I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 14192

Enhancing Human Capital at Scale

Francesco Agostinelli
Ciro Avitabile
Matteo Bobba

MARCH 2021 (THIS VERSION: SEPTEMBER 2022)

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 14192

Enhancing Human Capital at Scale

Francesco Agostinelli
University of Pennsylvania

Ciro Avitabile
World Bank

Matteo Bobba
University of Toulouse Capitole and IZA

MARCH 2021 (THIS VERSION: SEPTEMBER 2022)

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Enhancing Human Capital at Scale*

This paper provides new insights on the science of scaling. We study an educational mentoring program with a home visit component implemented at scale in Mexico, under different modalities (original and enhanced training for mentors) and different situations (field experiment and policy implementation). While the program was ineffective when implemented by the government in its original modality, the enhanced modality boosts children's outcomes, both in the field experiment and during the government implementation. Higher-quality home visits encourage parent/child and parent/community interactions, which in turn are found to promote the scalability of the program. Our work provides new knowledge on the socially determined nature of scaling educational programs.

JEL Classification: C90, C93, D02, I3, J1

Keywords: children's skills, parental investment and community engagement, science of scaling

Corresponding author:
Matteo Bobba
Toulouse School of Economics
University of Toulouse Capitole
21 Allée de Brienne
31015 Toulouse
France
E-mail: matteo.bobba@tse-fr.eu

* We are grateful to the *Consejo Nacional de Fomento Educativo* (CONAFE) for the generous collaboration throughout this project, Alonso Sanchez for his initial input into the project, and Miguel Angel Monroy for excellent research assistance. We thank the anonymous referees, as well as Jere Behrman, Horacio Larreguy, John List, Giuseppe Sorrenti, and Stephane Straub for helpful comments and discussions. Avitabile acknowledges financial support for data collection from the Strategic Impact Evaluation Fund (SIEF) of the World Bank and the *Consejo Nacional de Evaluación de la Politica de Desarollo Socia*l (CONEVAL). Bobba acknowledges financial support from the AFD, the H2020-MSCA-RISE project GEMCLIME-2020 GA No 681228, and the ANR under grant ANR-17-EURE-0010 (Investissements d'Avenir program). This study is registered in the AEA RCT Registry and the unique identifying number is AEARCTR-0001645.

1 Introduction

One of the main challenges in using scientific insights to inform policy decisions comes from the fact that small differences in the implementation of any given intervention can translate into substantial differences in outcomes. Even when programs display large and significant effect sizes in randomized control trials, their success in different situations is far from guaranteed (List, 2022).

This paper contributes to the recent debate about the challenges to scale-up interventions aimed at enhancing human capital in children. In particular, we provide a case study of a mentoring program implemented at scale in Chiapas, the poorest state in Mexico. The program assigns recent university graduates to remote and disadvantaged communities. Among other things, mentors encourage parental involvement in children's education through home visits. We evaluate the relative effectiveness of two program modalities that differ both in terms of content and the intensity of the training provided to the frontline mentors. The *Original* modality features a training module focused on curricular knowledge and pedagogical notions, which was initially implemented by the government. The *Plus* modality embeds a significant change in the training module, which was designed and tailored by our research team to guarantee its operational continuity in the event of a national rollout of the program. The new training protocol includes periodic peer-to-peer meetings during which mentors share their experiences regarding the home visits and their interactions with families.

The evidence on the relative effectiveness of the two program modalities is based on two independent field experiments. The first experiment was directly carried out by the government during the ongoing national rollout of the *Original* modality of the mentoring intervention. Assignment to the program was randomized across 80 program-eligible primary schools, with 40 getting access to mentors. The results show that the program had no discernible effect on children's achievement outcomes, as measured by standardized test scores. In the second experiment we randomly assigned both the *Original* and the *Plus* modality as well as a control group with no mentoring program across 230 primary schools. After two years of exposure to the mentoring program, the *Original* program modality displays relatively small and noisy effects on cognitive and socio-emotional scores, as well as on educational achievements when compared to the control group with no mentors. The *Plus* modality delivers sizable and significant gains in children's reading scores (+0.32 standard deviations), math scores (+0.24 standard deviations), and socio-emotional scores (+0.20 standard deviations)

as well as a large, albeit marginally significant, effect on the probability of enrolling in seventh grade (+12.7 percentage points, out of a basis of 62 percent enrollment in the control group).

The large difference in effect sizes between the two training modalities is corroborated by direct evidence on parental behavior. While both experiments unequivocally display inconclusive evidence on parents' investment under the *Original* modality, the *Plus* modality significantly increases parental engagement both toward the child's education activities and toward the school community—including volunteering activities, as well as in-kind donations. We further show some evidence that mentors with enhanced training engage in higher-quality interactions with parents during the periodic encounters. In particular, mentors with enhanced training are more likely to inform parents about their children's learning difficulties, to provide concrete advice to parents on how to tackle these difficulties, as well as to promote parenting styles that are centered around communicating with the child and learning activities. We complement these empirical patterns with qualitative evidence that confirms the role of the peer-to-peer sessions as the driving factor behind both the enhanced parent/mentor interactions and the increased parental engagement.

After the release of this evidence, the government autonomously decided to replace the *Original* modality of the program with the *Plus* modality for all its primary schools throughout the country, including those that were part of the experimental evaluation. This reform provides us with a unique opportunity to study the determinants and mechanisms of scaling. One of the key situational differences between the experimental setting and the policy implementation comes from the fact that several schools in this context are at risk of closure, an event that has disruptive consequences for children's learning and their educational trajectories.[1] While the intense monitoring during the experimental evaluation has minimized the extent of school closures, this high-stakes implementation feature may compromise the success of the program under the business-as-usual conditions. However, two years after the rollout, none of the schools that received a mentor during the government implementation closed, while approximately 10 percent of the other schools did.

We next zoom into the relationship between exposure to the mentors and school closures in order to study the sources of scalability of the program. Parents play an important role in the community-based schooling system under study (Gertler et al., 2012). We sketch a

[1]The importance of keeping schools open for the development of children has recently gained momentum in educational studies on the impact of the COVID-19 lockdowns on schooling outcomes (see, e.g., Agostinelli et al., 2022; Engzell et al., 2020; Maldonado and De Witte, 2020).

simple model of parental investment with local spillovers to formalize the idea that parents have an active role in promoting educational opportunities, and that educational investment at the community level are a socially determined outcome (List et al., 2020). We show that the extent to which an educational program preserves or loses impact at scale depends upon its ability to promote parental coordination and engagement in the local community. We empirically corroborate these predictions by leveraging the changes in community-level parental engagement induced by the experiment. Using this variation, we document that parents prevent schools from closing, and as a consequence promote the effectiveness of the mentoring intervention during the government implementation. We find that an increase of half standard deviations in parental engagement decreases the probability of school closures by 11 percentage points over the subsequent two years. Our qualitative data from in-depth surveys of mentors and local instructors further corroborate the role of parents in guaranteeing the continuation of educational activities in the communities, in a context with poor school infrastructure and where schooling activities are often disrupted.

Finally, we study the educational impacts of the policy reform across the overall population of schools in the state of Chiapas. The assignment of the mentoring program under the *Plus* modality at scale was done through a rotating scheme with a priority-based mechanism. We exploit the quasi-experimental variation in the program rollout once we condition on the set of eligibility criteria officially used by the government. After providing evidence that this variation appears conditionally "as-good-as random" via various placebo tests, we show that the program was successful in the schools that were previously part of the evaluation sample as well as in the rest of the schools in Chiapas. Within the evaluation schools, the marginal effect of the *Plus* modality after one year of government implementation on the probability of enrolling in seventh grade is +5.4 percentage points. The cumulative effect of continuous exposure to the program for three years (two years under the experiment and one year under the government) implies that the enrollment rates in these disadvantaged and rural areas achieve the secondary school enrollment rates in urban Mexico (95 percent). For the much larger sample of schools that did not participate in the experimental evaluation, the results show a positive effect on secondary school enrollment, with an average program impact of 4.5 percentage points under the *Plus* modality at scale. We further document positive effects of the program on child literacy, which imply a reduction of illiteracy rates by 20 percent with respect to the sample mean, as well as a decrease in school closures that is remarkably similar to the corresponding impact of the program under the experimental assignment. Taken together, these findings corroborate the effectiveness of the intervention in increasing

schooling opportunities under the new situation created by the policy implementation.

Relationship to Literature. There is a consensus in the literature that gaps in family investment and parent/child interactions are behind the gaps in children's achievements among different socio-economic groups (Cunha et al., 2010; Fryer et al., 2015; Agostinelli and Wiswall, 2016). Moreover, the literature provides ample evidence that successful home visit and mentoring programs, although implemented in very different contexts, share the common outcome of stimulating parental investment and parent/child interactions. Several studies in developing countries document increased cognitive and socio-emotional outcomes for children in the first years of life (see Attanasio et al., 2022b, for a complete review). In the United States, successful interventions like the Perry Preschool project, and the Carolina Abecedarian project show improvements in the home environment (see Heckman and Mosso, 2014, for a complete review). The quality of child/home-visitor interactions and parent/home-visitor interactions are found to be key ingredients for boosting the impact of home visiting programs (Carneiro et al., 2019; Heckman and Zhou, 2021; Zhou et al., 2021). However, little is known on the role of family and community interactions in sustaining program impacts at scale.[2] Our study fills this gap by providing direct evidence on how enhanced training for home visitors promotes higher parental engagement and interactions at the community level, which in turn promotes the success of the program at scale. To the best of our knowledge, we are the first to highlight that parents can act as means of scalability, which has implications for the design and the evaluation of scalable educational interventions that actively include parents in the learning process.

In recent years, scholars and policy makers alike have been increasingly concerned about the ability of field experiments to inform policy decisions, given that experimental interventions that have been found effective often fail to live up to their promises when implemented at scale by governments or firms (Bold et al., 2018; Cameron et al., 2019; Muralidharan and Singh, 2020). Based on the insights of recent work (Banerjee et al., 2017; Muralidharan and Niehaus, 2017; Davis et al., 2017; Mobarak and Davis, 2021), our educational intervention is well-equipped to overcome some of the major threats to scalability. We highlight the infor-

[2]For example, Zhou et al. (2021, p. 90) state: "*The body of research discussed above clearly identifies the key mechanism by which home visiting programs positively impact short-term and long-term outcomes for children: fostering engagement between caregiver and home visitor to improve the caregiver's quality and frequency of caregiver–child interaction, thereby fostering child development. This volume, including this chapter, seeks to move the field toward understanding how to effectively scale up promising interventions and inspire more research on the subject.*" In their recent review, Attanasio et al. (2022b, p. 886) raise another important issue: "*[S]calability does not only refer to the financial cost of running these interventions but also to the ownership and acceptability of the intervention by the community that is targeted. How should interventions be designed and delivered to take account of this important distinction?*"

mative features for scalability of our experimental design following the key points analyzed in Al-Ubaydli et al. (2020). First, we harness the value of replication by drawing joint inference from two independently run field experiments on different and representative samples of schools that share one of the two program modalities. Second, the field experiments were run while the program was already at scale and in close collaboration with the government agency that was later in charge of the program's rollout and policy implementation. Third, the government agency and the research team designed the *Plus* modality together, bearing in mind the financial and human resources constraints of the context under study. Finally, the relatively large units of randomization (schools/communities) take into account possible local spillover effects that often arise in the context of interventions evaluated at scale (Miguel and Kremer, 2004; Bobba and Gignoux, 2019; List et al., 2020).

2 Context and Experimental Design

The *Consejo Nacional de Fomento Educativo* (CONAFE) is a semi-autonomous government agency responsible for providing schooling services in highly marginalized communities of Mexico with a population below 2,500 inhabitants. In those communities, CONAFE offers all education services from pre-school until the end of lower secondary school (hereafter, we refer to the population of CONAFE schools as schools). In 2013, these schools accounted for 10 percent of the roughly 99,000 primary schools and 7 percent of the 38,000 lower secondary schools across the 31 Mexican states. About 20 percent of the schools are located in Chiapas, the Mexican state with the highest incidence of poverty in the country (CONEVAL, 2018).

