The impact of Enseña por México on student socioemotional skills

November 2017

Microanalitica

Acknowledgements

The evaluation presented in this report was possible thanks to the work of many people. We are very grateful to Andrés Peña for his tireless support in coordinating with educational authorities and providing information from Enseña por México. Robbie Dean and Laura Lewis played the essential role of coordinating the many facets of the project. We appreciate the comments and insights of Hunter Gelbach, Andy Sokatch, Koji Miyamoto, Patrick Kyllonen and Daniel Santos on the measurement of socioemotional skills, and the feedback on preliminary versions of this report form Dave Evans, Patrick McEwan, and Tim Kautz. All errors and omissions are our responsibility.

Executive summary

This report presents the results of an evaluation of the impact of Enseña por México on student socioemotional skills in the academic year 2016-17 in the states of Baja California, Chiapas, Hidalgo and Puebla. Using self-applied student questionnaires, we measured the four socioemotional scales of the CORE Districts Survey (self-management, growth mindset, self-efficacy and social awareness), grit, and locus of control. We also measured six behaviors and attitudes that could be interpreted as proxies of socioemotional skills: educational expectations, perceived general and pecuniary returns on education, tardiness and absenteeism, time devoted to homework, and community involvement.

We use a difference-in-differences approach to compare growth in the metrics of impact between the start and the end of the academic year across treated and non-treated students. Some non-treated students are in treated schools, while others are in comparable, non-treated or control schools. Control schools were found using a Coarsened Exact Matching technique. The sample includes over 56,000 student observations (baseline plus endline) in 1,194 classrooms in grades 4-12 in 328 schools.

We find that exposure to Enseña por México fellows is associated with short-run improvements in socioemotional skills in Secundaria (grades 7-9), as measured by the CORE scales. There is also evidence of a reduction in tardiness and absenteeism in all educational levels analyzed. The magnitudes are modest—the main estimates are below 0.15 standardized units. These results are in line with evaluations of other members of the Teach For All network.

We also measured Tripod's 7Cs (a set of scales of effective teaching that are predictive of students' academic achievement) in the student surveys, and teaching practices and attitudes using instructor surveys. Using those two sources we found that Enseña por México fellows differ from regular teachers. Fellows score between 0.15 and 0.30 standard deviations higher than regular teachers in Tripod's 7Cs. At the same time, fellows give more importance to developing student curiosity instead of student competencies. They think it is more important for students to set ambitious goals instead of making concrete, achievable plans. Fellows engage in extracurricular activities, use evaluations, and give feedback to students to a larger extent than regular teachers. In terms of the Big Five personality traits, fellows are more open to new experiences, more extraverted, and more agreeable.

In sum, the results indicate that Enseña por México fellows are more effective than regular teachers, and that they help foster student socioemotional skills.

1 Introduction

Teach For All is both a global network of 46 independent, locally led and governed partner organizations, inspired by the Teach For America and Teach First UK models, and a global organization that works to accelerate the progress of the network. According to Teach For All, these organizations "recruit and develop promising future leaders to teach in their nations' under-resourced schools and communities and, with this foundation, work with others, inside and outside of education, to ensure all children are able to fulfill their potential." Throughout we refer to Teach For All recruits as fellows. In Mexico, the member organization is Enseña por México. In the summer of 2016, Teach For All hired Microanalitica through a competitive process to conduct a short-term evaluation of the impact of Enseña por México fellows on student socioemotional skills. This report summarizes the methodology and the results of the evaluation carried out in the academic year 2016-17.

The term "socioemotional skills" encompasses concepts that, depending on the context, are referred to by different names, such as non-cognitive skills, soft skills, character skills, life skills or 21st century skills (Sanchez *et al.*, 2016). Socioemotional skills are: "(a) conceptually independent from cognitive ability, (b) generally accepted as beneficial to the student and to others in society, (c) relatively rank-order stable over time in the absence of exogenous forces […], (d) potentially responsive to intervention, and (e) dependent on situational factors for their expression" (Duckworth and Yeager, 2015). Socioemotional skills have become a central issue in the discussion on how to improve educational outcomes. A rich body of literature has established that these skills matter (Sanchez *et al*., 2016). Indeed, they may be as important as intelligence in determining academic and professional success (Heckman, Stixrud, and Urzúa, 2006). However, many questions remain about the extent to which these skills are malleable, exactly how they can be cultivated, and how to properly measure them for evaluation purposes (Farrington *et al*., 2012).

Although there is evidence indicating that socioemotional skills can be cultivated (Sanchez *et al*., 2016), results are hard to generalize. Some studies estimate the impact of programs specifically aimed at enhancing socioemotional skills of a well-defined target population, while others estimate the impact of general interventions such as attaining more years of schooling. The evaluation presented here falls somewhere between those two points. Enseña por Mexico's goal is to expand quality education and opportunities for all children, and as part of that, to also promote socioemotional learning. The impact of such intervention on student socioemotional skills cannot be inferred from studies like those mentioned above.

Two lines of research are relevant for situating the context of this evaluation. The first line includes studies that have produced evidence on whether teachers can have an impact on student socioemotional skills, and how a teacher's ability to enhance socioemotional skills relates to the ability to improve academic performance, as measured by test scores. In some instances, socioemotional skills are proxied by "non-test outcomes" such as unexcused absences or suspensions. Some studies have found that teachers' abilities to foster socioemotional learning and academic performance show little correlation (Kraft, 2017; Jennings and DiPrete, 2010; Jackson, 2016; Blazar and Kraft, 2017). In other words, the teachers who are good at adding value in terms of test scores are usually not the same ones who are good at fostering socioemotional skills. However, there is some evidence of a positive

correlation (Rusek *et al*., 2015; Ladd and Sorensen, 2017). In the context of Enseña por México, this means that even if the organization's fellows are good at improving academic performance, it is an open question whether they are good at fostering socioemotional skills.

The second relevant line of research consists of evaluations of Teach For All members. Most of the evaluations publicly available are for Teach For America, and the majority focus on academic performance as the metric of impact. Except for one study that finds no effects on academic achievement (Kane, Rockoff and Staiger, 2008), the studies of Teach For America find that fellows have a positive impact on math achievement, but not on reading (Xu, Hannaway and Taylor, 2011; Chiang, *et al*. 2017; Antecol, Eren and Ozbeklik, 2013). Other studies that focused on socioemotional skills and related behaviors, reveal mixed results. Some found decreases in unexcused absences and suspensions (Backes and Hansen, 2017), while others found no effect on attendance, promotion or disciplinary incidents (Glazerman, Mayer and Decker, 2006). There are two evaluations of other members of the Teach For All network. A study of Teach First UK found evidence indicative of improvements in test scores (Allen and Allnutt, 2013), and a study of Enseña Chile found positive effects on test scores, self-esteem and self-efficacy (Alfonso, Santiago and Bassi, 2010).

The present evaluation builds on previous studies and sheds light on whether the instructors selected and trained by Enseña por México have a positive impact on student socioemotional skills in the short term—specifically, one academic year. Enseña por México selects and trains its fellows to provide quality education and expand opportunities for students. Although comparisons with the evaluation results for Teach For America are inevitable, it must be noted that Enseña por México operates in a very different environment. Mexico is a middle-income country (its GDP per capita is one third that of the US.) Education is not organized in the same way. Teachers are trained and paid differently (OCDE, 2017). Whatever is observed for Teach For America may not necessarily be applicable to Mexico or any other country where the Teach For All network is present.

The way Enseña por México operates made a quasi-experimental difference-in-differences approach the best option for an impact evaluation. The difference-in-differences method "compares the changes in outcomes over time between a population that is enrolled in a program (the treatment group) and a population that is not (the comparison group)" (Gertler *et al*., 2011, p.95). Given a set of schools where Enseña por Mexico fellows were to be deployed in the academic year 2016-17 (the treatment schools), we found a set of similar schools to use as a comparison group (the control group.) Treatment schools included three educational levels: Primaria (grades 1-6), Secundaria (grades 7-9), and Bachillerato (grades 10-12).

As metrics of impact, we used six measures of socioemotional skills and six measures of attitudes or behaviors related to those skills. Four socioemotional scales were taken from the CORE Districts Survey: self-management, growth mindset, self-efficacy and social awareness. Those scales are positively related to academic performance and positive behaviors (West, 2016). We also included the scales of grit (Duckworth and Quinn, 2009) and academic locus of control (Arocha and Lezama, 2007). The scales for attitudes or behaviors related to socioemotional skills that we measured are: educational expectations, perceived pecuniary and general returns on education, tardiness and absenteeism in school, time devoted to homework, and community involvement. Those scales can be thought of as proxies for

socioemotional skills, but are more easily malleable and directly interpretable. All the scales rely on students' responses to questions about their attitudes and behaviors. They were measured at the beginning and the end of the academic year in treatment and control schools among students in grades 4-12. Our impact estimates are the result of comparing growth in socioemotional skills between students who were exposed to the fellows and students who were not exposed. For our estimation approach to be valid, "the comparison group must accurately represent the change in outcomes that would have been experienced by the treatment group in the absence of treatment" (Gertler *et al*., 2011, p.96). Hence the importance of finding comparable schools for the control group.

As part of the evaluation, Teach For All also requested a comparison of Enseña por México fellows and regular teachers within the same schools, as well as in different but comparable schools. For that purpose, we applied an instructor questionnaire to elicit teaching values and practices. In the student questionnaire we included some scales that measure effective teaching. Information about instructors' values and practices is crucial for gauging the room for potential impact. Positive differences between fellows and regular teachers may indicate that fellows can make a positive difference, and the larger the difference, the more room for impact. Additionally, the information provided by scales on teaching values and practices may help Enseña por México learn what works better and adapt its practices accordingly.

The evaluation produced two main findings. First, Enseña por México fellows differ from regular teachers in their teaching values and attitudes. Fellows give more importance to developing student curiosity instead of student competencies. They think it is more important for students to set ambitious goals instead of making concrete, achievable plans. Fellows engage in extracurricular activities, use evaluations, and give feedback to students to a larger extent than regular teachers. In terms of the Big Five personality traits (Goldberg, 1993), fellows are more open to new experiences, more extraverted, and more agreeable. Fellows seem to follow practices more conducive to better academic achievement, as measured by student surveys. They score between 0.15 and 0.30 standard deviations higher than regular teachers in Tripod's 7Cs, which are seven scales predictive of students' academic gains (Kane, McCaffrey, Miller and Staiger, 2013).

Second, we find evidence that exposure to fellows is associated with modest short-run improvements in socioemotional skills in Secundaria. We also find evidence of a reduction in tardiness and absenteeism in all educational levels. The magnitudes are modest—below 0.15 standardized units.

The report is organized as follows. Section 2 explains Enseña por México's treatment and lays out its theory of change. Section 3 presents the empirical strategy to compare fellows and regular teachers, estimate the impact of fellows on student socioemotional skills, and assess whether there is reference bias. Section 4 describes the questionnaires used, which are included in the online Appendix (microanalitica.com/ExM/Appendix). Section 5 describes the process employed to collect the data and the main features of the sample of analysis. Section 6 presents the results of comparing fellows and regular teachers, the estimates of impact, and the assessment of reference bias presence. Our findings and its implications are discussed in section 7.

2 Treatment

Enseña por México "selects and trains talented and driven college graduates to teach for two years in underperforming schools." Its fellows "are meant to be excellent instructors, interacting with their students inside and outside the classroom, and to be role-models." Fellows are recruited via an open call—anyone interested may apply. Through a competitive process involving tests and interviews, Enseña por México selects the best candidates based on several attributes. Among them are: leadership, perseverance, critical thinking, motivation and influence. Between 2013 and 2017, Enseña por México received 8,137 applications and hired 619 fellows—8% of the total. Two thirds of fellows active in the academic year 2016-17 are women, and roughly one third are Psychology or Education majors. Their ages range between 22 and 32 years, with an average of 26.

Selected candidates attend an intensive four-week training course to "develop basic skills and acquire tools to lead the classrooms they will teach." Once fellows join schools, Enseña por México offers them continuous training and tutoring. At the start of the academic year 2016-17, 232 fellows were teaching in 12 states in grades K-12.

A crucial part of the operation of Enseña por México is the selection of the schools where the fellows teach. The selection is made through case-by-case negotiations with state educational authorities. The intention is to send fellows to schools in need of high-quality instructors. It is not unusual for fellows to be sent to schools facing staff shortages.

Fellows teach in different educational levels across diverse types of schools. In the academic year 2016-17, Enseña por México sent some fellows to a system of Primarias (elementary schools, grades 1-6) administered by the National Council for Education Development (CONAFE). CONAFE schools operate in communities with fewer than 30 students. Those communities are in remote areas and lack basic services, such as electricity, running water or cellphone signal. Frequently, regular teachers in these schools are graduates of the schools where they teach and have no postsecondary education. Fellows sent to CONAFE schools usually stay with the families of the students—there are no restaurants, grocery stores or lodging options in those communities. They split their time between two CONAFE schools and switch locations every other week. Their role is to provide tutoring to underperforming students; their job title is "itinerant pedagogic counselor." In Chiapas, one of the two states where fellows worked in CONAFE schools, tensions between unionized teachers from regular schools and authorities have been a fixture in recent years. A teacher strike accompanied by road blocks prevented the 2016-17 academic year from starting normally ("Mexican Dissident Teachers Strike as School Year Starts" by Anthony Harrup, Aug. 22, 2016, *The Wall Street Journal*). In this context, it is fair to say that fellows sent to CONAFE schools face harsh conditions and retention in the program is a challenge.

Some fellows are sent to regular Primaria schools, where they assume the role of regular teachers. Others are sent to Secundaria schools (grades 7-9) or Bachillerato schools (grades 10-12). In those grades each subject could be taught by a different teacher. As a consequence, although fellows are full-time instructors and teach an average of 13 hours per week, they usually teach only a subset of the subjects across several classrooms. In Secundaria and Bachillerato, fellows usually teach Math, English or

Science. A distinction between Secundaria and Bachillerato is that courses are organized by semesters in the latter. Thus, exposure to fellows may be more heterogeneous in Bachillerato.

In Bachillerato and Secundaria, the distribution of each fellow's workload across subjects and classrooms is ultimately decided by the school principal. As a result, there is idiosyncratic variation in the level of exposure even among treated students in the same school. For instance, within the same school, students in one classroom may take one course with a fellow, while students in a different classroom may take three courses with the same fellow. The latter would have three times the exposure of the former.

2.1 Theory of change

The goal of Enseña por México is "to expand quality education and opportunity for all children and, as part of that, also promote socioemotional learning." The organization believes its fellows "are excellent instructors and a positive influence, who work with others inside and outside the classroom that can transform the way students see their world and themselves." For Enseña por México, "the mindsets and competencies used to screen its fellows, plus the training and support they receive from the organization, make them strong classroom leaders and determined advocates of their students." Thus, by being exposed to fellows—instead of regular teachers—students are more likely to change their academic mindsets and behaviors in a positive way.

