The Chi-square (and corresponding p-value below) is the result of a test testing for the individual coefficients being jointly equal to 0 using seemingly unrelated estimation.

Table A2: Balance of Estimation Sample at Three-year Follow-up

	Merit Scholarship					Poverty Scholarship				
	n	All	Treatment	Control	Difference	n	All	Treatment	Control	Difference
	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
Female	735	0.495	0.505	0.484	0.022	660	0.577	0.64	0.505	0.123***
		(0.5)	(0.501)	(0.5)	(0.034)		(0.494)	(0.481)	(0.501)	(0.034)
HH size	735	7.117	7.204	7.02	0.296	660	7.071	6.858	7.316	-0.34
		(2.572)	(2.502)	(2.647)	(0.238)		(2.397)	(2.149)	(2.636)	(0.235)
Own motorcycle	735	0.645	0.624	0.669	-0.007	660	0.536	0.521	0.554	0.02
		(0.479)	(0.485)	(0.471)	(0.041)		(0.499)	(0.5)	(0.498)	(0.037)
Own car/truck	735	0.022	0.021	0.023	0.001	660	0.024	0.014	0.036	-0.015
		(0.146)	(0.142)	(0.15)	(0.012)		(0.154)	(0.118)	(0.186)	(0.013)
Own oxen/buffalo	735	0.574	0.613	0.53	0.048	660	0.526	0.578	0.466	0.076
		(0.495)	(0.488)	(0.5)	(0.045)		(0.5)	(0.495)	(0.5)	(0.051)
Own pig	735	0.608	0.649	0.562	0.079	660	0.586	0.618	0.55	0.054
		(0.488)	(0.478)	(0.497)	(0.049)		(0.493)	(0.487)	(0.498)	(0.057)
Own ox or buffalo cart	735	0.253	0.263	0.242	-0.002	660	0.239	0.283	0.189	0.074
		(0.435)	(0.441)	(0.429)	(0.049)		(0.427)	(0.451)	(0.392)	(0.049)
Hard roof	735	0.849	0.84	0.859	-0.009	660	0.77	0.737	0.808	-0.061
		(0.358)	(0.367)	(0.349)	(0.031)		(0.421)	(0.441)	(0.395)	(0.041)
Hard floor	734	0.977	0.987	0.965	0.019	659	0.97	0.974	0.964	0
		(0.151)	(0.113)	(0.183)	(0.012)		(0.172)	(0.158)	(0.186)	(0.015)
Have pit toilet	735	0.049	0.039	0.061	-0.012	660	0.055	0.051	0.059	-0.007
		(0.216)	(0.193)	(0.239)	(0.025)		(0.227)	(0.22)	(0.235)	(0.026)
Electricity	735	0.427	0.438	0.415	0.022	660	0.352	0.351	0.352	0.005
		(0.495)	(0.497)	(0.493)	(0.051)		(0.478)	(0.478)	(0.478)	(0.053)
Piped water	735	0.287	0.289	0.285	-0.008	660	0.306	0.28	0.336	-0.069
		(0.453)	(0.454)	(0.452)	(0.054)		(0.461)	(0.45)	(0.473)	(0.057)
SES Index (2PL)	735	0.187	0.198	0.175	0.065	660	-0.098	-0.128	-0.063	0.004
		(0.85)	(0.891)	(0.804)	(0.09)		(0.82)	(0.836)	(0.802)	(0.075)
Joint significance: Ho: all coef. =0										
Chi-square					13.39					30.74
p-value					0.42					<.01

Notes: Column (1) presents the number of observations in the analysis sample (excluding observations with imputed baseline information). Colums (2) to (4) display the means for the full sample, the treatment group, and the control group, respectively. Standard deviations in parentheses. Column (5) is the difference between the treatment group mean and the control group mean. Differences in means are computed by OLS regression, controlling for province fixed effects. Standard errors in parentheses are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1. The Chi-square (and corresponding p-value below) is the result of a test testing for the individual coefficients being jointly equal to 0 using seemingly unrelated estimation.