Primary schools typically have a single multi-grade classroom with 10–15 students. Instructors are generally community residents between 15 and 29 years old. Only 2.6 percent report having a college degree, while 19 percent report having only completed lower secondary education. Instructors are supposed to receive between five and seven weeks of training, but more than half report four weeks of training or less. They receive a stipend of MXN $1,427 per month (US $95 in 2015). After one year of service in the community school, instructors become eligible to receive a scholarship of MXN $982 per month for up to 30 months, which is conditional on enrolling in a higher education institution. As a result of the very low compensation and extremely challenging conditions, about one quarter of the instructors drop out before completing the first school year (Bando and Uribe, 2016).

2.1 The Mentoring Program

In 2009, the government launched the "Mobile Mentors" (*Asesores Pedagogicos Itinerantes,* API henceforth) program as an attempt to improve the quality of education provision in primary schools located in the most-deprived communities. Initially, the program was implemented in 11 states, but starting in 2012, it was extended to all 31 states in Mexico. The mentors are selected from recent university graduates (the program was advertised both during on-campus visits and announcements through the media). Preference is given to applicants with degrees in pedagogy, psychology, sociology, and social services who have previous experience as community instructors and who speak an indigenous language. They are usually hired for a two–year period and receive a monthly salary of MXN $5,000 (US $332 in 2015).

After a week-long training session focused on curricular knowledge and basic notions of pedagogy, the mentors are assigned to schools on a rotating-based algorithm, which gives differential priority across the school communities according to four criteria: (i) at least 30 percent of the students are classified as "insufficient" in the National Standardized test; (ii) at least six students are enrolled, (iii) there are high levels of poverty and marginalization in the respective municipalities; and (iv) the school has not received a mentor in previous academic cycles. Mentors meet with their supervisors every two months in two-day sessions throughout the school year. In December 2018, there were 1928 mentors deployed throughout Mexico, and the largest share is in Chiapas (20 percent).

The mentors periodically conduct home visits to provide parents with information on their children's progress in school and promote their participation in school activities. In addition to working on behavioral issues directly with the children, the mentors are supposed to address them with parents as part of the home visits. Each mentor is assigned to a maximum of six students for individual (one-on-one) remedial education sessions that take place after the regular instructional time. Student eligibility for the remedial sessions is determined by a diagnostic evaluation at the beginning of the school year and an additional exam to assess the grade to which the student's knowledge corresponds. During regular school hours, the mentor is supposed to observe and take notes on the teaching practices of the community instructor, help her with the students who have learning difficulties, and work outside the classroom with students who are unable to attend the remedial sessions in the afternoon.

2.2 Two Field Experiments

In this section we describe the field experiments that we use to evaluate the impact of two different modalities of the API program. A full description of the different data sources used throughout the empirical analysis is provided in Appendix A. Because both experiments were run at scale within the infrastructure of the existing program, the recruitment process and the assignment mechanism of the mentors are the same between experimental and non-experimental schools.

First Experiment. As part of an effort to evaluate a broader set of interventions targeted to families and schools in disadvantaged communities (Martínez, 2012), in 2010, the government undertook the first impact evaluation of the API program. Eighty primary schools are selected among those that met the eligibility criteria for the program across four Mexican states (72 in Chiapas). Assignment to the API program was randomized at the school level using a block design, with the strata represented by the Mexican states where schools are located. Forty schools were assigned to receive the API program starting from the 2011–2012 school year while the remaining half of the schools were assigned to the control group without mentors.

A mid-line survey collected after one year of the API assignment recorded parental behaviors and investments for 208 parents in 73 schools (the enumerators were not able to reach the parents in seven schools). Student outcomes were measured two years after treatment assignment through the results in the national standardized test for students in grades three through six. Due to the incomplete take-up of the test—mainly due to the opposition from the teachers' unions in some states—we were able to match 70 schools with 599 test score records out of the subsample of 73 schools with parental outcomes. Both sources of sample attrition are orthogonal to treatment assignment. Five of the unmatched schools were in the treatment group and five were in the control group. Table B-1 shows the balance for the original sample as well as the nested samples with parental outcomes and student outcomes between the treatment and the control groups across mean community-level and school-level characteristics measured in the year before the start of the first experiment.

Second Experiment. In 2014, as part of a World Bank project, we designed and evaluated

an alternative training modality aimed at strengthening the original mentoring program.[3] The API *Plus* modality embeds all the features of the *Original* modality, with two significant changes in the training module. First, it entails two weeks rather than one week of initial training. The extra week is focused on hands-on strategies to improve students' reading and math competencies. Second, mentors attend an additional day during each one of the bimonthly meetings throughout the school year. The schedule of the extra day is organized around peer-to-peer sessions in which mentors share experiences and design common strategies to better address the most pressing issues. Among the large array of possible improvements in the design of the program, enhanced training was ex-ante scalable and yet a promising alternative from the perspective of the government. The *Plus* modality entails a cost per child of US $332 as opposed to US $285 per child for the *Original* modality. These cost figures are very much in line with another recent government-run intervention that targeted both children and parents through home visits in Colombia (Attanasio et al., 2022a).

We randomly selected 230 schools in rural Chiapas from a set of schools that were not previously part of the API program. Assignment of the program was carried out using a randomized block design at the school level, with the strata represented by the deciles of the 2012 school-average in a national standardized achievement score in the Spanish test (see Appendix A.1). As a result, 60 schools were assigned to the API *Plus*, 70 schools were assigned to the API *Original* modality, and the remaining 100 schools were in the control group with no API intervention.

We rely on both administrative and survey data sources as well as qualitative interviews for the second field experiment. Most of these variables are shown in Table B-2, and they are balanced with respect to treatment assignment. The data collection took place by the end of the second school year after the inception of the API program in the evaluation sample. By that time, two schools out of the original 230 schools in the evaluation sample had closed, while the program could not be put in place in another four schools due to high political instability. Within the remaining 224 schools, one quarter of the community instructors reported eight or fewer months of tenure in the school, and only 56 out of the original 126 mentors were working in the same schools to which they had been originally assigned. All

[3]The *Original* modality in the second experiment is meant to track the benchmark intervention with two minor differences. First, the ability to speak the main indigenous language in the community would become the most important criterion for the assignment of the mentors across program-eligible communities. Second, the supervisors of the mentors would receive a salary increase in exchange for a mandatory increase in the frequency of their visits to the targeted communities.

these outcomes are well balanced across treatment arms. Out of the six schools that dropped out of the sample, two schools were in the control group, two were in the *Original* group, and two in the *Plus* group. The p-values of the Komolgorov-Smirnoff statistic for the equality of the distributions of work experience in the school of the community instructors in each treatment arm and the control group are 0.773 and 0.892, respectively. The p-value of the *Plus-Original* difference in the share of mentors who drop out from the program during the experiment is 0.957. There is no evidence of composition changes between the *Original* and *Plus* groups induced by mentor turnover (see Table B-3).

The effective sample size of the second experiment is 1,045 children/parents and 224 schools (see Appendix A.2 for further details on the sampling design). We use the Early Grade Reading Assessment (reading score) and the Early Grade Math Assessment (math score) as our main measures of children's cognitive achievement. Those are individually administered student assessments that have been conducted in more than 40 countries and in a variety of languages (Dubeck and Gove, 2015; Platas et al., 2016). While these instruments are typically applied to students in first, second, or third grade, we administer them to third through six grade students to account for the large learning gaps of the children in our sample. The school-average standardized scores in math and Spanish as measured in the school year prior to the introduction of the second experiment are, respectively, 0.5 and 0.7 standard deviations below the national averages.[4] To measure the impact of the intervention on socio-emotional skills, we consider a collection of thirty-two behavioral issues as reported by a caregiver, which resembles the questionnaire in the Children section of the National Longitudinal Study of Youth (CNLSY-79), such as antisocial behavior, anxiety/depression, headstrongness, hyperactivity and peer conflicts (for details, see Appendix A.2). The resulting behavioral problem index is re-scaled in such a way that higher values are associated with fewer behavioral issues (socio-emotional score). The survey also contains a module on instructors' characteristics as well as pedagogical practices collected through an adapted version of the Stallings Classroom Snapshot (Bruns and Luque, 2015), a module on parental attitudes and investment toward children's education, as well as information about the mentors' activities in the communities, among others. To better interpret our results, we standardize most of the survey-based outcome variables using the mean and the standard

[4]Only 5 percent of the children in our sample score at the maximum of the scale in two or more subdomains of the reading score (out of eight subdomains) and in three or more subdomains of the math score (out of a total of seven subdomains). Unlike the first experiment, we cannot leverage the national standardized test scores for the second experiment since the test ceased to be universal during the period of interest (after 2014).

deviation observed in the control group.

In addition, we access separate administrative data on students' records that we use to construct an indicator for enrollment in seventh grade, which is the first grade in lower secondary school. The sample reduces to 468 sixth graders in 182 schools, who are deciding whether to transit to secondary school. This sample reduction is due to the multi-grade aspect of the schooling system, where student composition among grades in each school is not homogeneous in size. Missing schools in this analysis are balanced among treatment arms. The choice of this cohort of students is meant to maintain the same length of exposure to the API program of the main sample of the analysis.[5]

Finally, we conducted a series of in-depth interviews in the spring of 2022 for a small and representative subsample of 16 mentors and 12 community instructors who were part of our study.[6] This qualitative evidence proves useful to complement the quantitative analysis and to shed further light on the mechanisms through which the mentoring intervention affects students and parents as discussed in Section 3.3, as well as on the role of parents as means of scalability as discussed in Section 5.2.

3 Experimental Evidence

In this section, we report OLS estimates from separate regression models for each experiment on the treatment assignment indicators for the *Original* and *Plus* modality after two years of exposure to the mentoring program. All models include the strata indicators that account for the block randomization design (see Section 2.2) as well as few individual characteristics such as students' age and ethnicity. We control for interview week fixed effects, which account for changes in weather and political conditions, as well as indicators for the different teams of enumerators who administered the survey across the communities in our sample. The error terms are clustered at the school level, which represents the unit of randomization of the treatments. We complement the usual asymptotic inference with two alternative procedures. First, we display p-values based on randomization inference, which are accurate

[5]The distribution of missing schools in the analysis of transition to secondary school is 18 schools in the control group, 14 in the *Original* group and 16 in the *Plus* group. Due to the different individual identifiers, we are not able to match this dataset to the survey data. The estimates reported in Table B-6 document no program effects on grade repetition and attrition, which suggest that conditioning on grade attainment is not problematic in our context.

[6]Appendix A.3 reports more details about these interviews. Tables B-4 and B-5 show that the characteristics of these survey respondents are broadly comparable to those of the mentors and the local instructors in our main sample.

even with a small number of clusters. This may be especially relevant in the context of the first experiment, which had fewer schools per treatment arm than the second experiment. Second, given the large array of hypotheses considered throughout the analysis, we also provide p-values that are adjusted for multiple hypothesis testing across different families of outcomes (List et al., 2019).

3.1 Children's Outcomes

Table 1 and the first row of Table 2 display the impacts of the *Original* modality on children's outcomes, as measured by individual test scores collected two years after the introduction of the mentoring program in each experiment, respectively. For the first experiment, the outcome variables shown in Table 1 are based on administrative records of sixth graders in a national standardized test. For the second experiment, we collect our own measures of cognitive and socio-emotional skills (first to fourth columns of Table 2), as the national standardized test was terminated in 2014.[7]

In spite of the differences in measurement of the outcome variable, the separate analyses of the two experiments show consistently inconclusive evidence regarding the effectiveness of the *Original* modality of the mentoring intervention. Depending on the outcome, the effect of the program in the first experiment ranges from positive to negative and is not statistically different from zero. The effect size of the estimated treatment effect on the overall index for student achievement (column 4 of Table 1)—a Generalized Least Squares (GLS)-weighted average across the three subject tests that increases the power of the analysis (O'Brien, 1984)—is negative, small and imprecise.[8] Effect sizes are consistently positive and slightly more precise in the second experiment, although none of the estimated coefficients gets close to the conventional significance levels. The impact on the GLS-weighted overall index for student achievement across the two cognitive measures and the socio-emotional score is 0.12

[7] Another national standardized test was administered by the National Institute for the Evaluation of Education (INEE) starting in 2015, the PLANEA National Plan for Learning Evaluation. While the national test that we employ in the first experiment (ENLACE) was administered to all Mexican students in grades three through six through the year 2013 (see Appendix A.1), PLANEA scores are collected only on sixth graders in a random sample of students within schools.