The theory of change is consistent with the evidence of teachers having the ability to change student socioemotional skills and related attitudes or behaviors in the short-run (Kraft, 2017; Jennings and DiPrete, 2010; Jackson, 2016; Blazar and Kraft, 2017; Rusek *et al*., 2015; Ladd and Sorensen, 2017). An implicit assumption in the theory of change is that, on average, fellows are more effective at promoting socioemotional learning than regular teachers. Evaluations of Enseña Chile and Teach for America lend support to this assumption. They have found positive effects of fellows on self-esteem and self-efficacy, and negative effects on unexcused absences and suspensions (Alfonso, Santiago and Bassi, 2010; Backes and Hansen, 2017). However, there is also evidence from Teach For America showing no significant impact of fellows on attendance, promotion or disciplinary incidents (Glazerman, Mayer and Decker, 2006).

2.2 Prior evidence

Two previous studies provided some cross-sectional evidence that supports Enseña por México's theory of change. In the 2014-15 academic year, Microanalitica conducted a study in the state of Puebla using data from 6,889 students in 37 participating high schools and 71 non-participating high schools. Relative to teachers in the same schools and in comparable schools, fellows obtained average scores roughly 0.25 standard deviations higher in six of Tripod's 7Cs, and students exposed to fellows scored roughly 0.10 standard deviations higher in Gallup's engagement and wellbeing scales (Chacón and Peña, 2015).

In the 2015-16 academic year, Microanalitica carried out another study comparing 125 fellows and 125 quasi-randomly selected regular teachers teaching to the same 3,249 students in Middle and High schools. Regular teachers were included in the comparison group if they were teaching Monday's earliest class—excluding fellows' classes. In direct comparisons of fellows and regular teachers, students

ranked the former above the latter, and fellows scored 0.30-0.45 standard deviations higher in Tripod's 7Cs (Chacón and Peña, 2016).

The two studies suggest that Enseña por México fellows make a positive difference in the academic environment of the students they serve. However, it is an open question whether that difference translates into meaningful changes in student socioemotional skills. Hence the interest in further evidence.

3 Empirical strategy

The empirical strategy has three parts. First, we explore whether the evidence that we collected is in line with the assumptions of the theory of change—namely, that fellows provide a quality education, as well as support and opportunities for their students. Second, we estimate the impact of exposure to fellows on student socioemotional skills—this is the core of the analysis. Third, we assess whether there is a reference bias that could be affecting the impact estimation—one of the chief concerns that Teach For All expressed about the evaluation.

3.1 Empirical support to the theory of change

To test whether Enseña por México fellows appear to be different from regular teachers in the same schools or in comparable schools, we compare scales that measure effective teaching known as the Tripod's 7Cs. We also compare sociodemographic traits (e.g., gender, age, experience) and attitudes and practices (e.g., the percent of the days the instructor gives homework.) Those scales are explained in more detail in section 4.

By comparing fellows and regular teachers, we attempt to answer three questions. The first question is whether fellows differ from regular teachers in self-reported traits, attitudes and practices. The second question is whether fellows score higher than regular teachers in the Tripod's 7Cs. We test those hypotheses with t-tests for numerical or Likert-type scales. The third question is which self-reported scales (i.e., traits, attitudes and practices) explain differences between fellows and regular teachers in the Tripod's 7Cs—if any differences are found. To answer the third question, we follow four steps. First, we estimate a model in which each of the Tripod's 7Cs is a linear function of instructors' self-reported traits, attitudes and practices. Using only data for regular teachers, we estimate the following specification for each of the Tripod's 7Cs:

$$c_i = \mu + \lambda Z_i + \upsilon_i \qquad (1)$$

The variable c_i denotes one of the Cs for instructor i. The vector Z_i denotes self-reported instructor scales, which includes traits, attitudes and practices. Second, we compute differences in the averages for each of the components of Z_i between fellows and regular teachers. The differences in averages are denoted by the vector ΔZ. Third, we multiply the estimates of the vector λ by the vector ΔZ. The sign and magnitude of each term of the product $\lambda \Delta Z$ is informative of the source of the difference between fellows and regular teachers in each of the Tripod's 7Cs. Fourth, we compare predicted average differences (using the estimates of λ and ΔZ) and observed differences to assess whether there are

unobserved differences between fellows and regular teachers—not attributable to the self-reported instructor scales.

3.2 Impact estimation

The impact evaluation involved the application of questionnaires at two points in time, one at the beginning of the academic year, before exposure to fellows occurred in treatment schools (baseline), and another at the end of the academic year, after exposure to fellows took place (endline). Questionnaires were applied to all students in treatment and non-treatment schools in the sample of analysis.

In the difference-in-differences approach we adopted, the metric of impact is given by different socioemotional scales and some behaviors related to those scales. The first difference is across periods: baseline versus endline. The second difference is between students exposed to fellows and students not exposed to fellows. The difference-in-differences approach identifies the impact of exposure to fellows assuming that, absent the treatment, the metrics of impact would behave similarly in the treatment and control groups.

In principle, it is possible that not all students in treated schools (referred to as "EPM schools" throughout) are exposed to fellows during the period of analysis. As we mentioned in section 2, in schools with many classrooms, fellows may teach only a subset of those classrooms. That possibility provides an additional source of variation. We have an external control group given by students in non-EPM schools. We also have an internal control group given by students in EPM schools who were not exposed to fellows in the period of analysis. We use those two groups for our main estimates.

There are multiple ways to implement difference-in-differences estimators. The simplest way involves linking baseline and endline questionnaires for each student. That straightforward way could take the following form:

$$y_{i1} - y_{i0} = \alpha + \gamma X_i + \beta \tau_i + \varepsilon_i \qquad (2)$$

Where y_{i1} is the metric of impact for student i in the endline, and y_{i0} is the metric of impact for the same student in the baseline. The vector X_i represents controls such as the socio-economic status of the student, and scores in the Big Five personality traits. The variable τ_i denotes the treatment status: it takes the value of one for students exposed to fellows, and zero for students not exposed to fellows. The coefficient β is the parameter of interest: the impact of being exposed to fellows. The estimation of (2) poses a practical challenge: it requires identifying and linking students across baseline and endline questionnaires.

Any questionnaire requesting personal information from students (e.g., birthdate, full name, population registry number) could turn into an obstacle for the evaluation. Asking information that allows the identification of students is likely to face resistance from students, parents, teachers or principals. To avoid such a potential obstacle, we opted for a pragmatic alternative. In both the baseline and endline questionnaires we did not include any information identifying students. However, we did include

information to identify grade and group (i.e., classroom.) That information allows us to implement the difference-in-differences approach with an alternative yet equivalent specification.

The alternative to specification (2) that does not require linking every student's baseline and endline information is given by a fixed classroom-effect specification:

$$y_{it} = \theta_t + \eta_{g(i)} + \gamma X_{it} + \beta \tau_i + \varepsilon_{it} \quad (3)$$

Where y_{it} is the metric of impact for student i in period t. In this specification, the baseline and endline data are stacked. If we have N students in the sample and all of them are surveyed twice, then specification (3) has $2N$ observations. The coefficient θ_t denotes a fixed time-effect, which is meant to capture the general trend in the metric of impact. The coefficient $\eta_{g(i)}$ denotes classroom-fixed effects, that is, fixed effects for each combination of grade, group and school, denoted by $g(i)$. In this case τ_i has a slightly different definition: it is a dummy variable that takes the value of zero when the student has not been exposed to fellows, and a value of one when the student has been exposed to fellows. The coefficient β is the parameter of interest: the impact of exposure to fellows. The error term ε_i is clustered by classroom.

It is important to clarify the meaning of "classroom" in the context of this evaluation. In CONAFE schools, all students attend the same classroom because schools have only one room. Since we do not know who is exposed to a fellow within each CONAFE school, if a fellow tutored some students of a school, we count the entire school as treated. Thus, there is no distinction between school and classroom. In regular Primaria, a classroom is defined as a combination of grade (4 to 6) and group (usually denoted A, B, C, etc.) A combination of grade and group (e.g. 4B) denotes both a roster of students and a physical space—a classroom—in the school where those students take classes. In Secundaria and Bachillerato, the combination of grade (7-9 and 10-12) and group also defines a roster of students and a physical space. Since students in Secundaria and Bachillerato have different instructors teaching different subjects, instructors go from one classroom to the other to teach their classes. Thus, just as in regular Primaria, a classroom denotes a roster of students that together takes the same courses in the academic year in the same physical space. Students rarely switch across classrooms within the same academic year.

We estimate equation (3) using different socioemotional scales as well as the scales that measure attitudes and behaviors that can be interpreted as proxies of socioemotional skills. Since the sample of analysis incudes students in grades 4-12, we estimate (3) for the whole sample and by educational level: Primaria (grades 4-6), Secundaria (grades 7-9), and Bachillerato (grades 10-12).

As we explain in section 5, there are two issues observed in the data that affect our empirical strategy. The first issue is treatment heterogeneity across schools and within the same school. Exposure to fellows differs across treated classrooms. To incorporate this heterogeneity in our strategy we estimate specification (3) for Secundaria and Bachillerato using two alternative definitions of treatment: the number of fellows that taught the classroom, and the number of subjects taught to the classroom by fellows.

The second issue is imperfect treatment fidelity. Given that fellows may face harsh conditions—especially in CONAFE schools—retention is a challenge. Also, conditions in the field may change, and fellows may be sent to a different school from the originally planned. To address this issue, we estimate equation (3) in two ways: by Ordinary Least Squares using actual treatment status as the explanatory variable, and by Two-Stage Least Squares, instrumenting actual treatment status with assigned treatment status (when the sample was designed.) The instrumental variable is defined at the school level because that is all that is known in advance (before the academic year starts.)

3.3 Longitudinal reference bias

A concern in the measurement of socioemotional skills brought up by Teach For All is reference bias (Duckworth and Yeager, 2015). Most measures of socioemotional skills are based on subjective assessments susceptible to social desirability and reference bias. In our context, social desirability may bias students' answers because they may want to give answers that make them look better in the eyes of the person asking the questions. This type of bias is not a problem if social desirability is time-invariant. Since we adopt a difference-in-differences strategy, we are not concerned about the levels of the scales being inflated. We only care about changes in those scales in time—between endline and baseline. The same can be said about reference bias caused by different students using different references to judge themselves—what we can call "cross-sectional reference bias." Since we analyze changes in the scales between two points in time holding constant the pool of students, whatever reference is used is purged from the change in the scale.

The source of real concern is "longitudinal reference bias" caused by the treatment. Exposure to fellows may change the reference point used for the self-assessment. For instance, suppose we want to measure whether a student works hard. In a baseline survey, the student must answer "yes" or "no" to the question "Do you work hard?" Based on her experience, she answers "yes." Then the student is exposed to a fellow who convinces her that she should work much harder to achieve her goals, and she starts working harder. In an endline survey, after exposure to the fellow, the same student is asked the same question. Although she works harder than before, she may consider the new level of work not being hard enough, and answer "no." Her answers may suggest that exposure to the fellow had a negative effect on hard work, when it had the opposite effect.

To empirically determine whether there is longitudinal reference bias in the context of this evaluation, in the student endline and baseline questionnaires we included several items referring to a sibling. In our view, those items could be affected by longitudinal changes in the reference, but are not directly affected by exposure to fellows. We used those items as decoys or "pretend metrics of impact" in regressions with the same specification described by equation (3). In those cases, the estimates of β are informative of the presence of reference bias.

4 Questionnaires

An essential part of the evaluation was the development of instruments to measure socioemotional skills and related behaviors and attitudes. As part of the questionnaire design process, Teach For All brought in several experts to give us their opinions and suggestions. The design process also included

piloting questionnaires and doing cognitive interviews to determine whether respondents properly understood what was asked. All questionnaires were designed to be optically read, with multiple-option items. We tried to keep time of completion under 20 minutes because of time and quality concerns. Below, we separately describe the questionnaires for students and teachers. The actual questionnaires are included in the Appendix (microanalitica.com/ExM/Appendix), together with the formulae to compute the scales.

4.1 Student questionnaire

The items that comprise the student questionnaire can be split into three groups: controls, metrics of impact, and information about effective teaching. The controls include educational attainment of the mother, number of people living in the same dwelling, and number of books at home. We also included ten items to measure the Big Five personality traits and used them as controls. The Big Five constitute the best-known categorization of personality traits (Goldberg, 1993). They have long been used in psychology to study how personality relates to job performance and satisfaction, academic success, entrepreneurial status and financial attitudes and behaviors. The Big Five measure to what extent any person is conscientious (orderly, responsible, dependable), extraverted (talkative, assertive, energetic), agreeable (good-natured, cooperative, trustful), neurotic (not calm, easily upset), and open to new experiences (intellectual, imaginative, open-minded.) (John and Srivastava, 1999). The ten-item inventory can be applied quickly (Rammstedt and John, 2007) and has been applied to Mexican youths, ages 15 to 29 (Peña, 2016).

As metrics of impact we included six socioemotional scales and six scales for behaviors and attitudes related to those skills. Four of the socioemotional scales are from the California Office of Reform to Education (CORE) Districts Survey. CORE is a consortium of school districts that is incorporating socioemotional skills into school accountability systems. CORE and the organization Transforming Education selected those four scales because they are valid predictors of academic success, and they are likely to be malleable through school-based interventions. Additionally, those scales are assessed in less than 20 minutes (West, 2016). The four CORE scales are: self-management, growth mindset, self-efficacy, and social awareness. Self-management is the ability to regulate one's emotions, thoughts, and behaviors effectively in different situations. This includes managing stress, delaying gratifications, motivating one's self, and setting and working toward personal and academic goals. Growth mindset is the belief that one's abilities can grow with effort. Students with a growth mindset see effort as necessary for success, embrace challenges, learn from criticism, and persist in the face of setbacks. Self-efficacy is the belief in one's own ability to succeed in achieving an outcome or reaching a goal. Self-efficacy reflects confidence in the ability to exert control over one's own motivation, behavior, and environment. Social awareness is the ability to take the perspective of and empathize with others from diverse backgrounds and cultures, to understand social and ethical norms of behavior, and to recognize family, school, and community resources and supports. The use of the four scales in field studies indicates that they are positively related to key indicators of academic performance and behavior (West, 2016).

The remaining two socioemotional scales are grit and academic locus of control. Grit is defined as perseverance and passion for long-term goals and is predictive of academic and professional success

(Duckworth, Peterson, Matthews, and Kelly, 2007). We used the eight-item version of the scale (Duckworth and Quinn, 2009), which has been used in Mexico in large-scale assessments among 9th and 12th graders. We decided not to include grit in the questionnaires for students in grades 4-6 because of the complexity of the items in the scale. Locus of control can be internal or external. A person with an internal locus of control believes that he or she can influence events and their outcomes, while someone with an external locus of control blames outside forces for everything (Rotter, 1966). Academic locus of control translates the same logic to the academic dimension. It is measured with nine-items (Arocha and Lezama, 2007). Since the inclusion of grit and academic locus of control would make the student questionnaire long, we opted for a so-called spiral structure. Half of the respondents were given the grit scale, and the other half was given the academic locus of control scale. Which students were given what scale was quasi-randomly determined by alternating questionnaire versions when they were handed to students.