[8] The GLS weighting procedure increases efficiency when compared to other summary indices by ensuring that outcomes that are highly correlated with each other receive less weight, while outcomes that are uncorrelated and thus represent new information receive more weight. This procedure is more powerful than other popular tests in the repeated-measures setting. Also, missing outcomes are ignored when creating the GLS-weighted score. Thus this procedure uses all the available data, but it weights outcomes with fewer missing values more heavily.

Table 1: Children's Achievement—First Experiment

	Reading Score	Math Score	Science Score	Overall Index
API Original	-0.053	0.083	-0.082	-0.022
	[0.737]	[0.655]	[0.585]	[0.902]
	{0.750}	{0.669}	{0.591}	{0.910}
	(0.779)	(0.739)	(0.717)	(0.878)
Number of clusters	70	70	70	70
Observations	599	599	599	599

Notes: This table shows OLS estimates and the associated p-values on student outcomes measured after two years of exposure to the mentoring program under the first experiment run by the government. For detailed descriptions of the test scores used in this table, see Appendix A.1. The dependent variables are standardized with respect to their means and the standard deviations in the control group. p-values reported in brackets refer to the conventional asymptotic inference. p-values reported in braces are computed using randomization inference (randomization-t). p-values reported in parentheses are adjusted for testing the null impact of API Original across the five outcomes shown in the tyable through the step-wise procedure described in Romano and Wolf (2005a,b, 2016). All p-values account for clustering at the school level.

standard deviations—a non-negligible effect size that is nonetheless not statistically different from zero. The effect of the *Original* modality of the mentoring program on the transition rates to lower secondary school are shown in the last two columns of Table 2. Estimated effect sizes are noisy and relatively small in magnitudes, ranging between an increase of seven and eight percentage points out of a basis of 62 percent enrollment rate in seventh grade in the control group.

The second row of Table 2 displays the estimated coefficients for the average impact of the *Plus* modality of the API program when compared to the control group. Children who are enrolled in a school that received the *Plus* modality increased their reading scores by 0.32 standard deviations. We can reject the hypothesis of a null effect of API *Plus* at the 99 percent confidence level across all three different inference procedures. Quantitatively, the API *Plus* effect is approximately 2.5 times higher than the effect of the API *Original*. The difference between the two modalities is statistically different from zero at the conventional 95 percent level in two out of three inference procedures.

We find similar patterns when we look at math scores, which show a sizable effect of the *Plus* modality with an estimated treatment effect of 0.24 standard deviations. This effect is precisely estimated, and we can reject the hypothesis that the two treatment arms have the same effect at the 95 percent confidence level in two out of three cases. The API *Plus* program also generates a sizable improvement in the socio-emotional score of 0.2 standard deviations.

Table 2: Children's Achievement and Attainment—Second Experiment

	Survey-Based Test Scores				Enroll 7th Grade (1 = yes)	
	Reading	Math	Socio-emotional	Overall Index		age ≥13
API Original	0.126	0.056	0.071	0.124	0.073	0.081
	[0.104]	[0.455]	[0.418]	[0.187]	[0.255]	[0.519]
	{0.138}	{0.483}	{0.440}	{0.218}	{0.283}	{0.573}
	(0.147)	(0.554)	(0.554)	(0.234)	(0.312)	(0.469)
API Plus	0.315	0.237	0.199	0.366	0.124	0.298
	[0.001]	[0.008]	[0.022]	[0.001]	[0.074]	[0.030]
	{0.001}	{0.012}	{0.030}	{0.001}	{0.084}	{0.053}
	(0.001)	(0.005)	(0.011)	(0.001)	(0.032)	(0.032)
API Original = API Plus	[0.043]	[0.043]	[0.178]	[0.020]	[0.469]	[0.134]
	{0.086}	{0.115}	{0.225}	{0.024}	{0.570}	{0.229}
	(0.045)	(0.045)	(0.098)	(0.023)	(0.376)	(0.157)
Number of clusters	224	224	224	224	182	76
Observations	1044	1044	1045	1045	468	106

Notes: This table shows OLS estimates and the associated p-values on student outcomes measured after two academic years of exposure to the API program under the second experiment designed and implemented by the authors in collaboration with the government. For detailed descriptions of the test scores used in this table, see Appendix A.2. The dependent variables in the first four columns are standardized with respect to their means and the standard deviations in the control group. The dependent variable in the last two columns is computed from administrative school records (see Appendix A.1). p-values reported in brackets refer to the conventional asymptotic inference. p-values reported in braces are computed using randomization inference (randomization-t). p-values reported in parentheses are adjusted for testing each null hypothesis (null impact of API Original, API Plus, and the comparison) for the two different families of outcomes (survey-based and administrative data) through the stepwise procedure described in Romano and Wolf (2005a,b, 2016). All p-values account for clustering at the school level.

While the difference with respect to the *Original* modality is not statistically significant, the larger effect of the *Plus* modality is consistent with qualitative evidence documenting that mentors with enhanced training shared more effective strategies to best deal with children's emotions during the bimonthly peer-to-peer sessions (see Appendix A.3). The effect size of the *Plus* modality on the GLS-weighted index of achievement is very large, 0.37 standard deviations—precisely estimated, and statistically different at the 95 percent level from the effect of the *Original* modality.[9]

The last two columns in Table 2 report the estimated effects on the average transition rate to secondary school. Less than two-thirds of the sixth graders in the control group enroll in seventh grade, while the corresponding national average is 95 percent. The *Plus* modality

[9]In Table B-7 we report the results by sub-domains of the reading scores (panel A), math scores (panel B). While the estimates are erratic and not statistically significant for the *Original* modality, the *Plus* modality is shown to increase students' proficiency in reading across various domains (familiar-word reading, reading comprehension, and dictation). There are no improvements in sound-related questions (initial sound and initial name), which is probably due to the fact that children whose first mother tongue is an indigenous language might struggle to capture Spanish alphabet pronunciation. For math scores, the *Plus* modality seems particularly effective on numbers' identification and discrimination as well as additions. There are no improvements in more involved tasks such as problem solving and shape recognition. Similarly, in Table B-8 we report the effects of the two program modalities for each individual component of the socio-emotional score.

increases the probability of a child's enrolling in seventh grade by 13 percentage points. Although marginally significant, the effect is quantitatively sizable, as it represents a 20 percent increase in the share of students who transit to secondary school relative to the mean in the control group. This effect more than doubles in size when we focus on the subsample of over-aged sixth graders (13 years old or more, sixth column), and it persists one year after the second experiment (see Figure B-1). Given the prevalence of child labor in Chiapas, this result for older children is particularly important in terms of life-cycle opportunities. While the estimated coefficient of the *Plus* modality in the last column of Table 2 is significant at the 95 percent level for two out of three cases, the p-values reported in the third row show that we cannot reject that it is different from the corresponding estimate of the *Original* modality.

3.2 Parental Investment and Behavior

Home visits are a key component of the mentoring intervention under study. The goal of these home visits and repeated family/home-visitor interactions is to increase parental awareness about their children's educational trajectories. Table 3 presents the average impact of the program on GLS-weighted indices of parental behavior and investment in their children's education (see Appendix A.2). Panel A displays the estimates of the *Original* modality in the first experiment, while Panel B shows the corresponding figures for both the *Original* and *Plus* modality in the second experiment. Under the *Original* program, consistently across experiments, the estimates are not statistically different from zero, with signs of the coefficients that range from positive to negative and effect sizes on the overall index of -0.03 and 0.1 standard deviations. Instead, parents appear to be systematically more invested in their children's education activities under the *Plus* modality of the program. The estimates reported in the second row of Panel B document that mentors with enhanced training are more effective in boosting parental engagement, both toward the school and directly with the child. The point estimates are positive throughout; three out of four coefficients are statistically significant at the 95 percent level with a very large effect size for the overall index of parenting practices of 0.36 standard deviations. While we can reject the null hypothesis of equal treatment effects on most parental outcomes shown in Table 3, we refer the reader to Section 4.1 for a more thorough discussion on hypothesis testing under both experiments

Table 3: Parental Investment and Behavior

	Engage at School	Manage School Resources	Engage With Child	Overall Index
		Panel A: First Experiment		
API Original	0.198	-0.135	0.149	0.101
	[0.259]	[0.415]	[0.399]	[0.580]
	{0.261}	{0.422}	{0.399}	{0.578}
	(0.338)	(0.511)	(0.511)	(0.511)
Number of clusters	73	73	73	73
Observations	208	208	208	208
		Panel B: Second Experiment		
API Original	-0.188	-0.124	0.167	-0.034
	[0.049]	[0.176]	[0.015]	[0.684]
	{0.058}	{0.197}	{0.015}	{0.630}
	(0.067)	(0.205)	(0.021)	(0.704)
API Plus	0.217	0.087	0.353	0.359
	[0.034]	[0.344]	[0.001]	[0.001]
	{0.037}	{0.247}	{0.001}	{0.001}
	(0.055)	(0.388)	(0.001)	(0.002)
API Original = API Plus	[0.001]	[0.056]	[0.029]	[0.001]
	{0.001}	{0.056}	{0.158}	{0.001}
	(0.002)	(0.036)	(0.036)	(0.001)
Number of clusters	224	224	224	224
Observations	1045	1045	1045	1045

Notes: This table shows OLS estimates and the associated p-values on survey-based measures of parental behavior measured after two years of exposure to the API program. Panel A refers to the first experiment run by the government. Panel B refers to the second experiment designed and implemented by the authors in collaboration with the government. For detailed descriptions of the individual components of the summary measures of parental engagement used in this table, see Appendix A.2. p-values reported in brackets refer to the conventional asymptotic inference. p-values reported in braces are computed using randomization inference (randomization-t). p-values reported in parentheses are adjusted for testing each null hypothesis (null impact of API *Original*, API *Plus*, and the comparison) for the two different families of outcomes through the stepwise procedure described in Romano and Wolf (2005a,b, 2016). All p-values account for clustering at the school level.

(see also Table 5).[10]

Overall, the results presented in these two sections show that the API intervention had differential impacts according to the training received by the mentors. While the *Original* modality does not significantly boost any of the outcomes of interest across two independently run field experiments, the *Plus* modality is shown to generate sizable average effects on children's cognitive and socio-emotional scores, on schooling attainment, as well as on parental engagement toward their children's education. In the next section, we leverage de-

[10]We also estimate the impacts of both the *Original* and *Plus* modalities for each of the individual measures of the parental behavior collected in the survey that have been aggregated in the summary measures displayed in Table 3. Table B-9 reports he results, which are broadly comparable to the estimates discussed in the text. They show large and significant effects for the *Plus* modality on food donations to the instructors, the management of the school resources, help with homework, enrolling their children in extra-curricular activities, expecting their children to complete secondary education or more, and meet periodically with the instructor.

tailed survey information to provide direct evidence on the possible channels through which the API *Plus* program enhance children's outcomes.

3.3 *Plus* vs. *Original*: Channels

We start by evaluating the role of the remedial education sessions. The estimates displayed in Table B-10 suggest that there is no differential effect across the four children's outcomes (p-values = 0.766, 0.675, 0.639, and 0.937) in the relative impact of the two training modalities between children who are more or less likely to be eligible for the remedial sessions (see also Figure B-2). Although the design of the second experiment does not allow us to directly isolate the direct effect of the remedial education sessions within each API modality, this evidence suggests that such mediating factor is unlikely to explain the differential impact between the *Plus* and the *Original* documented in Table 2.[11]

We next consider the role of the pedagogical practices of the community instructors. Table B-11 reports estimates of the effect of the two API modalities using data at the instructor-school level (the sample average number of instructors per school is 1.2 in the school year prior to the start of the second experiment) on four summary measures of pedagogical practices based on GLS-weighted indices across an array of instructor-student interactions (for details, see Appendix A.2). The results show erratic patterns of positive and negative signs with no statistically significant effects of either API modality. The overall index of pedagogical practices reveals a non-negligible negative effect of 0.18 standard deviations for the *Plus* modality, indicating, if anything, a crowding-out effect of the presence of the mentors on instructors' job effort. The effects are quantitatively and statistically similar between API modalities across the different pedagogical practices, as shown in the third row of Table B-11.

Finally, we study the role of the mentor/parent interactions during the home visits as a potential mechanism behind the large and positive effect of the *Plus* modality. Panel A in Table 4 displays the estimated differences across the two API modalities on selected survey variables when 591 parents were asked about the frequency and content of their interactions with the mentors over a period of two months prior to the survey. The number of observations varies across the columns in Panel A due to some of the parents not responding to the survey questions. Missing values for each outcome are balanced with respect to the assignment of

[11] The correlation between the school-level rankings as implied by the average diagnostic test and the math and reading scores is 0.51 and 0.52, respectively. In the absence of randomization across the different components of the intervention within each modality, the direct effect of the remedial education sessions cannot be separately identified from heterogeneous treatment effects by academic achievement.

the *Plus* modality (p-values = 0.746, 0.183, 0.442, 0.517, 0.539, and 0.575). In spite of quite noisy estimates due to the sample attrition and the reduced sample size—parents in the control group cannot be part of this analysis by design—the evidence does show a systematic pattern. Over a two-month period, mentors in the *Plus* modality met one time more with parents at school and 0.7 times more at home compared to those in the *Original* modality (sample means in the *Original* group are five and three, respectively). The GLS-weighted index shown in the third column documents that the quantity of parent-mentor interactions increased by 0.36 standard deviations under the *Plus* modality, which is significant at the 10 percent level. Columns 4 and 5 of Panel A show marginally significant estimates on two measures of the quality of the interactions between parents and the mentors: (i) an indicator variable for whether the mentors have informed parents about their children's learning difficulties, (ii) and whether the mentors provide concrete advice to the parent on how to tackle these difficulties. The effect sizes are large for both outcomes, implying a 14 percent increase in the probability of informing parents relative to the respective sample means in the *Original* group (70 percent). The estimated coefficient for the GLS-weighted quality index is 0.25 standard deviations, which is significant at the 90–95 percent level depending on the inference procedure.

Panel B in Table 4 shows the effect of the *Plus* modality on different competencies, or "parenting style," that the mentors report to have promoted during their encounters with parents. This information was collected during a follow-up interview at the end of the field experiment. Of a total of 126 mentors between the *Original* and *Plus* modalities, enumerators were able to interview 107 of them. The attrition of survey participation of mentors is unrelated to the treatment assignment (p-value = 0.514). For further details on the survey of mentors, please refer to Appendix A.2. Mentors with enhanced training are more inclined to foster attitudes that are centered on educative parenting styles, such as communicating with the child (first column), as well as learning activities (second column). The overall educative style GLS-weighted index (third column) shows a sizable and significant effect (across the three inference procedures) of the *Plus* modality, with an increase of 0.49 standard deviations in the promotion of educative parenting styles to parents during the home visits. Other aspects of the parent-child relationship that are focused on emotional practices do not seem to systematically vary across the two program modalities.

The evidence presented in this section points toward cross-modality variation in the quality of both the parent/mentor interactions and parent/child interactions as a potential mechanism behind the observed difference in children's outcomes. While we cannot separately quantify

Table 4: The Role of Mentors in Fostering Parental Attitudes—Second Experiment

	Panel A: Parents and Mentors Interactions (as reported by the parents)					
	Quantity (Last 60 Days)			Quality		
	Meetings	Visits	Index	Inform About Child	Advise About Child	Index
API Plus	1.039	0.726	0.362	0.102	0.100	0.251
	[0.147]	[0.125]	[0.062]	[0.057]	[0.034]	[0.040]
	{0.194}	{0.171}	{0.094}	{0.097}	{0.056}	{0.070}
	(0.194)	(0.194)	(0.100)	(0.078)	(0.078)	(0.078)
Observations	482	491	504	354	353	357
Clusters	123	124	124	113	112	113

	Panel B: Parenting Styles that Are Promoted by the Mentors (as reported by the mentors)						
	Educative Style			Emotional Style			
	Communication	Learning	Index	Share Feelings	Self-Knowledge	Manage Transitions	Index
API Plus	0.178	0.168	0.494	0.049	0.030	0.142	0.194
	[0.038]	[0.077]	[0.018]	[0.627]	[0.756]	[0.123]	[0.312]
	{0.043}	{0.091}	{0.029}	{0.635}	{0.753}	{0.134}	{0.321}
	(0.074)	(0.075)	(0.043)	(0.843)	(0.843)	(0.308)	(0.558)
Observations	107	107	107	107	107	107	107

Notes: This table shows OLS estimates and the associated p-values of the API *Plus* modality on survey-based measures of interactions between parents and mentors (Panel A) and the different parenting styles that are promoted by the mentors during their interactions with the parents. For a detailed description of the outcome variables used in this table, see Appendix A.2. p-values reported in brackets refer to the conventional asymptotic inference. p-values reported in braces are computed using randomization inference (randomization-t). All p-values account for clustering at the school level. p-values reported in parentheses are adjusted for testing the effect of API *Plus* for the different families of outcomes (quantity and quality of interactions, parenting styles) through the stepwise procedure described in Romano and Wolf (2005a,b, 2016).

the relative contribution of each additional training module, the increase in the quality of the home visits is likely to originate from the mentors' peer-to-peer sessions, which had the exact role of helping the mentors with enhanced training to communicate more effectively with parents. As mentioned in Section 2.2, these workshops enable interactions and information sharing among the participants, while the extra week of initial training is instead focused on pedagogical practices targeted to children at school. Qualitative evidence seems indeed to corroborate this hypothesis, as summarized by the following quotes from mentors who have participated in the peer-to-peer meetings (see Appendix A.3 for more details):

- *"During the workshops I was told that I should be able to adapt to the context of the community and understand the local living arrangements in order to establish a dialog with the parents without modifying what they conceive as their environment."*
- *"It was recommended that we pay frequent home visits so as to establish a relationship with the parents and gain their trust."*
- *"[The workshops] exposed us to effective strategies of other mentors [for*

dealing with parents] that we could try and implement in our community."

4 Threats to Scalability

Over the summer of 2016, after learning about the results of the second experiment, the government decided to replace the *Original* program with the enhanced training modality. All its primary schools, including the 224 schools that were part of the evaluation sample, were deemed eligible to receive the *Plus* program modality. The overall scale of the operation of the mentoring intervention—including the total number of mentors that participated in the program—remained constant in the periods before, during, and after the experiments. This single policy change creates two interesting circumstances that are informative for our case study on scaling. On the one hand, schools that were part of our second experiment experienced a change in the situation—from the research setting to the government implementation. On the other hand, the rest of the schools, that were not part of the experiment but received the mentoring program under the *Original* modality, underwent a reform in program design within the same government situation. In this section and the next one, we focus on the sample of the experimental schools in order to zoom into the threats and mechanisms of scaling brought about by the new situation. In Section 6, we discuss policy impacts on education outcomes for experimental schools as well as for the overall population of schools in the state of Chiapas.

To study the scale-up problem in our context, we analyze three key aspects outlined in Al-Ubaydli et al. (2020), namely inference, representativeness of the population, and representativeness of the situation. Although it will not be part of our discussion, we also want to mention additional features of the experimental design that may speak to other threats to scaling. First, the relatively large units of randomization (school community) are robust to local general equilibrium/spillover effects, which are often relevant in field experiments that are implemented at scale. Second, our field experiment has been implemented by the research team in close collaboration with the government agency that was later in charge of its policy implementation. Hence, the design of the two program modalities bears in mind the supply-side considerations of scaling, as well as various financial and local institutional constraints.

Table 5: Joint Test of Significance Within and Across Experiments (p-values)

	First Experiment	Second Experiment	Both Experiments
Api Original = Control	0.828	0.411	0.707
Api Plus = Control	.	0.001	.
API Original = API Plus	0.114	0.001	0.002

Notes: This table reports randomization-inference (Randomization-t) p-values for the omnibus test of overall experimental significance of each separate hypothesis (Young, 2019). An asymptotic p-value is reported for the hypothesis that API Original = API Plus in the first column, which is tested across experiments. All p-values account for clustering at the school level.

4.1 Inference

Inference problems arise when researchers and practitioners want to learn to what extent existing evidence advocates for policy decisions. We focus on whether (i) the lack of effectiveness of the *Original* modality is indicative of a null result, and (ii) the large impact of the *Plus* modality on schooling outcomes for children and engagement outcomes for parents is a false positive." We jointly consider two key outcomes—the overall index of student achievement and the overall index of parental engagement—and compute p-values of overall statistical significance (Westfall and Young, 1993). Following the insights in Maniadis et al. (2014), we harness the value of the two experiments to bolster the credibility of our empirical evidence. We test hypotheses across experiments using Fisher's combined probability test, which is akin to the joint statistical significance test usually invoked in meta-analyses.[12]

Table 5 shows the results. Consistently within and across experiments, the *Original* modality does not generate actionable evidence (and yet, the government implemented such program modality at scale). The evidence displayed in the first row documents a lack of significance of such variant of the mentoring program on children's achievements and parental investment. The results in the second and third rows of Table 5 document a highly significant impact of the mentoring program under the *Plus* modality, both when compared to the control group with no mentors (p-value = 0.001) and the *Original* modality. The relatively noisier estimates of the first experiment do not allow us to reject the hypothesis of equal treatment effects between the two modalities when tested across experiments (p-value = 0.114). This result highlights the importance of the design of the second experiment, in which we replicate the

[12]The Westfall-Young procedure uses the joint distribution of p-values across all equations so as to minimize the loss of power brought about by the multiple testing adjustment within a given experiment. Combined p-values across experiments are obtained using Fisher's formula: $-2\sum_{i=1}^{k} log(p_i) \sim \chi^2_{2k}$, where $p_i \sim U[0,1]$ is the p-value for the i^{th} hypothesis test and $k = 2$ is the number of independent experiments being combined.

Original modality along with the new *Plus* modality. In the second experiment, we strongly reject that the *Plus* modality is equally effective to the *Original* modality (p-value = 0.001). A very similar result holds through a combined probability test across both experiments (p-value = 0.002), which is reported in the third column of Table 5.

The joint inference drawn from the two experiments seems to convincingly point toward the relative effectiveness of the *Plus* modality when compared to both the *Original* modality and the control group with no mentors. Given that the policy reform under study is a change from the *Original* to the *Plus* modality, we discard the first experiment in the rest of this section and focus our analysis on the schools that participated in the second experiment.

4.2 Representativeness of the Population

Research findings from field experiments may sometimes be difficult to generalize because, in the language of Al-Ubaydli et al. (2020), the properties of the study population may differ from the population of interest to policy makers. Heckman (1992) discusses selection into field experiments and finds that the characteristics of subjects who participate can be distinctly different from those of subjects who do not participate. In Table 6 we compare means in observable characteristics between our experimental sample and the overall population of schools in the state of Chiapas. The descriptive statistics for the sample of experimental schools are shown in column two, and they appear remarkably balanced when compared to the respective statistics in the overall population that are displayed in the first column. As shown in the third column, we cannot reject equal means across the several variables assessed. There is a very small imbalance in the number of local instructors, which is only marginally significant.