The other six metrics of impact are behaviors and attitudes related to socioemotional skills. They can be thought of as ways in which changes in socioemotional skills materialize. The first scale in this group is educational expectations. We measure this by eliciting desired and expected educational attainment, as well as the perceived likelihood of obtaining a college degree. The second scale is given by the perceived general returns on education. We measure it as the degree to which a student perceives education as a means to get a better job and attain greater job satisfaction. The third scale is defined as perceived pecuniary returns on education. We measure the scale by eliciting wage expectations at age 40 in money intervals, in two hypothetical scenarios: if the respondent gets a college degree, and if the respondent only gets a high school diploma. The difference between those two expectations (defined in money intervals) is a measure of the perceived pecuniary returns on education. The fourth and fifth scales in this group are two school-related behaviors: self-reported tardiness, absenteeism and class-skipping, and self-reported time devoted to homework. The last scale in this group is the three-item civic engagement scale of Doolittle and Faul (2013), which measures attitudes and behaviors of students regarding their communities, eliciting whether they care and do anything for them.

Teach For All expressed concern about reference bias (see section 3.3). In our chats with experts, it became clear that there is no infallible method to measure—let alone correct for—reference bias. Measuring reference bias usually involves the use of “anchoring vignettes”, which are hypothetical situations to be assessed by the respondent. Since the situations are the same across respondents, their responses are supposed to be informative of the reference used by each respondent (Kyllonen and Bertling, 2013). Given the skepticism expressed by the experts about the practical value of the information gathered through anchoring vignettes, we took a different—new and original—route. To assess whether reference bias could affect the results of this evaluation, we included seven items in the student questionnaire, both at the baseline and the endline. Those items were asked twice in each questionnaire. The first time they referred to the respondent, and the second time they referred to a respondent’s sibling attending Secundaria or Bachillerato. Respondents with no siblings in Secundaria or Bachillerato did not answer these items. We also asked whether the sibling mentioned attends the same school as the respondent. This information allows us to include only the responses about siblings who almost surely have had no contact with Enseña por México fellows. We arrived at the seven reference-bias items using three criteria. First, they had to be a subset of the items included in the scales that are

our metrics of impact. Second, they are good candidates to suffer from reference bias. In other words, we focused on items that have a reference that could, in principle, be modified by being exposed to a fellow. Third, those items should refer to observable behaviors, so that the respondent could objectively assess them for a sibling. The seven items are at the end of the questionnaires.

We included Tripod's 7Cs as measures of effective teaching (Bill & Melinda Gates Foundation, 2012). Tripod's survey was developed by Ronald Ferguson and has been refined over a decade as a research and professional development tool. Tripod's 7Cs are measured through student surveys and they reflect teaching practices. The 7Cs are: care (show concern for students' emotional and academic well-being), confer (encourage and value students' ideas and views), captivate (spark and maintain student interest in learning), clarify (help students understand content and resolve confusion), consolidate (help students integrate and synthesize key ideas), challenge (insist that students persevere and do their best work) and control (foster orderly, respectful, and on-task classroom behavior). Higher scores in Tripod's 7Cs are associated with greater teacher value-added in test scores (Kane, McCaffrey, Miller and Staiger, 2013). Tripod's 7Cs were included only in our endline questionnaire because they require students to assess teacher performance. At the time of the baseline students did not have enough elements for such assessment—the baseline took place near the start of the academic year. In contrast, students can assess teacher performance in the endline because it takes place close to the end of the academic year. We used the upper elementary version of the Tripod's 7Cs (35 items) for grades 4-9, and the secondary version (36 items) for grades 10-12.

4.2 Instructor questionnaire

The evaluation did not require applying questionnaires to instructors. However, the information that those questionnaires include may be helpful in interpreting its results. In principle, instructor questionnaires allowed us analyzing whether there is empirical support for the theory of change. Before describing how we arrived at the instrument applied in the field, it is worth noting that students are already reporting teacher practices and values through the Tripod's 7Cs in the questionnaires described in the previous section. The intention behind giving a questionnaire to instructors is to gain additional insights complementing the information on teacher performance (as assessed by students with the Tripod's 7Cs) with self-reported characteristics, practices and values.

In the context of our analysis, an appropriate instructor questionnaire must satisfy three requirements. First, it must be "palatable," i.e. it cannot ask questions that may be perceived as threatening to instructors or principals. It must be short and non-intrusive. Some teacher questionnaires (e.g. the Comprehensive Assessment for Leadership and Learning) are taken voluntarily and schools request—and pay—to be surveyed. Other questionnaires like those of the OECD Teaching and Learning International Survey (TALIS) are applied by mandate of the educational authorities (OECD, 2014). Our case is entirely different. Schools and individual teachers may not be interested in responding to the survey and could reject it without any consequence. After all, Enseña por México is an external organization requesting the application of questionnaires that offer no benefits for the control schools, or for regular teachers in treatment schools. Access to schools depends entirely on the good will of principals and teachers. The second requirement for the instructor instrument is that it should measure variables belonging to one of three categories: (1) easily-malleable behaviors, (2) non-malleable traits

observable at the recruitment stage, or (3) school-level factors that might be good complements for the presence of fellows. This requirement is for the utility of the answers from the perspective of Enseña por México. The third requirement is that the items in the questionnaire must be validated. Teach For All expressed a strong inclination for relying on items that had been used by other researchers, preferably in diverse contexts. For those two reasons we relied on TALIS. In 2013, TALIS teacher questionnaires were applied to thousands of teachers in 34 countries, including Mexico. In our revision of the TALIS teacher questionnaire we identified instances of "double-barreled" questions, and potential social desirability and reference biases. Since any attempt to address those issues would require changing TALIS' items and thereby lose comparability, we did not try to address them.

The contents of the instructor questionnaire can be divided into five groups. In the first group are demographic characteristics: gender, age, subject taught, and years of experience.

In the second group are personality traits and socioemotional skills. We included the ten-item Big Five inventory and the eight-point grit scale described in the previous section.

In the third group are eight questions that elicit teaching values while trying to avoid desirability bias by using a forced-choice structure. The forced-choice structure has been used when responses could be subject to social desirability bias, because it prevents respondents from endorsing the desirable items and rejecting the undesirable ones (Brown and Maydeu-Olivares, 2018). Forcing the choice does not prevent respondents from trying to choose the most desirable option. But it requires them forming hypotheses about what is more desirable (Meade, 2004). In our view, this is not necessarily a problem. When instructors try to choose the most desirable answer in a force-choice question, we may be extracting information about their teaching values—what they think is more socially desirable. In our eight forced-choice questions, instructors must choose what is more important in a Likert-type scale between two desirable situations that are, to some degree, in tension. For instance, in the first of those eight questions, the instructor must choose whether achieving "disciplined students" in the classroom is more important that "students that participate." Both options are desirable, but one may come at the expense of the other. The other pairs of situations in tension are: "fostering curiosity to learn" vs "developing skills and capacities", "teamwork" vs "independence of ideas", "students who respect the teacher" vs "students who trust the teacher", "students focused on success" vs "students focused on effort", "students who make concrete, achievable plans" vs "students who set very ambitious goals", "students who like learning" vs "students who know that learning is useful", and "developing the potential of the most advanced students" vs "helping students who are lagging behind to catch up with the rest of the class."

The fourth group consists of eleven scales that characterize teacher practices, most of them were taken from TALIS. The scales focus on curricular and extracurricular activities, expectations and perceptions about students' time devoted to the class taught be the instructor, perception of appropriate training and qualifications, promotion of independent thinking, self-perception of effectiveness, proactivity in the classroom, use of evaluation and feedback, and relationship to others in school.

The fifth group consists of two scales from TALIS that measure perceptions on culture of participation in the school, and perceptions on harmony in school.

5 Data

Our analysis uses the results of applying the instruments described in the previous section to over 30,000 students and 800 instructors. Below we describe each step in the process to arrive at our sample of analysis.

5.1 Sample design

Enseña por México does not unilaterally choose in what schools it operates. In any given academic year, the list of participating schools is the result of negotiations with state authorities. For that reason, we were not able to propose an experiment where fellows would be randomly allocated within a pool of schools. Instead, we were given: (a) a list of treatment schools, which we refer to as "EPM schools", and (b) a list of schools where state authorities, in principle, granted access for the evaluation.

5.1.1 Schools with access

Although Enseña por México currently operates in 12 states, access to schools was granted in just four states: Baja California Sur (BCS), Chiapas, Hidalgo and Puebla. Those four states accounted for 69% of the active fellows at the beginning of the 2016-17 academic year. Authorities in those states provided lists of schools of different educational levels with access. Table 1 shows the distribution of schools with access for the evaluation.

Table 1. Schools with access for the evaluation

School level	BCS	Chiapas	Hidalgo	Puebla
Primaria	396	829	240	4,310
Secundaria	189	0	0	2,175
Bachillerato	137	0	0	1,483
Total	722	829	240	7,968

It is important to note that in Chiapas and Hidalgo all schools with access are part of the CONAFE system. In general, CONAFE schools are comparable amongst themselves but they are not comparable to non-CONAFE schools. CONAFE schools are multi-grade and have very small student bodies (under 30). The small number of students is accompanied by poverty and a lack of public infrastructure. Additionally, fellows in CONAFE schools assumed the role of "itinerating pedagogic counselors," tutoring students who needed remedial work while splitting their time between two schools—switching locations every other week.

Leaving CONAFE schools aside, many other schools with access are not comparable to EPM schools. They serve students of a different socio-economic status or have a better academic performance. They also have student bodies of different sizes with different student-teacher ratios.

For the purposes of the evaluation, it is crucial to find non-EPM schools within the pool of schools with access that are comparable to EPM schools. For that reason, we restricted the list of schools in BCS and Puebla to schools with available performance records from PLANEA 2015, a national standardized test given to students in grades 6, 9 and 12. We use PLANEA test scores as a measure of academic

performance in our search for comparable schools. The exclusion of schools without PLANEA test scores disqualified recently created schools as well as schools that refused to take the test.

5.1.2 Coarsened exact matching

We applied the method known as Coarsened Exact Matching to the list of schools with access to find comparable schools. We used Stata's cem command (Blackwell *et al*., 2009). Depending on the state, we used different variables for the matching: the number of students and teachers, the poverty index of the municipality or town where the school is located, school infrastructure, and performance in the National Plan for the Evaluation of Learning (PLANEA) 2015 test.[1]

In general, the Coarsened Exact Matching method finds strata of schools that are similar in terms of the above-mentioned variables. If the matching is successful for the purposes of the evaluation, then in every stratum where there are EPM schools we would also find non-EPM schools. By construction, those non-EPM schools are comparable to the EPM schools in the same stratum. Therefore, they can potentially serve as control schools.

The Coarsened Exact Matching method can be adjusted to provide finer or coarser matches. There is a trade-off between how fine the match is and how many non-EPM schools we find that match the EPM schools. If we want schools to be very similar, the result is fewer schools per stratum. In the extreme, there would not be any match. On the opposite end, if the match is too coarse, we end up with many matched schools that are not very similar. We took a trial-and-error approach to balance the similarity of schools within the same stratum and number of matches. The process was done separately for each of the four states. In some states, we were willing to tolerate coarser matches.

In the case of Baja California Sur, the matching was based on performance in PLANEA 2015 in Language and Communication and in Mathematics. In Chiapas, the matching was based on the number of students, the number of teachers, and the type of floor in the school (concrete or other material). In Hidalgo, the matching was based on the number of students, the number of teachers, and the poverty index of the town. Lastly, the match in Puebla was based on performance in PLANEA 2015 in Language and Communication and in Mathematics, the number of students, the number of teachers, and the poverty index. In all cases we considered the match of schools within the same educational level: Primaria, Secundaria or Bachillerato.

Table 2 shows the results of the matching by state. The first two rows show the total number of strata created and the number of strata where EPM schools were matched to non-EPM schools. For instance,

[1] Information on the number of students and teachers comes from the administrative records provided by state authorities and from publicly available data from the National System of Statistical Information on Education of the Ministry of Education. Poverty indices come from public data bases of the National Council for the Evaluation of Social Development Policy and the National Population Council. School level results for PLANEA 2015 are publicly available from the federal Ministry of Education. Information about school infrastructure comes from public data found in the School Census 2015 of the National Institute of Statistics and Geography.

in the case of Baja California Sur (BCS) there are 40 strata; 14 of them include both EPM and non-EPM schools.

Table 2. Results of the Coarsened Exact Matching

	BCS		Chiapas		Hidalgo		Puebla	
Strata								
Total	40		35		15		1,308	
Matched	14		12		8		36	
Schools	Non-EPM	EPM	Non-EPM	EPM	Non-EPM	EPM	Non-EPM	EPM
Matched	225	26	627	36	187	38	151	39
Unmatched	384	0	87	2	14	1	5,085	9
Total	609	26	714	38	201	39	5,236	48

The lower part of Table 2 displays the number of schools matched and the number of schools unmatched, by status (EPM or non-EPM.) Many non-EPM schools are unmatched. The most notable case is Puebla, where access was granted to 5,236 non-EPM schools. However, 5,085 (97%) of those schools do not resemble EPM schools. We also have unmatched EPM schools, but their number is relatively modest: 12 out of 151 in the four states.

As Table 2 shows, we found 1,190 non-EPM matches for 139 EPM schools. However, due to cost considerations we cannot include all the matched non-EPM schools in the sample of analysis. From among all those pre-selected non-EPM schools (i.e., in matched strata), we proceeded to randomly choose the ones that were part of the sample of analysis.

5.1.3 Random sampling of control schools

We included in the sample two or three non-EPM schools for each matched EPM school—bringing the number of schools in the sample to over 400 schools. The selection of those schools was random and stratified by stratum. In the end, we obtained a sample of analysis wherein non-EPM schools resemble EPM schools in terms of the joint distributions of the variables used for the Coarsened Exact Matching. Table 3 displays the number schools after the random sampling.

Table 3. Number of schools in the sample

Schools	BCS	Chiapas	Hidalgo	Puebla	Total
Non-EPM	71	76	78	94	319
EPM matched	26	36	38	39	139
EPM unmatched	0	2	1	9	12
Total	97	114	117	142	470

The resulting sample of analysis is roughly balanced across states. In the four states we had a ratio of non-EPM schools to matched EPM schools of at least 2.05. Table 3 shows unmatched EPM schools because, even though they will not be part of the analysis, we planned to collect surveys there.

We tested differences between matched EPM and non-EPM schools in the resulting sample—we excluded EPM schools with no match. In the case of continuous variables (e.g., the number of students), we used t-tests. In the case of categorical variables, we used chi-2 tests. Of the 19 tests performed, none

showed significant differences at 95% confidence. In other words, using the variables on which we focused, the matching process resulted in the selection of similar schools.

5.2 Survey collection and attrition

The most important challenge we faced in the field during the survey collection was a combination of communication problems between authorities and schools.[2] As we explain below, the number of questionnaires collected in the endline is smaller than the number of questionnaires collected in the baseline. Half of the drop is explained by "lost" classrooms in schools surveyed in both the baseline and the endline. Since the field staff followed the same protocol for the baseline and the endline, we attribute the loss of classrooms to absences when the endline survey was collected.

Another challenge experienced in a few cases was the lack of control school officials and teachers had over their students. That translated into less than full cooperation with the evaluation. In some isolated incidents, students were openly hostile to the survey collection process—questionnaires were destroyed, and field staff were intimidated. Some students answered the questionnaires carelessly or with offensive messages.

In the baseline, we collected 45,723 questionnaires. By design, we did not identify students in the surveys. Thus, our estimation approach relies on matching baseline and endline data at the classroom level. As we explained in section 3.2, a classroom is defined as a combination of grade and group and it means both a roster of students taking the same courses together and the physical location where they take those courses. Our sample of analysis consists of the classrooms observed in both the baseline and the endline. We have a total of 1,194 such classrooms in 328 schools. In those classrooms, we collected 30,389 questionnaires in the baseline and 26,127 in the endline.

Table 4 shows the sources of attrition. We collected 45,723 questionnaires in the baseline (first row.) Some schools surveyed in the baseline refused to participate in the endline and some were not reached. If we omit those schools from the baseline, we would have collected 39,491 questionnaires in the endline (second row.) Within each school surveyed twice, some classrooms were surveyed in the baseline but not in the endline. If we only count classrooms surveyed twice, the number of questionnaires is 30,389 in the baseline (third row) and 26,127 in the endline (fourth row.)

[2] Some school principals slowed down the collection process a full day, arguing they hadn't been notified. For some principals, a direct, written instruction to fully cooperate with the survey wasn't enough. They requested to speak with their supervisors. In addition, planning was not very effective because, in many instances, schools took days off without informing the authorities.

Table 4. Attrition by educational level

	Primaria	Secundaria	Bachillerato	Total
Students in baseline	27,666	15,766	2,291	45,723
Students in schools in both surveys	23,705	13,894	1,892	39,491
Students in classrooms in both surveys:				
Students in baseline	17,442	11,333	1,614	30,389
Students in endline	14,734	9,881	1,512	26,127
Classrooms	554	443	197	1,194
Schools	105	70	153	328

Another way to explain the information in Table 4 is decomposing attrition starting with all students surveyed in the baseline. Of those 45,723 students, 13.6% were lost because their schools were not surveyed in the endline. Within schools surveyed in the endline, 23.0% of students were lost because their classrooms were not surveyed. Lastly, 14.0% of students in classrooms surveyed were lost because they did not show up the day of the endline survey. The resulting attrition is 42.9%.

Figure 1 presents a graphic version of attrition. For each educational level, there are two graphs. In one graph, each observation corresponds to one school. In the other graph, each observation corresponds to one classroom. The horizontal axis denotes the number of schools or classrooms in the baseline. The vertical axis denotes the number of schools or classrooms in the endline. The 45-degree line indicates the expected relationship in the absence of attrition.

In the school-level graphs, the observations on the horizontal axis represent schools surveyed in the baseline but not in the endline. For instance, there were a couple of Bachillerato schools with over 600 questionnaires in the baseline that refused to participate in the endline. In the classroom graphs, there are observations above and below the 45-degree line. Some classrooms had more observations in the baseline while others had more observations in the endline.

Table 5 presents the results of statistical tests to determine whether attrited students are similar to non-attrited students in terms of the six control variables (socio economic status and the Big Five) and the twelve metrics of impact. The results indicate that attrited students were different to non-attrited students in five control variables for Bachillerato, one for Secundaria, and one for Primaria. In terms of the metrics of impact, there are significant differences in six of them for Bachillerato, in four for Secundaria, and in two for Primaria. In sum, Table 5 indicates that attrition was not random. Among the more relevant results of the comparison are that attrited students in Bachillerato and Secundaria have higher socioeconomic status, greater self-management, lower growth mindset, and higher educational expectations.

Figure 1. Attrition of schools and classrooms by educational level

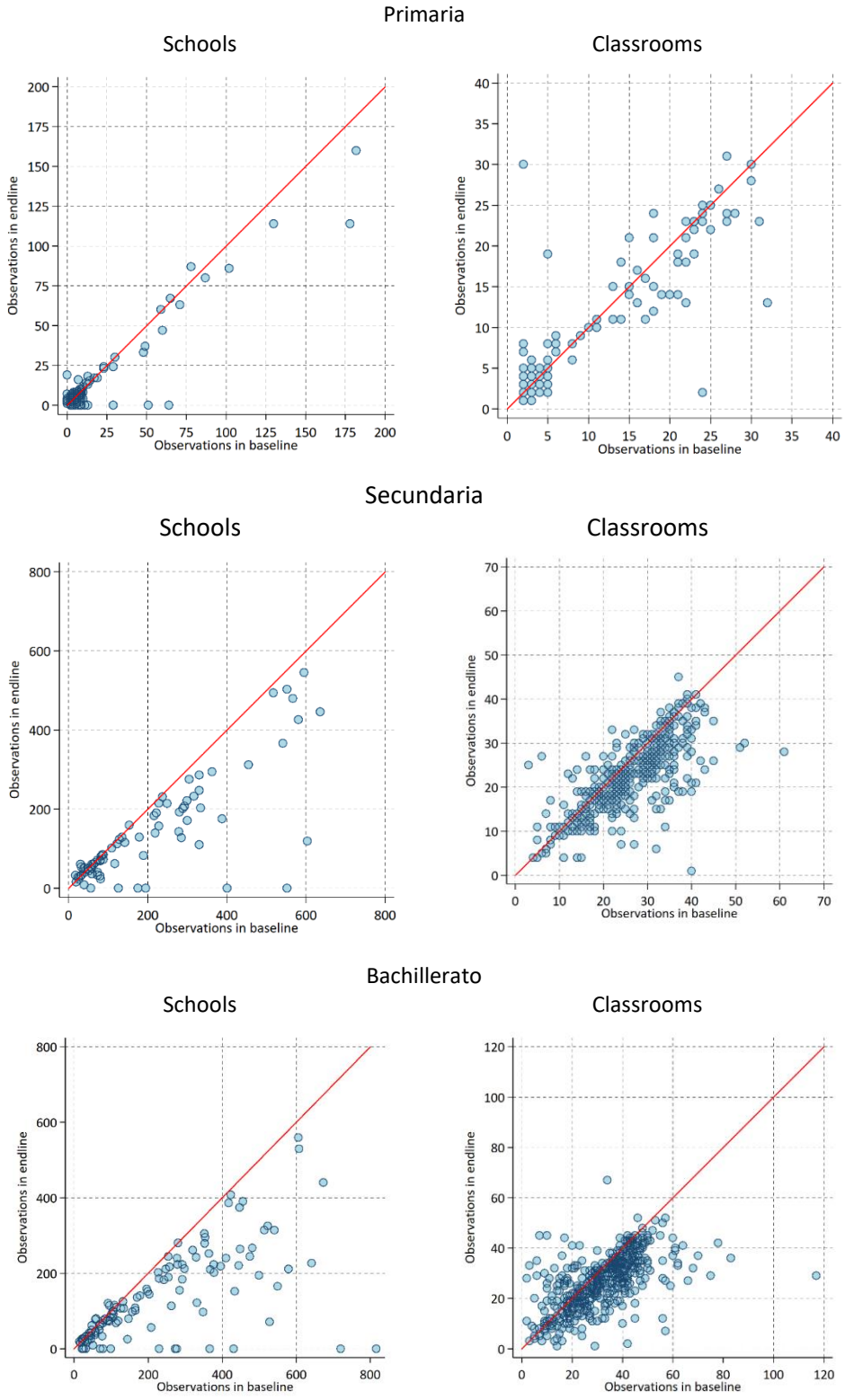

Table 5. Balance between attrited and non-attrited students

Scale	Bachillerato	Secundaria	Primaria
Socio Economic Status	0.198 ***	0.156 ***	-0.073
Big Five: Openness	-0.021	0.018	0.039
Big Five: Conscientiousness	-0.105 ***	0.042	-0.018
Big Five: Extraversion	-0.037 *	0.018	0.120 *
Big Five: Agreeableness	-0.096 ***	0.024	-0.006
Big Five: Neuroticism	-0.036 *	-0.034	0.070
CORE: Self-management	0.065 **	0.080 **	0.046
CORE: Growth mindset	-0.050 *	-0.058 *	0.047
CORE: Self-efficacy	0.050 *	0.027	0.146 *
CORE: Social awareness	0.007	0.049	-0.060
Educational expectations	0.109 ***	0.118 **	-0.081
Returns on education: General	-0.018	0.010	0.018
Returns on education: Pecuniary	-0.042 *	-0.026	
School-related behavior: Tardiness and absenteeism	0.044	-0.069 *	-0.140 *
School-related behavior: Time devoted to homework	0.082 ***	0.040	0.067
Community involvement	-0.008	0.038	
Locus of control	-0.002	0.041	-0.027
Grit	0.007	0.026	

*p < 0.05, **p < 0.01, and ***p < 0.001. Robust standard errors clustered by classroom.

5.3 Treatment fidelity

In the 2016-17 academic year, fellow retention in the program was 74%. Most of the fellows who dropped out could not adapt to the living conditions around CONAFE schools. As a point of reference, retention in Enseña Chile is 79% (Alfonso, Santiago and Bassi, 2010). Since Enseña por México had commitments to reach a given number of schools, some of the schools that were originally in the control group ended up being treated. Table 6 shows the treatment fidelity by assigned and actual treatment status. Actual and assigned statuses are the same for 298 of the 328 schools in the sample of analysis. For 30 schools, actual and assigned statuses differ. As we mentioned in section 3.2, to deal with imperfect treatment fidelity we use a Two-Stage Least Squares strategy to obtain alternative estimates.

Table 6. Treatment fidelity

Educational level	Assigned status			
	Control		Treatment	
	Actual status			
	Control	Treatment	Control	Treatment
Primaria	100	1	18	34
Secundaria	49	2	1	18
Bachillerato	62	3	5	35
Total	211	6	24	87

We use actual treatment to define the status of schools as control or treatment ("EPM school"). Table 7 presents the composition of the sample of analysis, that is, students in classrooms that were successfully visited in the baseline and the endline. The top panel shows the number of schools. There are 328 schools, 92 of which are EPM schools. The middle panel shows the number of classrooms. CONAFE schools are counted as a single classroom. There are 1,194 classrooms in the sample, 754 are in control

schools and 435 are in EM schools. Of the 440 classrooms in EPMs schools, 288 had a fellow and 152 did not. Lastly, the bottom two panels show the number of students in the baseline and the endline.

Table 7. Sample of analysis

	Bachillerato	Secundaria	Primaria	Total
Schools				
Control	68	51	117	236
EPM schools	37	19	36	92
Total	105	70	153	328
Classrooms				
Control schools	322	286	146	754
EPM schools, non-treated	68	78	6	152
EPM schools, treated	164	79	45	288
Total	554	443	197	1,194
Students in baseline				
Control schools	9,582	7,197	1,173	17,952
EPM schools, non-treated	2,331	2,240	75	4,646
EPM schools, treated	5,529	1,896	366	7,791
Total	17,442	11,333	1,614	30,389
Students in endline				
Control schools	8,130	6,319	1,113	15,562
EPM schools, non-treated	1,897	1,911	76	3,884
EPM schools, treated	4,707	1,651	323	6,681
Total	14,734	9,881	1,512	26,127

5.4 Balance

As we mentioned in section 5.2, attrition was a challenge. To explore whether attrition modified the balance between EPM schools and control schools, we compare school-level characteristics. Table 8 shows the results of tests for differences in the same school characteristics used for Coarsened Exact Matching described in section 5.1. The tests were implemented through a regression that includes fixed effects for each combination of state and school level.

Table 8. Differences in school-level characteristics (EPM schools – control schools)

Variable	Difference	p-value	Schools
Number of students	**51.97**	(0.004)	328
Student-to-teacher ratio	-0.74	(0.355)	327
Percent of students who are female	0.34	(0.804)	326
Percent of teachers who are female	0.65	(0.885)	327
Computers per student	-0.01	(0.493)	326
School has test data (PLANEA 2015)	-0.02	(0.553)	328
Percent who scored "insufficient" in Language	-1.34	(0.659)	221
Percent who scored "insufficient" in Math	-2.77	(0.437)	220

Estimates in bold type and shaded are significant at 95% confidence.

There is a significant difference in the average number of students in EPM schools, which are larger. That is explained by Secundarias and Bachilleratos. On average, fellows at those levels are sent to larger schools. In the rest of the variables, EPM and controls schools do not appear to be significantly different.

Table 9 presents the results of t-tests using student-level data for the controls and the metrics of impact in the sample of analysis in the baseline. The first set of columns shows the results including all schools in the sample (treated and non-treated), and the second set of columns shows the results including only EPM schools, that is, schools where there was at least one EPM fellow. In those cases, the comparison was made between treated and non-treated students. Within each set of columns, we present results for the three educational levels together and separately.

We find significant differences in socioeconomic status, three of the Big Five (extraversion, agreeableness and neuroticism), one of the CORE scales (social awareness), pecuniary returns on education, tardiness and absenteeism, time devoted to homework, and internal locus of control. In sum, treated and non-treated students do not look entirely similar in the baseline. However, no systematic pattern of differences is apparent.

Table 9. Baseline differences between treated and non-treated students

	All schools				EPM schools only			
Statistic	All levels	Bachillerato	Secundaria	Primaria	All levels	Bachillerato	Secundaria	Primaria
Socioeconomic status								
Estimate	0.046	0.056	0.025	0.048	**0.122**	**0.189**	0.010	-0.012
p-value	(0.193)	(0.205)	(0.700)	(0.648)	(0.042)	(0.018)	(0.922)	(0.954)
Observations	29,284	16,991	10,780	1,513	8,709	5,387	3,194	128
Classrooms	1,188	551	443	194	283	153	119	11
Openness								
Estimate	0.026	0.019	0.030	0.114	0.012	0.017	-0.009	0.416
p-value	(0.142)	(0.391)	(0.362)	(0.163)	(0.682)	(0.665)	(0.855)	(0.155)
Observations	29,794	17,118	11,143	1,533	8,885	5,431	3,324	130
Classrooms	1,191	552	443	196	283	153	119	11
Conscientiousness								
Estimate	0.017	0.026	-0.006	0.027	0.010	-0.006	0.036	0.104
p-value	(0.394)	(0.298)	(0.877)	(0.725)	(0.739)	(0.886)	(0.489)	(0.785)
Observations	29,364	16,993	10,866	1,505	8,769	5,405	3,239	125
Classrooms	1,190	552	443	195	283	153	119	11
Extraversion								
Estimate	**0.039**	**0.047**	0.017	0.075	0.040	**0.071**	0.012	-0.013
p-value	(0.026)	(0.036)	(0.597)	(0.372)	(0.130)	(0.046)	(0.789)	(0.955)
Observations	29,688	17,028	11,109	1,551	8,859	5,414	3,318	127
Classrooms	1,191	552	443	196	283	153	119	11
Agreeableness								
Estimate	**0.043**	0.035	0.068	0.034	0.043	0.053	0.072	-0.003
p-value	(0.026)	(0.137)	(0.080)	(0.698)	(0.179)	(0.183)	(0.221)	(0.990)
Observations	29,776	17,162	11,086	1,528	8,884	5,449	3,307	128
Classrooms	1,191	552	443	196	283	153	119	11
Neuroticism								
Estimate	**0.036**	**0.049**	0.019	-0.021	**0.065**	**0.083**	0.034	0.221
p-value	(0.032)	(0.015)	(0.582)	(0.776)	(0.023)	(0.010)	(0.550)	(0.400)
Observations	29,816	17,194	11,074	1,548	8,896	5,455	3,310	131
Classrooms	1,189	552	443	194	283	153	119	11
CORE: Self-management								
Estimate	0.012	0.030	-0.047	0.100	0.022	0.024	-0.021	0.309
p-value	(0.589)	(0.292)	(0.239)	(0.212)	(0.582)	(0.673)	(0.725)	(0.218)
Observations	28,359	16,624	10,344	1,391	8,457	5,252	3,086	119
Classrooms	1,190	552	443	195	283	153	119	11
CORE: Growth mindset								
Estimate	-0.014	-0.011	0.002	-0.145	-0.014	0.020	-0.058	-0.413
p-value	(0.546)	(0.698)	(0.968)	(0.159)	(0.696)	(0.665)	(0.345)	(0.251)
Observations	29,102	16,767	10,823	1,512	8,640	5,302	3,213	125
Classrooms	1,191	552	443	196	283	153	119	11
CORE: Self-efficacy								
Estimate	-0.002	-0.010	-0.012	0.167	-0.036	-0.022	-0.101	0.299
p-value	(0.932)	(0.730)	(0.748)	(0.051)	(0.317)	(0.652)	(0.082)	(0.190)
Observations	29,156	16,809	10,845	1,502	8,696	5,338	3,233	125
Classrooms	1,191	552	443	196	283	153	119	11

Estimates in bold type and shaded are significant at 95% confidence. p-values in parentheses.