We next study whether the average impacts of the *Plus* modality in the second experiment have ex-ante external validity with respect to the impact of the mentoring intervention at scale in the broader population of schools in Chiapas. To do this, we evaluate whether program impacts vary along the program eligibility criteria used by the government during the policy implementation (see Section 2.1). The idea behind this exercise is that any variation in treatment effects along those dimensions may be indicative of the extent to which program effects may change because of the underlying differences across populations. Table 7 displays heterogeneous treatment effects of both *Original* and *Plus* modalities along two criteria that are time invariant and hence plausibly unaffected by the intervention: whether the community where the school is located is categorized as having high or very

Table 6: Differences Across Populations

	All Chiapas Mean (SD)	Second Experiment Mean (SD)	Chiapas vs. Second Experiment p-value
		Panel A: Community Characteristics	
Number of households	34.625	29.329	0.486
	(109.863)	(50.234)	
Total population	140.293	121.389	0.494
	(394.371)	(240.562)	
Share economically active	0.298	0.303	0.361
	(0.073)	(0.070)	
Water connection (Y/N)	0.033	0.023	0.454
	(0.178)	(0.151)	
Sewer system (Y/N)	0.018	0.009	0.346
	(0.134)	(0.096)	
Share of illiterates	0.268	0.270	0.832
	(0.175)	(0.167)	
Share of dwellings with dirt floor	0.328	0.363	0.126
	(0.319)	(0.322)	
Garbage collection (Y/N)	0.025	0.023	0.842
	(0.158)	(0.151)	
		Panel B: School Characteristics	
Average test score (Spanish) 2010	425.173	431.340	0.158
	(57.245)	(60.810)	
Average test score (Math)	415.998	421.333	0.363
	(76.967)	(80.895)	
Number of students	14.023	14.770	0.205
	(8.403)	(7.069)	
Number of local instructors	1.216	1.279	0.054
	(0.449)	(0.514)	
Share students over-age	3.125	3.620	0.264
	(6.285)	(5.629)	
Observations	1,475	230	

Notes: Means and standard deviations in parentheses for various characteristics collected before the introduction of the API program. The last column shows asymptotic p-values for mean differences between the overall population and the experimental sample. Panel A shows community-level characteristics from the population census (2010), whereas Panel B displays school-level variables from the school census (2010). See Appendix A.1 for more details on the data sources.

high "marginality," as defined according to the National Population Council or CONAPO (Poverty 1), and whether the community was targeted by an anti-poverty program (Poverty 2).[13] In the sample of schools in the second experiment, approximately one-third satisfy the Poverty 1 criterion, 70 percent satisfy the Poverty 2 criterion, and 25 percent satisfy both criteria. We run separate regression models for three summary outcomes of the intervention on both students and parents: the overall index of student achievement, the indicator for enrollment in seventh grade, and the overall index of parental engagement. Estimation

[13]For details on the Poverty 1 index, refer to https://www.gob.mx/cms/uploads/attachment/file/685308/Nota_t_cnica_IML_2020.pdf, accessed on August 2022.

Table 7: Heterogeneity in the Impact of the Program by Eligibility Criteria

	Children's Outcomes		Parental Outcome
	Overall Score	Enrolled Secondary	Engagement Index
API Original	0.088	0.134	-0.018
	[0.419]	[0.205]	[0.887]
API Original× Poverty 1	0.090	-0.104	-0.046
	[0.637]	[0.420]	[0.816]
API Original× Poverty 2	0.091	-0.033	-0.009
	[0.490]	[0.767]	[0.949]
API Plus	0.320	0.273	0.464
	[0.047]	[0.024]	[0.000]
API Plus× Poverty 1	0.094	0.015	0.100
	[0.638]	[0.902]	[0.643]
API Plus× Poverty 2	0.010	-0.216	-0.173
	[0.963]	[0.128]	[0.386]
Original(Pov. 1)=Original(Pov. 2)	[0.995]	[0.681]	[0.873]
Plus(Pov. 1)=Plus(Pov. 2)	[0.816]	[0.297]	[0.444]
Number of clusters	224	182	224
Observations	1045	468	1045

Notes: This table shows OLS estimates and the associated p-values (in brackets) on student and parental outcomes measured after two academic years of exposure to the API program under the second experiment designed and implemented by the authors in collaboration with the government. For a detailed descriptions of the test scores used in this table, see Appendix A.2. The dependent variables in the first and third columns are standardized with respect to their means and the standard deviations in the control group. The dependent variable in the second column is computed from administrative school records (see Appendix A.1). All p-values account for clustering at the school level.

results reveal limited variation in program impacts along both poverty measures, with effect sizes for the interaction terms with the indicator variables for the program modalities that are not statistically different from zero, and not statistically different from each other.

Taken together, the evidence shown in this section documents that differences across populations (if any) are unlikely to represent a meaningful threat to scalability in this context. There is a very high degree of similarity in observable characteristics between the experimental sample and the overall population of schools in Chiapas. Program impacts are also not heterogeneous along the determinants of the rollout of the program under the government implementation, which is indicative of the fact that the community/school targeting process is unlikely to play a role for the scalability of the program.

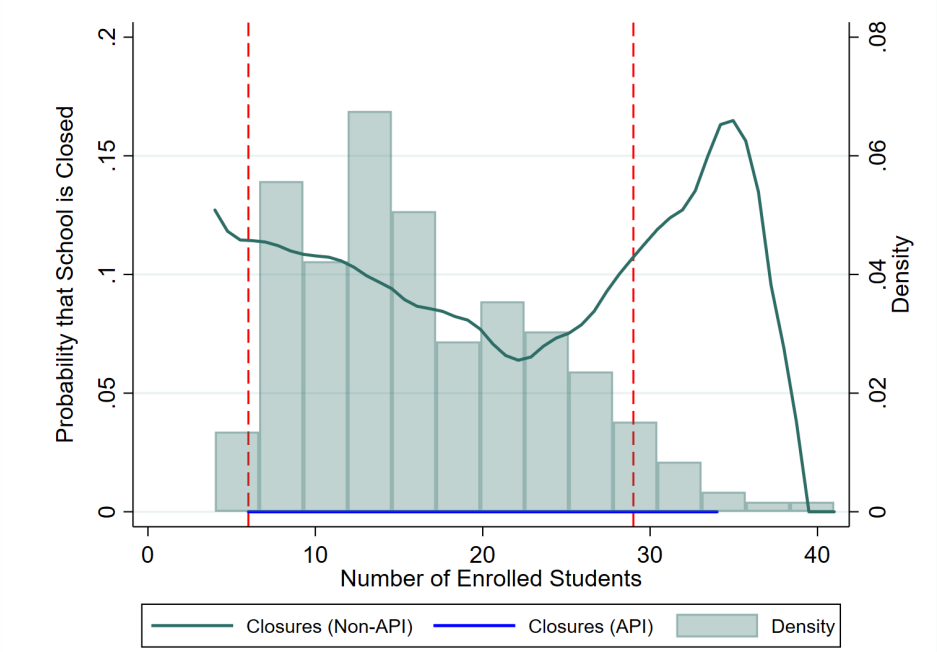
4.3 Possible Threats from the New Situation

Notwithstanding sizable and significant effects on a sample that is representative of the population of interest, the success of the mentoring intervention at scale is not guaranteed. The difference between the implementation protocol in our research setting—where we had an active role in guaranteeing the smooth progress of the field experiment—and the government operations can translate into contrasting aftermaths of the program. For example, the criteria for closing schools were very different in the two situations. Although the official enrollment threshold for closing schools is six students, schools in the second experiment were allowed to remain open if they had at least three enrolled students in either of the two school years when the experiment took place. In addition, children in schools with more than 29 enrolled students were required to transfer to schools in the regular public school system. As a result, only two schools closed in the sample of 230 schools in the second experiment (see Section 2.2).[14]

Hence, a threat to the success of the program in the new policy situation may come from the possible school closures as a result of the mentoring intervention under the government implementation. If instead the presence of a mentor increases the probability that schools remain open, this mechanism could represent an opportunity to learn something about the mechanisms behind the scalability of the program in our context. Figure 1 shows the relationship between school closures, as measured by the year-to-year presence in the school census two years after the second experiment, and school size, as measured by the number of enrolled students in the school year before the second experiment. The green (lighter) line shows how the probability of school closures varies by school size for schools that did not receive the program during the government intervention. The probability of school closure is bimodal, and its two modes are positioned around the two critical enrollment thresholds.

[14]While school closures are obviously a "non-negotiable" aspect (List, 2022) for the success of the API program, there may be other "negotiable" differences in the program implementation across the experimental and the policy settings that we cannot directly study due to a lack of monitoring data outside of the experimental sample/period. First, to avoid refusal of the assigned mentor among the communities of the evaluation schools, each mentor in the experimental sample was provided with two baskets of food, throughout the school year, as donations to the community leaders as well as for personal consumption. Second, as a way to attenuate the potentially detrimental consequences of mentors' dropping out of the program during the evaluation period, the government delegates in Chiapas arranged for a replacement within two weeks from the day of a mentor's departure from a community. If the dropout was part of the *Plus* group, the replacement would receive an additional three-day training session that would make up for the content covered during the extra week of the initial training session. Third, there might be differences between the experiment and the policy rollout in the implementation of the training module of the *Plus* modality, such as the number of extra days and the content of the curriculum of the initial training as well as the frequency of the peer-to-peer sessions.

Figure 1: School Closures and School Size



Notes: The figure shows a histogram with the distribution of the size of the 224 schools that participated in the second experiment (as measured by the number of enrolled students during the first two years of the government implementation). Overlaid on the histogram, it displays kernel-weighted local mean estimates of the relationship between the probability of closure for schools that receive (blue, darker line) and do not receive a mentor (green, lighter line) during the first two years of the government implementation of the *Plus* modality. The red vertical dashed lines represent the statutory enrollment thresholds for school closures in rural Chiapas.

School closures occur also for medium-sized schools, consisting of six and 29 students. In total, twelve schools out the 122 schools without mentors closed during the first two years of the government implementation of the program under the *Plus* modality. Over the same time period, none of the 102 schools that received a mentor during the government implementation closed (blue, darker line).

5 Pathways to Scale

What plausible mechanism can explain the strong and positive correlation between the mentors' presence and the probability that schools remain open? Parents organize local associations aimed at promoting community education, to which they contribute by maintaining the school's facilities and distributing school materials. The parents' association also plays a role in the decision to keep the school open as well as whether to require children enrolled

in schools with more than 29 students to transfer to schools that are part of the regular public school system. Given the evidence on the effect of the mentoring intervention on parental engagement (see Section 3), in this section we analyze the role of parents, and more broadly, of the community-level parental engagement, as a potential mechanism behind the scalability of the program. We first lay out a simple model of skill formation and parental investment, where the individual incentives for parents to invest in educational activities are jointly determined at the community level. The model provides us with a framework to study the threat to scalability from changes in situations. We then document empirical evidence that is consistent with the key model predictions and that is difficult to reconcile with alternative, more direct, channels of influence of the mentors on school closures.

5.1 Theory

There are various local communities (c), each composed of a number of families N_c. Each family $i \in \{1, \ldots, N_c\}$ decides whether to engage in parenting, $I_i \in \{0, 1\}$. The returns to parenting are defined by the following technology of skill formation:

$$\theta_i = \theta_{i,0} + A \cdot \left[I_i^{\phi} + \left(\frac{1}{N^c - 1} \sum_{j \neq i} I_j \right)^{\phi} \right]^{\frac{1}{\phi}}, \tag{1}$$

where θ_i is a child's skills, while $\frac{1}{N_c - 1} \sum_{j \neq i} I_j$ represents the average parental engagement among other parents in the same community. The parameter ϕ characterizes the degree of substitutability between individual parental engagement and community-level parental engagement in the process of child development. Parameter A is total factor productivity, while $\theta_{i,0}$ represents the child's initial skills.[15]

We model parents as paternalistic over their children's skills, and we assume that there is a cost of parenting so that the parental utility function is $\mu(I_i) : U_i = -\mu(I_i) + \theta_i$. The parental decision problem is defined as follows:

$$\max_{I_i \in \{0,1\}} U_i(I_i, I_{-i}) := -\mu(I_i) + \theta_{i,0} + A \cdot \left[I_i^{\phi} + \left(\frac{1}{N_c - 1} \sum_{j \neq i} I_j \right)^{\phi} \right]^{\frac{1}{\phi}}.$$

[15] We omit in equation (1) the shares of the parental investment since the total factor productivity term, A, allows us to appropriately re-scale the constant elasticity of substitution function without loss of generality.

In this framework, the incentives for parents to engage with the education of their children depend upon the community-average parental engagement as well as on the productivity of those investments. An equilibrium in this economy is defined as the optimal choices of families (I_i^*) that are consistent with the endogenously determined community-level engagement: $\left\{I_i^*(I_{-i}^*)\right\}_{i=1}^{N_c}$.