Table 9. Baseline differences between treated and non-treated students—continued …

	All schools				EPM schools only			
Statistic	All levels	Bachillerato	Secundaria	Primaria	All levels	Bachillerato	Secundaria	Primaria
CORE: Social awareness								
Estimate	**0.061**	**0.086**	-0.009	0.137	0.069	**0.139**	-0.053	0.618
p-value	(0.008)	(0.002)	(0.825)	(0.244)	(0.074)	(0.011)	(0.340)	(0.135)
Observations	29,158	16,861	10,827	1,470	8,684	5,335	3,234	115
Classrooms	1,188	551	443	194	283	153	119	11
Educational expectations								
Estimate	-0.015	-0.042	0.048	-0.024	0.079	0.103	0.020	0.276
p-value	(0.639)	(0.332)	(0.350)	(0.827)	(0.097)	(0.105)	(0.807)	(0.468)
Observations	29,441	16,907	11,000	1,534	8,768	5,368	3,272	128
Classrooms	1,190	551	443	196	283	153	119	11
Returns on education: General								
Estimate	-0.011	-0.018	0.001	0.003	0.031	0.066	-0.004	0.453
p-value	(0.521)	(0.432)	(0.961)	(0.974)	(0.300)	(0.137)	(0.912)	(0.207)
Observations	29,641	17,212	10,876	1,553	8,841	5,460	3,254	127
Classrooms	1,189	551	443	195	283	153	119	11
Returns on education: Pecuniary								
Estimate	-0.026	-0.032	-0.014		-0.011	**-0.046**	0.044	
p-value	(0.201)	(0.213)	(0.679)		(0.732)	(0.283)	(0.352)	
Observations	27,730	16,834	10,896		8,599	5,338	3,261	
Classrooms	994	551	443		272	153	119	
School-related behavior: Tardiness and absenteeism								
Estimate	0.029	0.012	0.044	**0.192**	0.014	0.085	-0.103	0.406
p-value	(0.235)	(0.675)	(0.414)	(0.042)	(0.749)	(0.110)	(0.217)	(0.448)
Observations	29,697	17,114	11,032	1,551	8,870	5,431	3,311	128
Classrooms	1,190	551	443	196	283	153	119	11
School-related behavior: Time devoted to homework								
Estimate	**-0.043**	-0.049	-0.052	0.061	0.028	0.034	-0.004	0.348
p-value	(0.039)	(0.065)	(0.175)	(0.527)	(0.424)	(0.477)	(0.948)	(0.531)
Observations	30,010	17,219	11,226	1,565	8,942	5,458	3,357	127
Classrooms	1,190	551	443	196	283	153	119	11
Community involvement								
Estimate	0.000	0.010	-0.025		0.053	0.067	0.034	
p-value	(0.986)	(0.676)	(0.463)		(0.089)	(0.124)	(0.483)	
Observations	28,248	17,159	11,089		8,785	5,463	3,322	
Classrooms	994	551	443		272	153	119	
Grit								
Estimate	-0.032	-0.039	-0.017		-0.057	-0.027	-0.061	
p-value	(0.194)	(0.191)	(0.685)		(0.165)	(0.599)	(0.367)	
Observations	12,910	7,990	4,920		4,009	2,535	1,474	
Classrooms	982	544	438		271	153	118	
Internal locus of control								
Estimate	**-0.054**	-0.035	**-0.101**	-0.121	-0.060	-0.038	**-0.137**	-0.354
p-value	(0.046)	(0.312)	(0.024)	(0.295)	(0.201)	(0.552)	(0.045)	(0.277)
Observations	13,339	8,188	4,475	676	4,025	2,614	1,354	57
Classrooms	1,173	548	438	187	281	153	117	11

Estimates in bold type and shaded are significant at 95% confidence. p-values in parentheses.

5.5 Heterogeneity in intensity

Leaving Primaria aside, treated students do not have the same degree of exposure to fellows, even within the same school. There is variation in the number of fellows teaching different classrooms, and even in the number of courses they teach. We define two measures of treatment intensity. The first measure is the number of different fellows teaching to a classroom the full academic year. Since Bachillerato is organized in semesters, it is possible for a fellow to teach in the Fall semester but not in the Spring semester, or vice versa. Hence, the number of fellows teaching a classroom in Bachillerato is defined in multiples of 0.5. The second measure of treatment intensity is the number of courses taught by fellows to a classroom in an academic year—regardless of the number of fellows. The semester organization means that the number of courses taught by fellows to one classroom is also defined in multiples of 0.5. Based on the administrative records of Enseña por México, we computed these two metrics of treatment intensity for the classrooms in our sample of analysis. Table 10 shows descriptive statistics for the two measures of intensity, by the status of the classroom and its educational level. Classrooms in the external control belong to control schools. Classrooms in the internal control belong to EPM schools but were not treated.

Table 10. Heterogeneity in treatment intensity

	Bachillerato			Secundaria			Both levels		
	External control	Internal control	Treatment	External control	Internal control	Treatment	External control	Internal control	Treatment
Measure of intensity: Number of fellows who taught the classroom									
Distribution of classrooms									
0.0	320	68		286	78		606	146	
0.5			70						70
1.0			71			67			138
1.5			16						16
2.0			7			12			19
Total	320	68	164	286	78	79	606	146	243
Average	0.00	0.00	0.88	0.00	0.00	1.15	0.00	0.00	0.97
Std. dev.	(0.00)	(0.00)	(0.40)	(0.00)	(0.00)	(0.36)	(0.00)	(0.00)	(0.41)
Measure of intensity: Number of subjects taught by fellows									
Distribution of classrooms									
0.0	320	68		286	78		606	146	
0.5			67						67
1.0			45			60			105
1.5			32						32
2.0			15			11			26
2.5			4						4
3.0						7			7
3.5			1						1
10.0						1			1
Total	320	68	164	286	78	79	606	146	243
Average	0.00	0.00	1.04	0.00	0.00	1.43	0.00	0.00	1.16
Std. dev.	(0.00)	(0.00)	(0.58)	(0.00)	(0.00)	(1.16)	(0.00)	(0.00)	(0.83)

We use the heterogeneity in treatment intensity to obtain an additional set of estimates. Although the intensity within a school may not be random, it is not clear how it could be systematically related to socioemotional learning. Fellows may be teaching subjects where there are staff shortages. Some classrooms may have more intensive coursework in the subjects taught by fellows.

5.6 Reliability of the scales

The metrics of impact are scales based on self-reported assessments. Ten of the twelve scales are built based on two or more items. A potential concern is their internal consistency or reliability. Table 11 presents Cronbach's alpha for those ten scales in the baseline data. Not surprisingly, reliability is higher for older students—except for social awareness, for which it is basically the same across educational levels. That fact may reflect a better understanding of the questions. The alphas are relative low for conventional standards. In the case of the CORE scales, growth mindset shows the lowest reliability, which is consistent with evidence from the CORE districts (West, 2016).

Table 11. Reliability of the metrics of impact in the baseline: Cronbach's alpha

Scale	Number of items	Educational level		
		Bachillerato	Secundaria	Primaria
CORE: Self-management	9	0.810	0.752	0.662
CORE: Growth mindset	4	0.644	0.579	0.525
CORE: Self-efficacy	4	0.869	0.847	0.741
CORE: Social awareness	8	0.717	0.727	0.711
Educational expectations	3	0.795	0.784	0.726
Returns on education: General	2	0.638	0.583	0.549
School-related behavior: Tardiness and absenteeism	3	0.617	0.581	0.548
Community involvement	3	0.678	0.643	
Grit	8	0.669	0.630	
Locus of control	9	0.641	0.635	0.520

5.7 Socioemotional skills and the attitudes and behaviors of students

Six of the metrics of impact we analyze are not socioemotional scales. They are attitudes and behaviors that to some extent may be affected by changes in socioemotional skills: educational expectations, perceived general and pecuniary returns on education, tardiness and absenteeism, time devoted to homework, and community involvement. To empirically show that those attitudes and behaviors are associated with the six socioemotional scales used as metrics of impact (self-management, growth mindset, self-efficacy, social awareness, grit and internal locus of control), we compute correlation coefficients using the baseline data. Table 12 shows Spearman correlation coefficients and their p-values, by educational level.[3]

[3] Spearman correlation coefficients are invariant to monotone transformation of the variables.

Table 12. Spearman correlation coefficients between students' socioemotional skills and their attitudes and behaviors, baseline data

	Attitudes and behaviors					
		Returns on education		School-related behavior		
Educational level and socioemotional skill	Educational expectations	General	Pecuniary	Tardiness and absenteeism	Time devoted to homework	Community involvement
Primaria						
CORE: Self-management	**0.204** (0.000)	**0.172** (0.000)		**-0.131** (0.000)	**0.066** (0.015)	
CORE: Growth mindset	**0.055** (0.035)	0.019 (0.464)		**-0.145** (0.000)	**-0.082** (0.002)	
CORE: Self-efficacy	**0.210** (0.000)	**0.193** (0.000)		0.011 (0.663)	**0.106** (0.000)	
CORE: Social awareness	**0.315** (0.000)	**0.317** (0.000)		-0.005 (0.851)	**0.193** (0.000)	
Secundaria						
CORE: Self-management	**0.320** (0.000)	**0.145** (0.000)	**0.119** (0.000)	**-0.278** (0.000)	**0.277** (0.000)	**0.305** (0.000)
CORE: Growth mindset	**0.173** (0.000)	0.005 (0.623)	**0.079** (0.000)	-0.017 (0.079)	**0.064** (0.000)	0.018 (0.061)
CORE: Self-efficacy	**0.406** (0.000)	**0.185** (0.000)	**0.142** (0.000)	**-0.121** (0.000)	**0.250** (0.000)	**0.285** (0.000)
CORE: Social awareness	**0.348** (0.000)	**0.203** (0.000)	**0.121** (0.000)	**-0.130** (0.000)	**0.257** (0.000)	**0.401** (0.000)
Grit	**0.273** (0.000)	**0.069** (0.000)	**0.116** (0.000)	**-0.182** (0.000)	**0.172** (0.000)	**0.145** (0.000)
Internal locus of control	**0.309** (0.000)	**0.134** (0.000)	**0.210** (0.000)	**-0.096** (0.000)	**0.182** (0.000)	**0.167** (0.000)
Bachillerato						
CORE: Self-management	**0.280** (0.000)	**0.101** (0.000)	**0.087** (0.000)	**-0.294** (0.000)	**0.333** (0.000)	**0.257** (0.000)
CORE: Growth mindset	**0.240** (0.000)	0.010 (0.196)	**0.096** (0.000)	**-0.019** (0.014)	**0.102** (0.000)	**0.076** (0.000)
CORE: Self-efficacy	**0.418** (0.000)	**0.109** (0.000)	**0.113** (0.000)	**-0.081** (0.000)	**0.238** (0.000)	**0.222** (0.000)
CORE: Social awareness	**0.281** (0.000)	**0.132** (0.000)	**0.087** (0.000)	**-0.148** (0.000)	**0.235** (0.000)	**0.299** (0.000)
Grit	**0.299** (0.000)	**0.083** (0.000)	**0.096** (0.000)	**-0.194** (0.000)	**0.194** (0.000)	**0.157** (0.000)
Internal locus of control	**0.253** (0.000)	**0.073** (0.000)	**0.154** (0.000)	**-0.089** (0.000)	**0.174** (0.000)	**0.205** (0.000)

Estimates in bold type and shaded are significant at 95% confidence. p-values in parentheses.

As expected, the attitudes and behaviors scales are strongly associated with the six socioemotional scales we use. Of 88 coefficients presented, only seven are not significant; one is significant but has the sign opposite to what we would expect. Students who scored higher in the socioemotional scales in the baseline, also have higher educational expectations, perceive greater returns on education, show less tardiness and absenteeism, devote more time to work, and are more involved with their communities.

The attitudes and behaviors scales are probably more responsive in the short run than socioemotional scales. That possibility makes them appealing for the purposes of measuring changes in one academic year.

6 Results

6.1 Empirical support to the theory of change

As we mentioned in section 3.1, the theory of change is supported by the data if we find differences between fellows and regular teachers that favor the former. The comparison is only relevant for Secundaria and Bachillerato, not for Primaria. As explained in section 2, fellows in Primaria in Chiapas and Hidalgo did not take on the role of teachers. In Baja California Sur and Puebla, only a handful of fellows worked at that level.