Proposition 1 *This economy exhibits two types of equilibria:*

1. *Free Riding Equilibrium: parents free ride on each other, and in equilibrium there is no community engagement ($I_i^* = 0 \ \forall i$).*
2. *Collaborative Equilibrium: all parents in the community are engaged in the process of children's learning ($I_i^* = 1 \ \forall i$).*

We introduce in this framework an educational intervention $\tau \in \{0, 1\}$—such as the API program—that directly affects both the process of skill formation of children and the incentives of parents as follows:

$$\theta_{i,0}(\tau; \gamma_0) = \theta_{i,0} \cdot (1 - \tau) + (\theta_{i,0} + \gamma_0) \cdot \tau \tag{2}$$

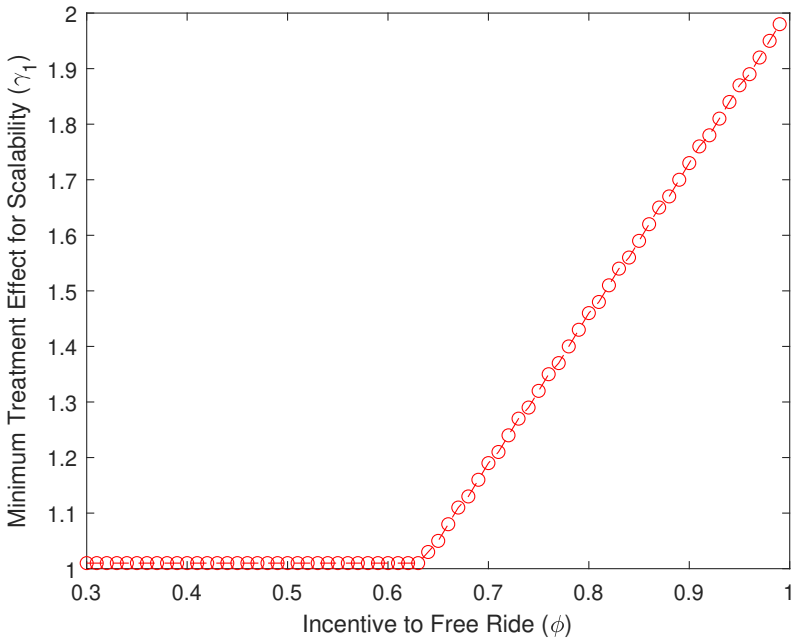
$$A(\tau; \gamma_1) = 1 \cdot (1 - \tau) + \gamma_1 \cdot \tau. \tag{3}$$

The impact of the intervention not only hinges on the direct effect on children's skill γ_0, but also on its ability to shift the incentives of parents through higher productivity of investment (γ_1), which in turn affects the community-level parental engagement. The total effect on children's outcomes $\Delta\theta_i = \theta_i(\tau = 1) - \theta_i(\tau = 0)$ can be written as follows:

$$\Delta\theta_i = \begin{cases} \gamma_0 & \text{in the } \textit{Free Riding Equilibrium} \\ \gamma_0 + \gamma_1 & \text{if the intervention induces a new } \textit{Collaborative Equilibrium}. \end{cases}$$

We define a threat to scalability in this framework as a deterioration of the direct impact of the program. More precisely, for two given situations s and s'—such as the field experiment and the government implementation in our context—we posit that $\gamma_0(s') < \gamma_0(s)$. The extent to which this translates into a deterioration of the overall effectiveness of the program depends on the endogenous response of parents. In one scenario, the change in the situation induces the program to fail at scale because of the lack of parental responses. In the second scenario, the program is able to promote coordination and engagement of parents in the local

Figure 2: Pathways To Scale



Notes: The figure shows the relationship between the incentive to free ride among parents (ϕ) and the program's impact on the productivity of parental investment as predicted by a calibrated version of the model ($\mu(I_i) = -2 \cdot I_i$, $N_c = 1000$).

community, therefore possibly offsetting the threat to scalability with respect to changes in the situations.

The parameter ϕ, which captures the degree of complementarity between parents in the technology of skill formation of children, lies at the core of the equilibrium selection in the model (Free Riding vs. Collaborative, see Proposition 1). This parameter pins down the incentives for parental cooperation in the local educational activities. Low values of ϕ represent high incentive for cooperation, while positive levels of ϕ induce free-riding among parents. Figure 2 shows how the impact of a program at scale across different situations depends upon the degree of parental collaboration (ϕ) and its ability to trigger local spillover effects. The figure shows that in economies with a high degree of parental collaboration (small ϕ), a relatively small impact on the productivity of investment (γ_1) can induce a shift toward the Collaborative Equilibrium. Economies with a lower degree of parental cooperation (high ϕ) need larger impacts of the program on the productivity of parental investment to trigger the social determination of human capital investment at the community level.

5.2 Evidence on Parents as Means of Scalability

We test the model's key prediction using variations across situations (the first experiment was run by the government, while the second experiment was run by the research team) and across program modalities (API *Original* and API *Plus*). In particular, we study whether a different response of parents in terms of their engagement both at the local school and with their children (see Table 3) subsequently triggers a differential impact of the mentoring intervention on school closures. School closures, which in our context represent a major disruption to the program's continuity and effectiveness, should be interpreted as an outcome of the deterioration effect of the program once it is implemented at scale.

The first two columns of Table 8 show the reduced-form effects of the two randomized program modalities—in both the first experiment (first column) and the second experiment (second column)—on the probability that schools close in the second year of the national rollout of both programs. The *Original* modality displays small and noisy effects on school closures in both experiments, which are not statistically different from zero. This result suggests that situations that do not promote parental engagement do not differ from the status quo rates of school closures, which range between 5 percent (first experiment) and 8 percent (second experiment) in the experimental control groups with no mentors.

The second column of Table 8 shows that the *Plus* modality, as Table 3 shows significantly boosts parental engagement, has a significant impact on school closures. Schools that were assigned to the enhanced modality during the second experiment experience are less likely to close permanently (-8.3 percentage points) two years after the *Plus* modality was adopted by the government, which is statistically different from zero at the 95 percent confidence level. This result echoes previous evidence on the relationship between the probability of closures for schools that receive a mentor during the government implementation of the *Plus* modality, which is shown in Figure 1. Given that the probability of receiving a mentor during the government implementation is orthogonal with respect to the randomized API assignment of the second experiment, this evidence rules out channels other than parental engagement through which the presence of the mentors can keep the schools open.[16]

The IV estimates shown in the third column of Table 8 go a step further and quantify the extent to which parental engagement affects the probability of school closures. Because of

[16] Approximately half of the schools in any of the treatment arms and the control group of the second experiment received a mentor by the second year of the national rollout of the *Plus* modality. This share is balanced across treatment arms after controlling for the program eligibility criteria (see Section 2.1): p-value($Original$) $= 0.367$, p-value($Plus$) $= 0.660$.

Table 8: School Closures and Parental Engagement

	Outcome: School Closures		
	First Experiment	Second Experiment	Second Experiment, IV
API Original	0.031	-0.031	-0.031
	[0.549]	[0.396]	[0.410]
API Plus		-0.083	
		[0.030]	
Overall Parental Engagement			-0.217
			[0.021]
Observations	80	224	1045
Clusters	.	.	224
F-Stat (Excl. Instruments)			13.833

Notes: This table reports the estimates for the reduced-form effects of the API modalities during the two experiments (columns 1 and 2) on the probability of school closures, as well as the instrumental variable estimates of the impact of parental engagement on school closures. In the third column, the randomized API *Plus* modality during the second experiment is used as an instrumental variable, while the randomized API *Original* modality is included as a control variable. The dependent variable is an indicator variable for whether the school is closed in the fall of 2014 (column 1) or in the fall of 2018 (columns 2 and 3). The variable "Overall Parental Engagement" is the same variable used in the last column of Table 3. All p-values account for clustering at the school level. p-values reported in brackets account for clustering at the school level.

the contextual information on the role of the parental association in deciding school closures discussed previously in this section, we posit that parents are the main channel through which the *Plus* modality of the API program affects school closures. The differential impacts of the two program modalities on both parental investment and school closures shown in the first two columns of Table 8 are consistent with this exclusion restriction. We find that an increase of half a standard deviation in the overall parental engagement index is causally associated with a reduction of 11 percentage points in the probability that their children experience a school closure. This effect is both statistically and quantitatively significant.

We complement these findings with qualitative evidence on the role of parents in ensuring continuity in schooling activities (see Appendix A.3). As reported by the community instructors, parents may have more at stake in keeping the schools open as they invest in durable goods for the local school:

- *"[Parents] help manage the school and contribute by improving the fencing, painting the walls, fixing the toilets, as well as buying school materials."*
- *"[Parents] serve the needs of the school with construction works and they provide food to the local instructor."*

Table 9: Parental Investment by Proxies of Community-Level Collaboration

	Engage at School	Manage School	Engage With Child	Engagement Index
API Plus× No Conflict	0.161	0.159	0.370	0.364
	[0.176]	[0.103]	[0.000]	[0.001]
API Plus× Conflict	0.926	0.281	0.592	0.914
	[0.000]	[0.183]	[0.005]	[0.000]
Conflict=No Conflict	[0.000]	[0.596]	[0.329]	[0.020]
Mean Control (Conflict)	-0.006	-0.059	0.152	0.011
Mean Control (No Conflict)	-0.023	0.013	-0.013	0.003
Number of clusters	224	224	224	224
Observations	1045	1045	1045	1045

Notes: This table shows OLS estimates on parental outcomes measured after two academic years of exposure to the API program under the second experiment designed and implemented by the authors in collaboration with the government. The variable Conflict takes the value of one if least on hostile event related to land property, religion, elections, crime, or drug addiction is reported at the locality level in the population census (2010). For a detailed descriptions of the variables used in this table, see Appendices A.1–A.2. The dependent variables are standardized with respect to their means and the standard deviations in the control group. All p-values account for clustering at the school level. Asymptotic p-values reported in brackets are clustered at the school level.

As reported by the mentors, parents follow up with their children on homework and other pedagogical material whenever the mentor is busy attending tasks outside of the community:

> "*Parents used to provide support with homework whenever mentors are visiting other communities ensuring pedagogical support, so that upon the return of the mentors they are able to make progress in the schooling activities without set-backs.*"

We next empirically investigate the second prediction of the model, as depicted in Figure 2 and discussed above. We take advantage of information from the 2010 locality-level census on the degree of social hostility in the community (see Appendix A.1), which is based on hostile event related to land property, religion, elections, crime, or drug addiction. This measure proxies the degree of collaboration in the community. In the 224 communities that form part of the second experiment, we construct an indicator variable for the presence of a conflict in the community if at least one of these hostile events is reported. We then interact this variable with the experimental API *Plus* assignment and regress the various survey-based measures of parental engagement collected during the second experiment on these interaction terms.

Table 9 displays the estimation results. In line with the prediction from the model, communities with higher hostility display a higher parental response to the *Plus* program when compared to communities with no conflicts, as parents need to overcome the higher incentive to free ride.[17] The impacts on parental engagement at school are shown in the first column, and it is approximately eight times larger in communities with conflicts than in communities without conflicts. The impacts are twice as large when considering activities related to managing school resources and engaging directly with children in educational activities, although in this case they are not statistically different from the corresponding effects without conflicts (second and third columns). In the last column we look at the overall parental engagement index. Communities with conflicts exhibit impacts on parental behavior (+0.91 standard deviations) that are 2.5 times larger than communities with no conflicts (+0.36 standard deviations)—this difference is statistically significant at the 95 percent confidence level.

Previous literature has highlighted how parental investments and parenting styles are responsive to the environments that families face (Doepke and Zilibotti, 2017; Agostinelli, 2018; Agostinelli et al., 2020). Our results shed light on how the success of an educational program depends upon the local engagement of parents in educational activities and how its scalability is effectively a socially determined outcome.

6 Policy Impacts at Scale

In this last part of the analysis, we discuss the impacts of the government-run program on various educational outcomes. After 2016, the government fully converted the mentoring program into the *Plus* modality, and all schools were in principle eligible to receive the mentors. Schools were assigned a score between one and four, with one denoting the highest priority level. The scores are based on a combination of criteria that includes school performance in the national learning assessment, whether the school has six or more primary students enrolled, whether the school received the API program in the period between 2009 and 2015, the level of marginalization of the community where the school was based, and whether the community was targeted by an anti-poverty program (see Section 2.1). The size of the new program did not change relative to the previous implementation of the *Original* modality, including the number of available mentors. In the fall of 2016 there were 535

[17]Notice that few to no school closures were detected in villages that were part of the *Plus* modality, independently of whether the community experienced any conflict.

mentors in Chiapas, and given these constraints, mentors are allocated across communities on a rotating basis.

6.1 The Exposure Effect of the Program

We start by analyzing the initial transition of the schools in the evaluation sample from the experimental situation to the policy at scale. We do this by estimating the following regression model:

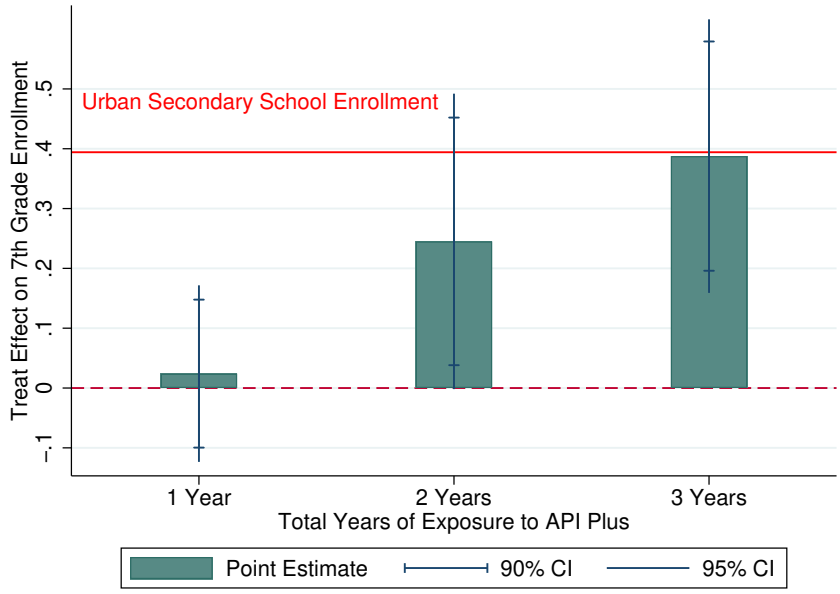
$$Y_j = \beta_0 + \sum_{k=1}^{K} \beta_{1,k} \mathbb{1}\{ExpPlus_j = k\} + \boldsymbol{\gamma}' \boldsymbol{X}_j^{Criteria} + u_j \,, \tag{4}$$

where $\mathbb{1}\{ExpPlus_{j(i)} = k\} \in \{0, 1\}$ represents an indicator variable for whether school j is exposed to $k \in \{0, 1, \ldots, K\}$ years of API *Plus* program, while $\boldsymbol{X}_{j(i)}^{Criteria}$ is a vector of indicator variables for the program eligibility criteria. Our outcome of interest (Y_j) represents the 2016–2017 school-level transition rates to seventh grade. This variable is constructed from the same administrative source we used for our experimental evaluation. The sample includes 207 schools of the 224 that were part of the experiment. Beyond a school that permanently closed, the sample attrition is caused by schools not having sixth graders during that school year. This fact is consistent with the multi-grade nature of the CONAFE system. Attrition is balanced with respect to the total years of exposure to API *Plus* (p-values = 0.467, 0.812, and 0.568).

We exploit the change in situation to analyze the exposure effect to the program one year after the government rollout, when school closures were still minimal (only one school closed among our 224 schools by the fall of 2017). After the first year of the rollout, schools started to close more intensively (11 additional schools closed during the second year); hence, outcome variables based on survey or administrative school data during this period would suffer from endogenous censoring.[18] By the spring of 2017, the total years of exposure to API *Plus* range from zero to three depending on whether a school received the *Plus* modality for two years during the second experiment, as well as on the API *Plus* government assignment in the year after the experiment. The underlying assumption of this approach is that, once controlling for the official government criteria, the remaining variation in the program's

[18]For this reason, in Section 6.2 we take advantage of the 2019 Census information that is not subject to this issue. The Mexican Census is available only every ten years, and hence we cannot use this information to estimate the model in equation (4).

Figure 3: The New Situation in the Experimental Sample of Schools



Notes: This figure shows OLS estimates of the years of exposure to the mentoring program on the probability of enrolling in seventh grade during the transition from the second experiment to the government implementation of the *Plus* modality. Vertical lines overlaid on each bar display the 95 percent and 90 percent confidence intervals, respectively. Confidence intervals are based on asymptotic inference.

assignment across localities in the year after the experiment is as good as random:

$$E[\boldsymbol{u}|\boldsymbol{X}_j^{Criteria}, ExpPlus_j = k] = E[\boldsymbol{u}|\boldsymbol{X}_j^{Criteria}],\ \forall k \in \{0,1,2,3\}\,. \tag{5}$$

While this assumption is obviously not testable, we examine whether the API *Plus* assignment is conditionally balanced with respect to predetermined educational outcomes. Table B-12 shows the results of this placebo test. The outcomes we use are the 2013 scores in the national standardized test (Spanish, math, and science, see Appendix A.1).[19] The estimated coefficients of the years of exposure to API *Plus* are not statistically different from zero. The point estimates are relatively small, especially after controlling for the assignment criteria, and their signs do not suggest any pre-existing positive trend. For all these reasons, we believe this evidence provides some support to the plausibility of (5) in our setting.

Figure 3 plots the estimated $\beta_{1,k}$ coefficients shown in equation (4), where zero years of exposure represents the reference category. The results show a positive exposure effect to the program one year after the government rollout. The effect of the *Plus* modality goes from 3 percentage points after one year to more than 35 percentage points after three years

[19]The year 2013 is the last year in which the national standardized test was applied universally in Mexico.

of exposure. The average effect of three years of exposure on the probability of enrolling in seventh grade is very large and precisely estimated (p-value < 0.001). The magnitude of this effect implies that the enrollment rates in these disadvantaged and rural areas achieve the secondary school enrollment rates in urban Mexico (95 percent). The average marginal effect of an extra year of the program is +10 percentage points (p-value = 0.006) in the probability of enrolling in seventh grade. The effect of two years of exposure, although not statistically different, is larger than the experimental estimate displayed in Table 2 (fifth column) suggesting that the impact does not fade out after one year.

6.2 The Effect of the Program Beyond the Evaluation Sample

We now broaden our analysis to the entire population of schools in the state of Chiapas. We examine whether the *Plus* modality of the program at scale has promoted educational opportunities for children in these disadvantaged communities, possibly resembling the results from the second field experiment (see Section 3). To do so, we match administrative records on the government rollout of the program during the fall of 2017 with village-level educational outcomes from the population census data (data collection in the fall of 2019) for the majority of the schools and communities, which include those that were part of the second experiment.[20] Our first outcome is the village-level lower-secondary enrollment rates among children between 12 and 14 years old. This variable is available in the census for 1,417 communities in Chiapas. It is not immediately comparable with our previous measure of enrollment in seventh grade for two reasons. First, the census-based information represents the stock (rates) of children enrolled in secondary school in a given year, while our previous measure represents the flow of new students enrolling in secondary schools. Second, the census-based variable includes children in the village who are enrolled in primary schools that are not eligible for the API program, which converts the analysis of the program at scale into an intent-to-treat analysis. Another educational outcome from the population census that we use is the rate of child literacy for children between eight and fourteen years old, which represents an available measure of children's achievement. This variable is available

[20]The match between the universe of schools and the localities of the population Census is one to one, as each village has at most only one primary school. The coverage of the census data is not universal, and we were not able to match nearly one-third of the 2,063 schools in Chiapas in 2019. For both educational outcomes in the census data we cannot reject the hypothesis that the probability of missing observations is balanced with respect to the program assignment (p-values for secondary school enrollment and child literacy are 0.728 and 0.430, respectively). For further details on the census sampling design, please refer to: https://www.inegi.org.mx/contenidos/productos/prod_serv/contenidos/espanol/bvinegi/productos/nueva_estruc/702825197629.pdf, accessed on August, 2022.

in the census for 1,440 communities in Chiapas. Finally, we match information about school closures from the school census in the same year (2019).

This set of educational outcomes is particularly conducive for our analysis. Secondary school is a critical period for the educational outcomes of the disadvantaged population under study, as more then a quarter of the 12 to 14 year olds in Chiapas are out of school. Likewise, 13 percent of school-aged children are still illiterate. The year of the data collection in the census (2019, two full school years after the rollout of the *Plus* modality at scale) is consistent with the length of exposure to the API program in the second experiment. These census-based outcomes cover the quasi-universe of the localities in Mexico and, unlike other survey-based or administrative test score measures, they are not subject to any censoring during the data collection due to school closures. This allows us to avoid the concerns about selection bias due to differential school closures induced by the program at scale (see Section 4.3).

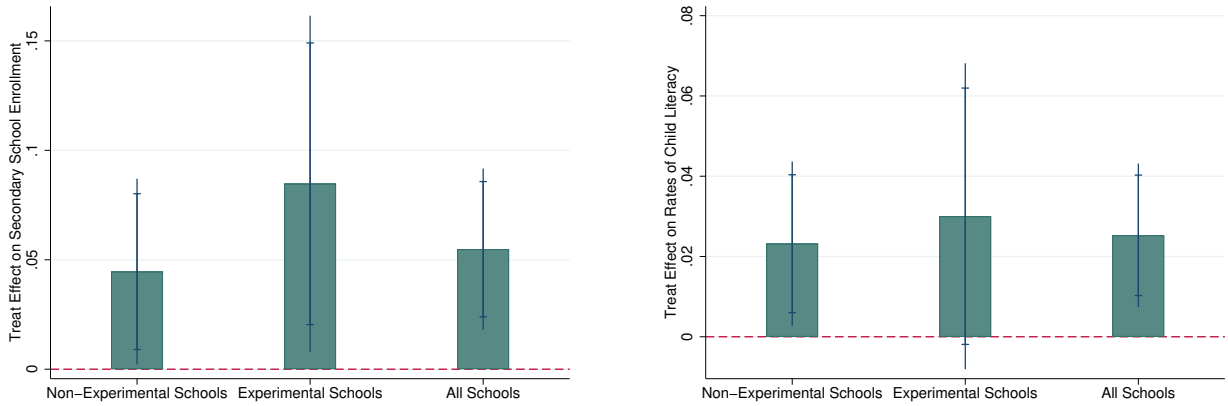
We analyze the impact of the policy implementation of the program using the following linear regression model:

$$Y_j = \alpha_0 + \alpha_1 Plus_j + \boldsymbol{\delta'}\, \boldsymbol{X}_j^{Criteria} + \epsilon_j \,, \tag{6}$$

where Y_j is a school-level outcome for school j, while $Plus_j$ takes a value of one if school j receives a mentor during the government implementation of the *Plus* modality, and zero otherwise. The vector $\boldsymbol{X}_j^{Criteria}$ consists of all the criteria used for the assignment of the program. The parameter of interest, α_1, represents the effect of the program during the government implementation on the outcome of interest. As was the case for the exposure analysis discussed in the previous section and formalized by equations (4)-(5), to causally interpret the estimates we need the assignment of the program across communities to be *conditionally* as good as random. In other words, conditional on the assignment criteria, schools that receive and do not receive the program at scale are similar in terms of unobserved characteristics. As before, we run some placebo tests to bolster the credibility of this identification assumption in our setting. Table B-13 shows the results. The 2017 government assignment is not unconditionally random (odd columns of the table), as priority is given to more disadvantaged communities. Instead, when we control for the vector $\boldsymbol{X}_j^{Criteria}$, the estimated coefficients displayed in the even columns of Table B-13 are very small and statistically insignificant.