Table 13 shows estimates of the differences between fellows and regular teachers in 30 self-reported metrics. Since the number of fellows in the sample is modest, we analyze together Secundaria and Bachillerato. Except for gender, age and experience, all metrics were standardized using the mean and

the average for regular teachers. Although we have information for 833 instructors (fellows and teachers), the sample of analysis includes only 27 fellows.[4]

Table 13. Differences between fellows and teachers in self-reported metrics

Metric	Difference	p-value	Observations	EPM fellows
Female	0.066	(0.494)	833	27
Age in years	**-13.1**	(0.000)	833	27
Experience in years	**-11.4**	(0.000)	829	27
Big Five: Openness	**0.568**	(0.006)	778	25
Big Five: Conscientiousness	0.364	(0.067)	795	26
Big Five: Extraversion	**0.539**	(0.007)	790	26
Big Five: Agreeableness	**0.496**	(0.013)	797	26
Big Five: Neuroticism	-0.125	(0.530)	789	26
Grit	0.226	(0.274)	751	24
Forced choice: Participaction over discipline	-0.266	(0.175)	809	27
Forced choice: Competencies over curiosity	**-0.591**	(0.003)	809	27
Forced choice: Independence over teamwork	0.125	(0.523)	806	27
Forced choice: Trust over respect	0.172	(0.378)	801	27
Forced choice: Effort over success	0.280	(0.150)	806	27
Forced choice: Ambitious goals over concrete plans	**0.670**	(0.001)	806	27
Forced choice: Useful over liking learning	0.023	(0.906)	809	27
Forced choice: Attention to laggards over advanced	-0.248	(0.202)	807	27
Curricular activities	0.065	(0.745)	758	26
Expectations about students' dedication	-0.317	(0.103)	799	27
Perceptions about students' dedication	-0.100	(0.606)	806	27
Extracurricular activities	**0.483**	(0.014)	799	27
Perception of appropriate training	**-0.503**	(0.011)	801	27
Perception of qualifications	-0.057	(0.776)	804	26
Promotion of independent thinking	0.251	(0.207)	788	27
Self-perception of effectiveness	-0.186	(0.372)	723	24
Proactivity in classroom	0.188	(0.354)	744	25
Use of evaluation and feedback	**0.594**	(0.003)	747	27
Relationship to others in school	0.120	(0.537)	740	27
Perceptions on culture of participation in school	**-0.625**	(0.002)	792	27
Perceptions on harmony in school	**-0.728**	(0.000)	792	27

Estimates in bold type and shaded are significant at 95% confidence.

We found significant differences in twelve of the 30 self-reported metrics. Not surprisingly, fellows are over ten years younger, and also over 10 years less experienced. In terms of personality traits, fellows are more open, extraverted and agreeable. In terms of teaching values elicited though forced-choice questions, fellows put the development of curiosity over the development of competencies, and they prefer students having very ambitious goals over concrete and achievable plans. We found differences

[4] According to Enseña por México, there were 97 fellows in the sample of analysis in Secundarias and Bachilleratos, where they teach only 13 hours per week on average, out of a maximum of 30 hours. Thus, we can assume that the probability of finding a particular fellow teaching at the time of the survey was 0.43 (= 13 ÷ 30), which, when multiplied by 90 fellows, equals 42. Ultimately, the number of fellows surveyed was 65% (= 27 ÷ 42) of the expected number.

in extracurricular activities: fellows participate more in those activities. They also see themselves as not having appropriate training for the courses they teach. They rely more on evaluations and give more feedback. Lastly, they have a more negative perception of the school where they work: with a less participative culture and less harmony.

The analysis above has limitations. First, although we can *a priori* hypothesize that showing higher scores in some variables is desirable (e.g., giving feedback to students about their performance), there are other variables where desirability is not clear *a priori* (e.g., emphasizing very ambitious goals over concrete and achievable plans). Second, the variables in Table 13 are self-reported. The answers given by respondents may not reflect actual behaviors. Third, we do not know which of the variables matter more for more effective teaching.

As an attempt to address those limitations, we can complement the analysis with what students think of fellows and teachers by the end of the academic year. As described in section 4.1, we collected information on the Tripod's 7Cs, which are supposed to be informative of effective teaching. Table 14 shows estimates of the differences between fellows and regular teachers in the Tripod's 7Cs. The scores were standardized using the mean and the standard deviation for regular teachers.

It should be clear that there is a small sample issue. We observe Tripod's 7Cs scores for less than 30 fellows. Despite that, we do find significant differences in all Tripod's 7Cs at 95% confidence. If we perform one-side tests, the differences for all Tripod's 7Cs are significant using Bachillerato and Secundaria together, both with all schools and EPM schools only. It is worth emphasizing that the estimates in Table 14 rely on metrics reported by students. On average, students perceive EPM fellows as more effective than regular teachers (as measured by the Tripod's 7Cs) by around a fifth of a standard deviation.

Table 14. Differences between fellows and teachers in the Tripod's 7Cs

	All schools			EPM schools only		
	All	Bachillerato	Secundaria	All	Bachillerato	Secundaria
Care						
Estimate	**0.274**	**0.309**	0.151	**0.236**	**0.281**	0.097
p-value	(0.000)	(0.000)	(0.129)	(0.001)	(0.001)	(0.382)
Students	18,370	11,517	6,853	4,302	2,723	1,579
Instructors	786	451	335	178	110	68
Fellows' students	584	455	129	584	455	129
Fellows	24	17	7	24	17	7
Control						
Estimate	0.151	0.129	**0.236**	0.141	0.136	0.198
p-value	(0.052)	(0.172)	(0.012)	(0.092)	(0.185)	(0.062)
Students	17,636	11,155	6,481	4,134	2,643	1,491
Instructors	786	451	335	178	110	68
Fellows' students	565	449	116	565	449	116
Fellows	24	17	7	24	17	7
Clarify						
Estimate	**0.242**	**0.248**	0.222	**0.211**	**0.241**	0.154
p-value	(0.005)	(0.014)	(0.174)	(0.023)	(0.029)	(0.371)
Students	18,276	11,353	6,923	4,260	2,679	1,581
Instructors	786	451	335	178	110	68
Fellows' students	576	453	123	576	453	123
Fellows	24	17	7	24	17	7
Challenge						
Estimate	**0.210**	**0.244**	0.089	0.175	**0.217**	0.044
p-value	(0.019)	(0.011)	(0.668)	(0.065)	(0.039)	(0.838)
Students	17,805	11,216	6,589	4,197	2,664	1,533
Instructors	786	451	335	178	110	68
Fellows' students	572	445	127	572	445	127
Fellows	24	17	7	24	17	7
Captivate						
Estimate	**0.219**	0.223	0.204	0.183	0.199	0.150
p-value	(0.032)	(0.078)	(0.106)	(0.090)	(0.140)	(0.275)
Students	17,604	10,808	6,796	4,110	2,551	1,559
Instructors	786	451	335	178	110	68
Fellows' students	551	428	123	551	428	123
Fellows	24	17	7	24	17	7
Confer						
Estimate	0.210	**0.260**	0.029	0.196	**0.264**	-0.012
p-value	(0.056)	(0.030)	(0.903)	(0.090)	(0.043)	(0.961)
Students	17,660	11,104	6,556	4,148	2,649	1,499
Instructors	786	451	335	178	110	68
Fellows' students	567	443	124	567	443	124
Fellows	24	17	7	24	17	7
Consolidate						
Estimate	**0.202**	**0.244**	0.053	**0.173**	**0.230**	0.001
p-value	(0.006)	(0.002)	(0.755)	(0.029)	(0.009)	(0.993)
Students	18,437	11,462	6,975	4,309	2,711	1,598
Instructors	786	451	335	178	110	68
Fellows' students	579	452	127	579	452	127
Fellows	24	17	7	24	17	7

Estimates in bold type and shaded are significant at 95% confidence. p-values in parentheses.

We can use the information from regular teachers to explore two questions. First, how much of the difference in Tripod's 7Cs between EPM fellows and regular teachers can be attributed to observable variables, and how much is due to factors not observed in the data. Second, we can try to determine which observed variables contribute more to the difference in the Tripod's 7Cs between fellows and teachers.

Using the methodology described in section 3.1, we compute a predicted difference between fellows and regular teachers for each of the Tripod's 7Cs, and compare it with the observed difference. We can say that the gap between the observed and predicted differences is attributable to unobservable variables. We also compute the contribution of each observed variable to the predicted difference.

Table 15 shows the results of the analysis. The top panel shows observed and predicted differences, as well as the part of the difference attributable to unobservable traits of fellows. The bottom panel shows the contribution of each variable to the predicted difference in the average score of each of the Tripod's 7Cs. The highlighted rows indicate the variables that contribute the most to the predicted difference across the Tripod's 7Cs.

Table 15. Tripod's 7Cs and self-reported metrics among regular teachers

Metric	Difference	Care	Control	Clarify	Challenge	Captivate	Confer	Consolidate
Difference in seven Cs								
Observed difference (EPM fellows - regular teachers)		0.230	0.109	0.164	0.166	0.171	0.157	0.138
Predicted difference based on observables (EPM fellows - regular teachers)		0.099	0.140	0.144	0.134	0.119	0.181	0.141
Difference attributable to unobservables of EPM fellows		0.131	-0.031	0.020	0.032	0.053	-0.024	-0.003
Contribution to the predicted difference								
Female	0.112	0.010	-0.008	0.008	0.008	0.005	0.011	0.010
Age in years	-13.39	-0.032	0.018	-0.012	-0.036	-0.017	-0.025	-0.019
Experience in years	-11.45	0.049	0.073	0.041	0.056	0.019	0.053	0.042
Big Five: Openness	0.559	0.015	0.020	0.031	0.028	0.027	0.030	0.023
Big Five: Conscientiousness	0.343	0.006	0.014	0.010	0.007	0.008	0.001	0.003
Big Five: Extraversion	0.451	0.007	-0.007	0.013	0.005	0.013	0.006	0.007
Big Five: Agreeableness	0.524	-0.013	0.007	-0.010	-0.014	-0.007	-0.015	-0.017
Big Five: Neuroticism	-0.207	-0.008	-0.010	-0.003	-0.006	-0.003	-0.002	-0.003
Grit	0.240	0.002	0.002	-0.001	0.002	-0.002	0.002	0.004
Forced choice: Participaction over discipline	-0.303	0.008	0.014	0.020	0.015	0.022	0.017	0.013
Forced choice: Competencies over curiosity	-0.573	0.025	0.009	0.031	0.028	0.041	0.036	0.027
Forced choice: Independence over teamwork	0.155	0.000	-0.003	-0.002	-0.003	-0.003	-0.001	-0.001
Forced choice: Trust over respect	0.189	0.006	0.016	0.009	0.009	0.011	0.009	0.007
Forced choice: Effort over success	0.284	-0.003	0.002	-0.010	-0.002	-0.005	-0.004	-0.003
Forced choice: Ambitious goals over concrete plans	0.725	0.004	0.000	-0.006	-0.009	0.002	0.003	-0.003
Forced choice: Useful over liking learning	-0.044	0.000	0.000	-0.001	-0.001	-0.001	-0.001	0.000
Forced choice: Attention to laggards over advanced	-0.219	0.009	0.006	0.009	0.008	0.006	0.002	0.005
Curricular activities	0.082	0.004	0.001	0.006	0.006	0.005	0.005	0.005
Expectations about students' dedication	-0.335	-0.018	-0.020	-0.023	-0.022	-0.018	-0.013	-0.016
Perceptions about students' dedication	-0.099	0.003	0.000	0.005	0.003	0.005	0.001	0.001
Extracurricular activities	0.448	-0.005	-0.043	-0.020	-0.011	-0.030	-0.020	-0.015
Perception of appropriate training	-0.423	-0.004	0.008	0.003	0.006	0.003	0.007	0.004
Perception of qualifications	-0.030	0.001	0.000	0.002	0.001	0.001	0.001	0.002
Promotion of independent thinking	0.306	-0.002	-0.001	0.000	-0.002	-0.008	-0.002	0.001
Self-perception of effectiveness	-0.187	0.001	-0.003	0.000	0.002	-0.001	0.001	0.002
Proactivity in classroom	0.233	-0.002	0.006	0.001	0.001	0.003	0.003	0.001
Use of evaluation and feedback	0.587	0.016	0.012	0.010	0.017	0.020	0.026	0.016
Relationship to others in school	0.204	0.003	0.005	0.002	0.001	0.002	0.002	0.003
Perceptions on culture of participation in school	-0.658	-0.024	0.010	-0.015	-0.026	-0.019	-0.021	-0.007
Perceptions on harmony in school	-0.721	-0.009	0.002	-0.012	0.000	-0.017	0.013	-0.005

Estimates in bold type and shaded are significant at 95% confidence.

In sum, the evidence supports Enseña por México's theory of change. Its fellows are different from regular teachers and they are perceived by students as more effective. The difference in effectiveness appears to be driven primarily by their lack of experience, their openness to new experiences, and their emphasis on developing students' curiosity instead of competencies and skills. In the case of "care", there seems to be an important unobservable effect for fellows.

6.2 Impact estimates

We first present the estimates defining a classroom as treated if it was taught by an Enseña por México fellow. We also show alternative results summarizing the twelve metrics of impact in one single metric, defined as the first factor using factor analysis. Then we show results using two different definitions of treatment that take into account treatment heterogeneity: (a) the number of fellows who taught the same classroom, and (b) the number of subjects taught to each classroom by fellows. Lastly, we present Two-Stage Least Squares estimates instrumenting actual treatment status with assigned treatment status.

Table 16 presents our main impact estimates according to the specification (3) in section 3.2. Each panel shows a separate metric of impact and four statistics from a separate regression: the impact estimate, its p-value in a two-side test, the number of observations (baseline plus endline), and the number of classrooms included. There are two sets of columns. The first set includes all schools in the sample: treated schools and untreated schools in the terms described in section 3.2. The second set of columns only includes treated schools, denoted as EPM schools. There are five columns within each set. The first column includes the three educational levels (Primaria, Secundaria and Bachillerato) without socioeconomic and personality controls. The second column adds socioeconomic and personality controls. The remaining three columns present separate estimates by educational level, also including socioeconomic and personality controls. In all cases, the standard errors are clustered by classroom.

Table 16 shows some evidence of a positive short-run impact of Enseña por México fellows after one academic year of exposure. When we use all schools, we find a positive estimate (95% confidence on two-sided tests) for self-efficacy (all levels together with controls), and negative estimates for tardiness and absenteeism (all levels together with and without controls) When we only include EPM schools, we find positive estimates for self-management and growth-mindset (Secundaria), and negative estimates tardiness and absenteeism (all levels together, and Bachillerato). For one metric we find estimates against the theory of change: social awareness (EPM schools only, all levels together without controls and Bachillerato). Lastly, we also find a negative estimate for time devoted to homework (Bachillerato), although this case does not necessarily go against the theory of change—it is possible that a more effective instructor could rely less on giving homework. All significant estimates have a rather modest magnitude: between 0.04 and 0.15 standardized units.

Of the 114 estimates presented in Table 16, 26 are significant or marginally significant (90% confidence on two-sided tests). Sixteen of them point in the direction of fellows causing an improvement in the metrics analyzed, seven point in the opposite direction, and three correspond to time devoted to homework—which we argue is hard to interpret. It is worth noting that five estimates for social awareness are negative and marginally significant. In principle, it is possible that exposure to fellows

could be shifting the reference of that scale. However, as we will show below, that hypothesis is not supported by the evidence.

It could be argued that the socioemotional scales and related attitudes and behaviors in Table 16 belong to the same domain, i.e. to some degree they measure the same underlying skill. If so, presenting separate estimates for each scale using the same treatment and control groups would inflate the number of significant estimates. In we believe they measure the same underlying variable, we could adjust the significance of the estimates using the correction proposed by Benjamini and Hochberg (1995). If we perform the correction, only the results for tardiness and absenteeism remain significant. However, we do not believe all twelve scales belong to the same domain, therefore the correction would to be too conservative.

Table 16. Impact estimates

	All schools					EPM schools only				
Statistic	All levels, no controls	All levels	Bachillerato	Secundaria	Primaria	All levels, no controls	All levels	Bachillerato	Secundaria	Primaria
CORE: Self-management										
Estimate	0.007	0.014	-0.005	0.056	-0.042	-0.006	0.021	-0.028	**0.102**	-0.175
p-value	(0.729)	(0.472)	(0.851)	(0.072)	(0.601)	(0.876)	(0.531)	(0.556)	(0.033)	(0.318)
Observations	52,675	47,666	28,011	17,295	2,360	15,524	14,029	8,780	5,035	214
Classrooms	1,193	1,189	554	443	192	283	283	153	119	11
CORE: Growth mindset										
Estimate	0.001	0.007	0.003	0.081	-0.057	0.008	0.027	-0.018	**0.152**	-0.120
p-value	(0.972)	(0.780)	(0.928)	(0.058)	(0.663)	(0.832)	(0.481)	(0.683)	(0.020)	(0.779)
Observations	54,139	48,760	28,371	17,930	2,459	15,955	14,354	8,891	5,231	232
Classrooms	1,194	1,190	554	443	193	283	283	153	119	11
CORE: Self-efficacy										
Estimate	0.038	**0.042**	0.017	0.055	0.041	0.023	0.049	0.015	0.097	-0.029
p-value	(0.057)	(0.027)	(0.463)	(0.132)	(0.676)	(0.497)	(0.113)	(0.709)	(0.068)	(0.931)
Observations	54,637	49,205	28,711	18,037	2,457	16,132	14,523	9,035	5,261	227
Classrooms	1,194	1,191	554	443	194	283	283	153	119	11
CORE: Social awareness										
Estimate	-0.035	-0.034	-0.046	0.046	0.034	**-0.073**	-0.058	**-0.101**	0.072	-0.455
p-value	(0.095)	(0.104)	(0.077)	(0.208)	(0.768)	(0.049)	(0.099)	(0.044)	(0.151)	(0.408)
Observations	54,440	48,959	28,599	17,926	2,434	16,040	14,430	8,966	5,247	217
Classrooms	1,193	1,189	554	443	192	283	283	153	119	11
Educational expectations										
Estimate	0.003	-0.002	0.001	-0.005	0.100	-0.010	-0.012	-0.006	0.027	-0.601
p-value	(0.890)	(0.893)	(0.981)	(0.890)	(0.339)	(0.738)	(0.656)	(0.881)	(0.531)	(0.075)
Observations	54,645	49,087	28,465	18,146	2,476	16,113	14,487	8,964	5,289	234
Classrooms	1,194	1,190	554	443	193	283	283	153	119	11
Returns on education: General										
Estimate	0.023	0.020	0.029	0.029	0.155	0.013	0.017	0.038	0.043	-0.633
p-value	(0.255)	(0.315)	(0.250)	(0.429)	(0.162)	(0.704)	(0.631)	(0.465)	(0.385)	(0.056)
Observations	55,150	49,530	29,034	17,990	2,506	16,267	14,620	9,125	5,258	237
Classrooms	1,193	1,190	554	443	193	283	283	153	119	11
Returns on education: Pecuniary										
Estimate	-0.012	-0.012	0.020	0.024		-0.033	-0.036	0.043	-0.043	
p-value	(0.607)	(0.616)	(0.501)	(0.566)		(0.374)	(0.350)	(0.366)	(0.446)	
Observations	51,645	46,568	28,589	17,979		15,836	14,244	8,984	5,260	
Classrooms	997	997	554	443		272	272	153	119	
School-related behavior: Tardiness and absenteeism										
Estimate	**-0.068**	**-0.078**	-0.032	-0.083	-0.200	**-0.113**	**-0.134**	**-0.145**	-0.076	-0.111
p-value	(0.002)	(0.001)	(0.254)	(0.058)	(0.099)	(0.003)	(0.001)	(0.002)	(0.257)	(0.849)
Observations	54,888	49,296	28,604	18,172	2,520	16,183	14,540	8,988	5,316	236
Classrooms	1,194	1,190	554	443	193	283	283	153	119	11
School-related behavior: Time devoted to homework										
Estimate	-0.033	-0.037	**-0.061**	-0.008	0.096	-0.054	-0.047	-0.067	-0.006	-0.461
p-value	(0.107)	(0.083)	(0.024)	(0.840)	(0.497)	(0.096)	(0.168)	(0.143)	(0.915)	(0.381)
Observations	55,657	49,910	28,968	18,436	2,506	16,403	14,718	9,097	5,386	235
Classrooms	1,194	1,189	554	443	192	283	283	153	119	11
Community involvement										
Estimate	0.014	0.020	0.029	0.041		0.010	0.026	0.041	0.051	
p-value	(0.516)	(0.350)	(0.261)	(0.262)		(0.777)	(0.418)	(0.348)	(0.283)	
Observations	52,313	47,147	28,853	18,294		16,061	14,426	9,087	5,339	
Classrooms	997	997	554	443		272	272	153	119	
Grit										
Estimate	-0.026	-0.033	-0.060	0.089		-0.040	-0.022	-0.051	0.110	
p-value	(0.395)	(0.278)	(0.103)	(0.079)		(0.434)	(0.656)	(0.386)	(0.157)	
Observations	23,729	21,534	13,245	8,289		7,283	6,573	4,139	2,434	
Classrooms	995	995	553	442		272	272	153	119	
Internal locus of control										
Estimate	0.018	0.014	0.014	0.118	0.195	-0.045	-0.064	-0.071	0.147	-0.376
p-value	(0.633)	(0.713)	(0.759)	(0.055)	(0.212)	(0.489)	(0.323)	(0.383)	(0.116)	(0.530)
Observations	24,794	22,449	13,817	7,500	1,132	7,395	6,712	4,383	2,222	107
Classrooms	1,175	1,160	554	443	163	283	283	153	119	11

Estimates in bold type and shaded are significant at 95% confidence. p-values in parentheses.

An alternative way to estimate impact, under the assumption that the metrics in Table 16 belong to the same domain, is to summarize them in one single metric defined as the first factor of the socioemotional and related scales. We computed that single metric using baseline data, and expressed it in standardized units. Table 17 presents impact estimates under the single-metric approach. The estimates for Secundaria are positive and significant: 0.07 and 0.13 standard deviations considering all schools and EPM schools only, respectively.

Table 17. Estimates of impact using as a metric the first factor of the socioemotional ad related scales

	All schools					EPM schools only				
	All levels, no controls	All levels	Bachillerato	Secundaria	Primaria	All levels, no controls	All levels	Bachillerato	Secundaria	Primaria
Estimate	0.021	0.022	0.003	**0.072**	0.112	-0.001	0.029	-0.003	**0.126**	-0.662
P-value	(0.330)	(0.247)	(0.909)	(0.035)	(0.374)	(0.976)	(0.348)	(0.944)	(0.006)	(0.220)

Estimates in bold type and shaded are significant at 95% confidence. p-values in parentheses.

As we mentioned in section 5.5, treatment classrooms in Bachillerato and Secundaria were exposed to a different number of fellows and were taught a different number of subjects by fellows (see Table 10). Table 18 shows impact estimates for Bachillerato and Secundaria using the specification with controls and three distinct definitions of treatment: (a) a dummy indicating whether a fellow taught a subject to the classroom (which is our basic estimate presented in Table 16), (b) the number of fellows who taught the classroom, and (c) the number of subjects taught by fellows.

When we differentiate by intensity and we focus only on treatment schools (internal control), we find positive effects of having more exposure to fellows in Secundaria on self-management, growth mindset, self-efficacy, social awareness, educational expectations, perceived returns on education, and locus of control. In Bachillerato, we find negative effects on social awareness and time devoted to homework when we use the external control. If we use the internal control, we find negative effects on social awareness, and tardiness and absenteeism.

Table 18. Estimates of the impact of treatment intensity

	Bachillerato						Secundaria					
	All schools			Internal control			All schools			Internal control		
	Definicion of treatment											
	A fellow taught a subject to the classroom	Number of fellows teaching the classroom	Number of subjects taught by fellows	A fellow taught a subject to the classroom	Number of fellows teaching the classroom	Number of subjects taught by fellows	A fellow taught a subject to the classroom	Number of fellows teaching the classroom	Number of subjects taught by fellows	A fellow taught a subject to the classroom	Number of fellows teaching the classroom	Number of subjects taught by fellows
CORE: Self-management	-0.005	0.002	-0.003	-0.028	-0.004	-0.008	0.056	0.038	0.018	**0.102**	**0.098**	**0.045**
	(0.851)	(0.933)	(0.892)	(0.556)	(0.939)	(0.825)	(0.072)	(0.143)	(0.203)	(0.033)	(0.030)	(0.004)
CORE: Growth mindset	0.003	0.006	0.009	-0.018	0.003	0.018	0.081	**0.087**	**0.060**	**0.152**	**0.148**	**0.074**
	(0.928)	(0.835)	(0.731)	(0.683)	(0.963)	(0.681)	(0.058)	(0.023)	(0.006)	(0.020)	(0.025)	(0.002)
CORE: Self-efficacy	0.017	0.022	0.012	0.015	0.034	0.004	0.055	0.039	0.024	0.097	**0.101**	**0.048**
	(0.463)	(0.259)	(0.483)	(0.709)	(0.395)	(0.915)	(0.132)	(0.230)	(0.125)	(0.068)	(0.037)	(0.010)
CORE: Social awareness	-0.046	**-0.054**	**-0.052**	**-0.101**	-0.095	**-0.095**	0.046	0.029	0.026	0.072	0.075	**0.059**
	(0.077)	(0.041)	(0.027)	(0.044)	(0.061)	(0.029)	(0.208)	(0.347)	(0.193)	(0.151)	(0.103)	(0.000)
Educational expectations	0.001	-0.011	0.008	-0.006	-0.045	0.002	-0.005	-0.011	-0.001	0.027	0.043	**0.034**
	(0.981)	(0.636)	(0.714)	(0.881)	(0.326)	(0.955)	(0.890)	(0.742)	(0.968)	(0.531)	(0.307)	(0.004)
Returns on education: General	0.029	0.029	0.036	0.038	0.057	0.067	0.029	0.026	0.033	0.043	0.045	**0.053**
	(0.250)	(0.223)	(0.106)	(0.465)	(0.303)	(0.164)	(0.429)	(0.421)	(0.075)	(0.385)	(0.318)	(0.000)
Returns on education: Pecuniary	0.020	0.019	0.021	0.043	0.068	0.073	0.024	0.014	0.002	-0.043	-0.021	-0.008
	(0.501)	(0.508)	(0.477)	(0.366)	(0.231)	(0.230)	(0.566)	(0.699)	(0.899)	(0.446)	(0.699)	(0.581)
School-related behavior: Tardiness and absenteeism	-0.032	-0.020	-0.024	**-0.145**	**-0.138**	**-0.114**	-0.083	-0.062	-0.023	-0.076	-0.073	-0.017
	(0.254)	(0.488)	(0.302)	(0.002)	(0.021)	(0.007)	(0.058)	(0.102)	(0.315)	(0.257)	(0.224)	(0.546)
School-related behavior: Time devoted to homework	**-0.061**	**-0.057**	**-0.051**	-0.067	-0.061	-0.062	-0.008	-0.032	-0.019	-0.006	-0.025	-0.011
	(0.024)	(0.039)	(0.030)	(0.143)	(0.207)	(0.128)	(0.840)	(0.314)	(0.295)	(0.915)	(0.629)	(0.443)
Community involvement	0.029	0.048	0.033	0.041	0.071	0.037	0.041	0.016	0.004	0.051	0.038	0.007
	(0.261)	(0.062)	(0.115)	(0.348)	(0.142)	(0.316)	(0.262)	(0.623)	(0.801)	(0.283)	(0.402)	(0.671)
Grit	-0.060	-0.011	0.008	-0.051	0.049	0.069	0.089	0.044	0.022	0.110	0.075	0.096
	(0.103)	(0.753)	(0.798)	(0.386)	(0.491)	(0.275)	(0.079)	(0.308)	(0.578)	(0.157)	(0.321)	(0.176)
Locus of control	0.014	0.042	0.024	-0.071	-0.003	-0.019	0.118	0.098	**0.058**	0.147	0.157	**0.058**
	(0.759)	(0.314)	(0.526)	(0.383)	(0.969)	(0.789)	(0.055)	(0.075)	(0.011)	(0.116)	(0.086)	(0.026)

Estimates in bold type and shaded are significant at 95% confidence. p-values in parentheses.

Table 19 presents Two-Stage Least Square (2SLS) estimates of impact using the specification with controls, side by side with the corresponding Ordinary Least Squares (OLS) estimates from Table 16 (with controls). The instrumented variable is the actual treatment status, and the instrument is the assigned treatment status. The estimate for self-efficacy pooling the three educational levels remains positive and significant. The negative estimates for tardiness and absenteeism for all levels and Secundaria remain negative and significant. The 2SLS estimate for Primaria turns negative and significant. Lastly, the negative and significant OLS estimate for time devoted to homework in Bachillerato becomes positive and insignificant under the 2SLS approach. The 2SLS approach does not alter the conclusions. If anything, the resulting estimates are more favorable to the hypothesis of positive impact.

Table 19. Two-Stage Least Square estimates of impact instrumenting actual treatment status with assigned treatment status

	All levels		Bachillerato		Secundaria		Primaria	
	OLS	2SLS	OLS	2SLS	OLS	2SLS	OLS	2SLS
CORE: Self-management	0.014	0.033	-0.005	0.038	0.056	0.002	-0.042	0.052
	(0.472)	(0.284)	(0.851)	(0.333)	(0.072)	(0.975)	(0.601)	(0.646)
CORE: Growth mindset	0.007	-0.015	0.003	0.011	0.081	-0.031	-0.057	0.119
	(0.780)	(0.679)	(0.928)	(0.800)	(0.058)	(0.706)	(0.663)	(0.512)
CORE: Self-efficacy	**0.042**	**0.058**	0.017	0.007	0.055	0.097	0.041	0.101
	(0.027)	(0.049)	(0.463)	(0.832)	(0.132)	(0.134)	(0.676)	(0.465)
CORE: Social awareness	-0.034	-0.027	-0.046	0.002	0.046	0.007	0.034	-0.060
	(0.104)	(0.406)	(0.077)	(0.964)	(0.208)	(0.915)	(0.768)	(0.668)
Educational expectations	-0.002	-0.008	0.001	0.001	-0.005	-0.013	0.100	0.133
	(0.893)	(0.780)	(0.981)	(0.981)	(0.890)	(0.830)	(0.339)	(0.345)
Returns on education: General	0.020	-0.034	0.029	-0.017	0.029	-0.035	0.155	0.160
	(0.315)	(0.320)	(0.250)	(0.673)	(0.429)	(0.613)	(0.162)	(0.302)
Returns on education: Pecuniary	-0.012	-0.008	0.020	0.024	0.024	0.089		
	(0.616)	(0.822)	(0.501)	(0.577)	(0.566)	(0.228)		
School-related behavior: Tardiness and absenteeism	**-0.078**	**-0.086**	-0.032	0.035	**-0.083**	**-0.203**	-0.200	**-0.372**
	(0.001)	(0.017)	(0.254)	(0.415)	(0.058)	(0.012)	(0.099)	(0.042)
School-related behavior: Time devoted to homework	-0.037	0.009	**-0.061**	0.006	-0.008	0.015	0.096	-0.046
	(0.083)	(0.802)	(0.024)	(0.892)	(0.840)	(0.838)	(0.497)	(0.828)
Community involvement	0.020	0.006	0.029	0.033	0.041	0.005		
	(0.350)	(0.847)	(0.261)	(0.406)	(0.262)	(0.942)		
Grit	-0.033	-0.012	-0.060	-0.036	0.089	0.137		
	(0.278)	(0.788)	(0.103)	(0.506)	(0.079)	(0.155)		
Locus of control	0.014	0.040	0.014	0.047	0.118	0.234	0.195	0.145
	(0.713)	(0.470)	(0.759)	(0.481)	(0.055)	(0.053)	(0.212)	(0.522)

Estimates in bold type and shaded are significant at 95% confidence. p-values in parentheses.

6.3 Reference bias

A potential concern expressed by Teach For All was the presence of reference bias in the measurement of socioemotional skills and related metrics of impact (as explained in section 3.3). Table 20 shows the results of the regressions to assess the presence of reference bias. The structure of the table is the same as that of Table 16. However, in this case there are two samples. One sample (top panel) includes all the siblings reported by respondents. It includes respondents' siblings in the school attended by the respondent and who therefore may have interacted with fellows. The other sample (bottom panel) only includes siblings who do not attend the school of the respondent. We assume that siblings in this second sample did not interact with fellows. We do not find evidence of reference bias. In the top panel there are no significant estimates. In the bottom panel there is only one significant estimate. It is important to mention that the five items of the social awareness scale shown at the bottom of Table 20 do not show evidence of reference bias. Thus, the negative estimates for that scale in Table 16 cannot be attributed to this type of bias.

Table 20. Reference bias estimates

	All schools					EPM schools only				
	All levels, no controls	All levels	Bachillerato	Secundaria	Primaria	All levels, no controls	All levels	Bachillerato	Secundaria	Primaria
Any sibling in middle or high school										
CORE: Self-management item 1										
Estimate	0.035	0.029	0.023	0.075	-0.055	0.039	0.045	0.021	0.113	-0.481
p-value	(0.094)	(0.196)	(0.424)	(0.069)	(0.571)	(0.283)	(0.231)	(0.686)	(0.070)	(0.099)
Observations	43,638	39,075	21,770	15,756	1,549	12,840	11,497	6,782	4,558	157
Classrooms	1,154	1,135	554	443	138	283	283	153	119	11
CORE: Self-efficacy item 3										
Estimate	0.028	0.016	0.008	0.058	0.003	0.028	0.030	0.017	0.094	-0.365
p-value	(0.283)	(0.575)	(0.825)	(0.227)	(0.975)	(0.505)	(0.501)	(0.759)	(0.211)	(0.135)
Observations	29,117	26,168	15,148	9,799	1,221	8,875	7,974	4,870	2,978	126
Classrooms	1,149	1,130	554	442	134	283	282	153	118	11
CORE: Social awareness items 1, 3, 6, 7 & 8										
Estimate	-0.007	-0.009	-0.032	0.036	-0.010	0.001	0.016	-0.042	0.086	0.087
p-value	(0.730)	(0.681)	(0.240)	(0.282)	(0.947)	(0.972)	(0.634)	(0.355)	(0.082)	(0.879)
Observations	43,586	39,031	21,747	15,740	1,544	12,822	11,487	6,778	4,552	157
Classrooms	1,154	1,135	554	443	138	283	283	153	119	11
Any sibling middle or high school, not in respondent's school										
CORE: Self-management item 1										
Estimate	-0.007	-0.018	-0.027	-0.007	0.044	-0.032	-0.024	-0.044	-0.002	0.098
p-value	(0.778)	(0.510)	(0.434)	(0.874)	(0.783)	(0.430)	(0.552)	(0.408)	(0.976)	(0.904)
Observations	29,106	26,156	15,142	9,796	1,218	8,867	7,968	4,872	2,970	126
Classrooms	1,148	1,129	554	442	133	283	282	153	118	11
CORE: Self-efficacy item 3										
Estimate	0.012	0.009	-0.003	0.057	-0.096	0.043	0.071	0.041	**0.139**	0.172
p-value	(0.602)	(0.697)	(0.915)	(0.142)	(0.330)	(0.255)	(0.080)	(0.468)	(0.016)	(0.725)
Observations	42,883	38,475	21,510	15,469	1,496	12,604	11,307	6,687	4,468	152
Classrooms	1,150	1,135	554	443	138	283	283	153	119	11
CORE: Social awareness items 1, 3, 6, 7 & 8										
Estimate	0.016	0.009	0.026	-0.001	0.049	0.013	0.033	0.054	0.037	0.639
p-value	(0.576)	(0.777)	(0.493)	(0.981)	(0.709)	(0.775)	(0.501)	(0.418)	(0.586)	(0.280)
Observations	28,661	25,794	14,985	9,628	1,181	8,717	7,842	4,803	2,916	123
Classrooms	1,142	1,125	554	441	130	283	282	153	118	11

Estimates in bold type and shaded are significant at 95% confidence. p-values in parentheses.

7 Discussion of findings

The impact evaluation presented here constitutes an unprecedented effort in the measurement of socioemotional skills in Mexico and in the Teach For All network. We visited (twice) 1,194 classrooms in 328 schools across four states, and collected 56,516 student surveys and 833 instructor surveys. The main lessons from the evaluation can be summarized in three points.

First, Enseña por México fellows differ from regular teachers in their teaching values and attitudes. Fellows give more importance to developing student curiosity instead of student competencies. They think it is more important for students to set ambitious goals instead of making concrete, achievable plans. Fellows engage in extracurricular activities, use evaluations, and give feedback to students to a larger extent than regular teachers. In terms of the Big Five personality traits, fellows are more open to new experiences, more extraverted, and more agreeable. Fellows score between 0.15 and 0.30 standard

deviations higher than regular teachers in Tripod's 7Cs, a set of scales based on student surveys that are predictive of students' academic achievement.

Second, taking together the main estimates and the estimates considering (a) a single metric, (b) heterogeneity in intensity and (c) treatment fidelity, the image that emerges is that exposure to fellows is associated with modest short-run improvements in socioemotional skills in Secundaria, as measured by the CORE scales: self-management, growth mindset, self-efficacy and social awareness. There is also evidence of a reduction in tardiness and absenteeism in all educational levels. In general, the magnitudes are modest—the main estimates are below 0.15 standardized units. These results are in line with evaluations of Teach For America (Backes and Hansen, 2017) and Enseña Chile (Alfonso, Santiago and Bassi, 2010). Based on our analysis of potential reference bias, we conclude this is likely not an issue in the estimates we present.

Third, accounting for heterogeneity in treatment intensity produces more estimates of improvements in Secundaria. We interpret this finding as indicative of a window of greater potential impact on students in their early teens (grades 7-9) relative to older teenagers (grades 10-12). It also suggests that Enseña por Mexico could gain from sending more fellows to fewer schools—increasing the intensive margin of the program at the expense of the extensive margin.

In sum, the results indicate that Enseña por México fellows are more effective than regular teachers, and that they help foster student socioemotional skills. We advise taking the results with caution. First, the evaluation is quasi-experimental and relies on the assumption that the trends in the metrics of impact absent the treatment would have been the same in the treatment and control groups. Second, the estimates are limited to the short-run, and one year of exposure to fellows. The evaluation is silent about what could happen with a longer exposure to fellows or with multi-year follow ups of students. Third, the partnerships that Enseña por México establishes with local educational authorities to deploy the program are very idiosyncratic. Fellows in some states assume roles that greatly differ from roles assumed in other states—CONAFE schools in Chiapas and Hidalgo are a clear example. Extrapolating the results in such an idiosyncratic context may not be entirely appropriate.

References

Alfonso, M., Santiago, A. & Bassi. 2010. Estimating the Impact of Placing Top University Graduates in Vulnerable Schools in Chile. Inter-American Development Bank, Education Division, Technical Notes, No. IDB-TN-230.

Allen, R. and Allnutt, J. 2017. The impact of Teach First on pupil attainment at age 16. British Educational Research Journal, 43: 627–646.

Antecol, H., Eren, O., & Ozbeklik, S. 2013. The effect of Teach for America on the distribution of student achievement in primary school: Evidence from a randomized experiment. Economics of Education Review 37, 113-125.

Arocha, M.J. & Ledezma, L.E. 2007. Construcción, Validación y Confiabilidad de un Inventario de Locus de Control Académico (ILC-A). Revista Iberoamericana de Diagnóstico y Evaluación, Nº 24, Vol. 2, pp. 151–175.

Backes, B. & Hansen M. 2017. The Impact of Teach for America on Non-Test Academic Outcomes, Education Finance and Policy.

Benjamini, Y., & Hochberg, Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B Methodological, 57(1), 289–300.

Bill & Melinda Gates Foundation. 2012. Asking Students about Teaching. Student Perception Surveys and Their Implementation.

Blackwell, M., Iacus, S., King, G., & Porro, G. 2009. cem: Coarsened exact matching in Stata", The Stata Journal 9, Number 4, pp. 524–546.

Brown, A. & Maydeu-Olivares, A. 2018. Modeling forced-choice response formats. In Irwing, P., Booth, T. & Hughes, D. (Eds.), The Wiley Handbook of Psychometric Testing. London: John Wiley & Sons.

Blazar, D. & Kraft, M.A. 2017. Teacher and Teaching Effects on Students' Attitudes and Behaviors." Educational Evaluation and Policy Analysis, 39, 146-170.

Chacón, A. & Peña, P.A. 2015. A Cross-Sectional Impact Evaluation of Enseña por Mexico in High Schools of the State of Puebla. Microanalitica.

Chacón, A. & Peña, P.A. 2016. Evaluación de desempeño de los Profesionales Enseña por México. Una comparación de corte transversal. Microanalitica.

Chiang, H. S., Clark, M. A. and McConnell, S. 2017. Supplying Disadvantaged Schools with Effective Teachers: Experimental Evidence on Secondary Math Teachers from Teach For America. J. Pol. Anal. Manage., 36: 97–125.

Doolittle, A., & Faul, A. C. 2013. Civic Engagement Scale: A Validation study. Sage Open Journal, 3, 1-7.

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. Journal of Personality and Social Psychology, *92*(6), 1087-1101.

Duckworth. A.L. & Quinn, P.D. 2009. Development and Validation of the Short Grit Scale (Grit–S), Journal of Personality Assessment, 91:2, 166-174

Duckworth, A.L. & Yeager, D.S. 2015. Measurement Matters: Assessing Personal Qualities Other Than Cognitive Ability for Educational Purposes, Educational Researcher Vol 44, Issue 4, pp. 237-251

Farrington, C.A., Roderick, M., Allensworth, E., Nagaoka, J., Keyes, T.S., Johnson, D.W., & Beechum, N.O. 2012. Teaching adolescents to become learners. The role of noncognitive factors in shaping school performance: A critical literature review. Chicago: University of Chicago Consortium on Chicago School Research.

Glazerman, S., Mayer, D. & Decker, P. 2006. Alternative routes to teaching: The impacts of Teach for America on student achievement and other outcomes, Journal of Policy Analysis and Management, Volume 25, Issue 1, Winter, Pages 75–96.

Gertler, P.J., Martinez, S., Premand, P., Rawlings, L.B., Vermeersch, C.M.J. 2011. Impact Evaluation in Practice, First Edition. World Bank.

Goldberg, L. R. (1993). The structure of phenotypic personality traits. American Psychologist, 48(1), 26-34.

Heckman, J.J., Stixrud, J. & Urzua, S. 2006. The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior. Journal of Labor Economics 24 (3): 411–48.

Jackson, C.K. 2016. What Do Test Scores Miss? The Importance of Teacher Effects on Non-Test Score Outcomes. NBER Working Paper No. 22226.

Jennings, J.L. & DiPrete, T.A. 2010. Teacher Effects on Social and Behavioral Skills in Early Elementary School." Sociology of Education, 83, 135-159.

John, O.P. & Srivastava, S. 1999. The Big Five Trait Taxonomy: History, measurement, and Theoretical Perspectives. Handbook of Personality, Second Edition: Theory and Research. Lawrence A. Pervin, Oliver P. John, Elsevier.

Kane, T.J., McCaffrey, D.F., Miller, T. & Staiger, D.O. 2013. Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment. MET Project Research Paper.

Kane T.J., Rockoff J.E., Staiger D.O. 2008. What does certification tell us about teacher effectiveness? Evidence from New York City. Economics of Education Review, 27, 615-631.

Kraft, M.A. 2017. Teacher Effects on Complex Cognitive Skills and Social-Emotional Competencies. Brown University.

Kyllonen P.C. & Bertling J. P. 2013. Innovative questionnaire assessment methods to increase cross-country comparability. In Rutkowski L., von Davier M., Rutkowski D. (Eds.), Handbook of international large-scale assessment: Background, technical Issues, and methods of data analysis (pp. 277-285). London, England: Chapman & Hall.

Ladd, H.F. & Sorensen, L.C. 2017. Returns to Teacher Experience: Student Achievement and Motivation in Middle School. Education Finance and Policy.

Meade, A.W. 2004. Psychometric problems and issues involved with creating and using ipsative measures for selection. Journal of Occupational and Organizational Psychology 77(4):531-551.

OECD. 2017. Education at a Glance 2017: OECD Indicators, OECD Publishing, Paris.

OECD. 2014. TALIS 2013 Technical Report.

Peña, P.A. 2016. Personality and Financial Culture: A Study of Mexican Youths, International Handbook of Financial Literacy, Springer, pp. 465-493.

Rammstedt, B. & John, O.P. 2007. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. Journal of Research in Personality, Volume 41, Issue 1, pp. 203-212.

Rotter, J.B. 1966. Generalized Expectancies for Internal Versus External Control of Reinforcement. Psychological Monographs: General and Applied. Vol. 80, No. 1, pp. 1-28.

Ruzek, E.A., Domina, T., Conley, A.M., Duncan, G.J. & Karabenick, S.A. 2014. Using value-added models to measure teacher effects on students' motivation and achievement," The Journal of Early Adolescence, 1-31.

Sánchez Puerta, M.L., Valerio, A. & Gutiérrez Bernal, M. 2016. Taking Stock of Programs to Develop Socioemotional Skills: A Systematic Review of Program Evidence. Directions in Development—Human Development. Washington, DC: World Bank.

West, M.R. 2016. Should non-cognitive skills be included in school accountability systems? Preliminary evidence from California's CORE districts. Economics Studies at Brookings. Evidence Speaks Reports, 1(13).

Xu, Z., Hannaway, J., & Taylor, C. 2011. Making a difference? The effects of Teach For America in high school. Journal of Policy Analysis and Management, 30(3), 447-469.