Panel A of Figure 4 shows the results for secondary school enrollment after two years from

Figure 4: Policy Impact on Education Outcomes



Panel A: Secondary School Enrollment

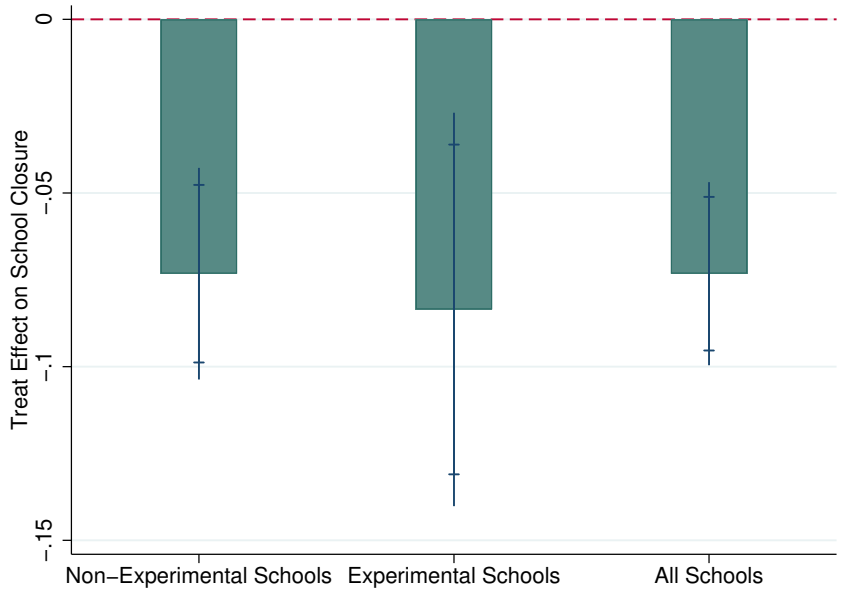
Panel B: Child Literacy

Notes: The bars in the figure represent the OLS estimates of the assignment to the API program during the government implementation of the *Plus* modality on school-level secondary school enrollment rates (Panel A) and child literacy rates (Panel B). Vertical lines overlaid on each bar displays the 95 percent and 90 percent confidence intervals, respectively. Confidence intervals are based on asymptotic inference.

the assignment of the mentors under the *Plus* modality at scale. Each bar represents the treatment effect of the policy (α_1 in equation (6)) for a different sample of schools in the state of Chiapas. For the sample of schools that did not participate in the second experiment, we find that the program increases the fraction of children who enroll in secondary schools by 4.5 percentage points (p-value = 0.039), which represents an increase of 6.5 percent with respect to the sample mean. For the schools that were part of the experiment, the impact of receiving the program during the government implementation is larger (+8.5p.p., p-value = 0.031), although the two estimates are statistically similar. These effects on secondary school enrollment are in line with the experimental findings on the enrollment in seventh grade (+12.4 percentage points, see Table 2). We interpret this result as evidence that the program at scale is effective in increasing schooling opportunities despite the change created by the policy implementation. Finally, the pooled estimates of the entire population of schools in Chiapas (+5.5 percentage points, p-value = 0.004) are in line with the estimated effects of each subpopulation of schools.

The results for child literacy are shown in Panel B of Figure 4. After two years of rollout of the program, we find that villages that received mentors under the *Plus* modality at scale display a 2.3 percentage points (p-value = 0.026) increase in child literacy rates when compared to villages without mentors. The magnitude of this effect implies a reduction of illiteracy rates by 20 percent with respect to the sample average. The estimated program effect for the

Figure 5: Policy Impact on School Closures



Notes: The bars in the figure represents the OLS estimates of the assignment to the API program during the government implementation of the *Plus* modality on the rate of school closures as measured over the subsequent two years. Vertical lines overlaid on each bar display the 95 percent and 90 percent confidence intervals, respectively. Confidence intervals are based on asymptotic inference.

subsample of experimental schools is quantitatively similar, although a bit noisier (+3.0 percentage points, p-value = 0.122). The pooled result for all schools (+2.5 percentage points, p-value = 0.006) mirrors the analysis for the two subpopulations. Overall, our results confirm the conclusion that the program is scalable and it enhanced an achievement outcome for children in these disadvantaged communities.

We conclude the analysis of the impact of the mentoring program at scale by looking at school closures. As previously described, school closures represent a major threat to the scalability of the program as schools are the means through which the mentors reach the children and their families. At the same time, we also show that parents act as a channel of scalability by preventing schools from closing through increased parental engagement. To gauge whether this mechanism persists during the government implementation of the *Plus* modality, it is important to test whether the impact of the policy on school closures is consistent across different situations for the experimental sample, as well as across different subpopulations within the state of Chiapas.

Figure 5 shows the results. The government implementation of the *Plus* modality induces a significant and sizable effect on school closures across all the subpopulations considered.

When we focus on the set of schools outside of the experimental sample, we find that the program reduces the probability of a school closing by 7.3 percentage points (p-value $<$ 0.001). Schools that were part of the experimental sample also experience a sizable decrease in school closures during the policy implementation at scale, with an average impact of the mentoring program of -8.3 percentage points (p-value $= 0.004$). The magnitude of this effect is remarkably similar to the corresponding impact of the API *Plus* program on school closures after two years under the experimental situation *for the same set of schools* (-8.3 percentage points., see Table 8). The pooled effect in the overall population of schools in Chiapas is -7.3 p.p. (p-value < 0.001). This last piece of evidence strongly suggests that the underlying mechanism through which schools are more likely to remain open with the API program persists across different situations. By enhancing parental engagement, the policy implementation of the *Plus* modality prevented disruptions in the school environment, which is a necessary condition for the success of the program at scale.

7 Discussion and Conclusion

We study a school mentoring program with a home visit component in the state of Chiapas, Mexico. By exploiting two independently run field experiments, we show that relatively small differences in program design (the training module in our case) can spur substantial differences in final outcomes. We confirm that the program as it was originally implemented by the government is largely ineffective. One alternative modality of the program (*Plus*) that features enhanced training for the mentors/home visitors is successful in enhancing test scores and educational attainment for the students in our sample. Parents not only increased their interactions and investment with children—a shared result among past successful interventions—but also they intensified their engagement at the school and community level.

The enhanced program modality is found to be effective after the government scale-up. The national rollout of the mentoring program, which fully converts the *Original* modality into the *Plus* modality with enhanced training for all the schools, provides us with an opportunity to study the mechanisms through which education interventions can be successfully scaled-up. Even when the evaluation sample is broadly representative of the targeted population, changes in the implementation protocol during the transition between the field experiment and the policy rollout can threaten the success of the program at scale. In our context, school

closures represent a major concern during this transition. We document that the exposure to the mentoring program at scale practically eliminates this issue. Parental responses are shown to be the key mechanism through which schools remain open, thereby ensuring the viability of the mentoring program as implemented by the government. The magnitudes of the estimated impacts are remarkably comparable across situations (field experiment versus government implementation) for our experimental sample as well as for the rest of the schools in Chiapas that experienced a change in program modality (from *Original* to *Plus*) during the government rollout.

Beyond the specific context of the analysis, we believe our case study can provide broader lessons for scholars who are interested in designing and evaluating scalable interventions. Whenever possible, we reiterate the importance of evaluating programs "at scale." This represents both an opportunity to exploit the existing infrastructure of the program as well as a restriction in terms of program design, since researchers have to consider various institutional constraints and supply-side issues. While we do not necessarily advocate sampling the entire population of beneficiaries, the composition of the evaluation sample needs to reflect the targeting criteria of the intervention, and the units of analysis should be large enough so as to encompass local spillover/general equilibrium effects that likely arise in those situations. It is also crucial to consider the joint impact of the intervention on the targeted actors. In the context of mentoring and educational programs, parental responses need to be taken into account jointly with children's outcomes both ex ante (e.g., for power calculations) and ex post (e.g., when adjusting inference procedures for multiple hypothesis testing and when conducting omnibus statistical significance tests). This is key for scalability since we have shown that it is precisely the interplay in the behavioral responses between these actors that determines the success of the programs at scale.

Finally, our work stresses the importance for scalability of the local parental incentives necessary to achieve a "Collaborative Equilibrium" in the community, as discussed in Section 5.1. Our model highlights a key trade-off between the degree of complementarity among parents, their incentive to cooperate, and the minimal actionable impact on parents that favors scaling. This result sheds light on the current debate on how to design mentoring interventions aimed at promoting better educational opportunities for children in disadvantaged contexts, including poor neighborhoods in the United States. While every parent is certainly unique to her own child (and hence not scalable), engaged communities of parents are potentially available at scale to promote the success of educational programs.

References

Agostinelli, Francesco, "Investing in Children's Skills: An Equilibrium Analysis of Social Interactions and Parental Investments," 2018.

_ **and Matthew Wiswall**, "Estimating the Technology of Children's Skill Formation," Working Paper 22442, National Bureau of Economic Research July 2016.

_ **, Matthias Doepke, Giuseppe Sorrenti, and Fabrizio Zilibotti**, "It Takes a Village: The Economics of Parenting with Neighborhood and Peer Effects," Working Paper 27050, National Bureau of Economic Research April 2020.

_ **,** _ **,** _ **, and** _ , "When the Great Equalizer Shuts Down: Schools, Peers, and Parents in Pandemic Times," *Journal of Public Economics*, 2022, *206*, 104574.

Al-Ubaydli, Omar, John A. List, and Dana Suskind, "2017 Klein Lecture: The Science of Using Science: Toward an Understanding of the Threats to Scalability," *International Economic Review*, 2020, *61* (4), 1387–1409.

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

Attanasio, Orazio, Helen Baker-Henningham, Raquel Bernal, Costas Meghir, Diana Pineda, and Marta Rubio-Codina, "Early Stimulation and Nutrition: The Impacts of a Scalable Intervention," *Journal of the European Economic Association*, 01 2022.

_ **, Sarah Cattan, and Costas Meghir**, "Early Childhood Development, Human Capital, and Poverty," *Annual Review of Economics*, 2022, *14* (1).

Bando, Rosangela and Claudia Uribe, "Experimental Evidence on Credit Constraints," Working Paper 670, Inter-American Development Bank February 2016.

Banerjee, Abhijit V., Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukerji, Marc Shotland, and Michael Walton, "From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application," *Journal of Economic Perspectives*, November 2017, *31* (4), 73–102.

Bobba, Matteo and Jérémie Gignoux, "Neighborhood Effects in Integrated Social Policies," *World Bank Economic Review*, 2019, *33* (1), 116–139.

Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Alice Ng'ang'a, and Justin Sandefur, "Experimental Evidence on Scaling Up Education Reforms in Kenya," *Journal of Public Economics*, 2018, *168* (C), 1–20.

Bruns, Barbara and Javier Luque, *Great Teachers : How to Raise Student Learning in Latin America and the Caribbean*, Washington, DC: World Bank, 2015.

Cameron, Lisa, Susan Olivia, and Manisha Shah, "Scaling Up Sanitation: Evidence from an RCT in Indonesia," *Journal of Development Economics*, 2019, *138*, 1–16.

Carneiro, Pedro, Emanuela Galasso, Italo Lopez Garcia, Paula Bedregal, and Miguel Cordero, "Parental Beliefs, Investments, and Child Development: Evidence from a Large-Scale Experiment," IZA Discussion Papers 12506, Institute of Labor Economics (IZA) July 2019.

CONEVAL, "Medición de Pobreza 2008-2018, Estados Unidos Mexicanos," 2018.

Cunha, Flavio, James J. Heckman, and Susanne M. Schennach, "Estimating the Technology of Cognitive and Noncognitive Skill Formation," *Econometrica*, 2010, *78* (3), 883–931.

Davis, Jonathan M.V., Jonathan Guryan, Kelly Hallberg, and Jens Ludwig, "The Economics of Scale-Up," Working Paper 23925, National Bureau of Economic Research October 2017.

Doepke, Matthias and Fabrizio Zilibotti, "Parenting With Style: Altruism and Paternalism in Intergenerational Preference Transmission," *Econometrica*, September 2017, *85*, 1331–1371.

Dubeck, Margaret M. and Amber Gove, "The Early Grade Reading Assessment (EGRA): Its Theoretical Foundation, Purpose, and Limitations," *International Journal of Educational Development*, 2015, *40*, 315–322.

Engzell, Per, Arun Frey, and Mark Verhagen, "Learning Inequality During the Covid-19 Pandemic," 2020. Mimeo.

Fryer, Roland G. Jr., Steven D Levitt, and John A. List, "Parental Incentives and Early Childhood Achievement: A Field Experiment in Chicago Heights," Working Paper 21477, National Bureau of Economic Research August 2015.

Gertler, Paul J., Harry Anthony Patrinos, and Marta Rubio-Codina, "Empowering Parents to Improve Education: Evidence from Rural Mexico," *Journal of Development Economics*, 2012, *99* (1), 68–79.

Heckman, James, "Randomization and Social Policy Evaluation," in "Evaluating Welfare and Training Programs. Edited by C. F. Manski and I. Garfinkel," Harvard University Press, 1992.

_ **and Jin Zhou**, "Interactions as Investments: The Microdynamics and Measurement of Early Childhood Learning," Working Paper, Center for the Economics of Human Development, University of Chicago 2021.

Heckman, James J. and Stefano Mosso, "The Economics of Human Development and Social Mobility," *Annual Review of Economics*, 2014, *6*, 689–733.

List, John A., *The Voltage Effect: How to Make Good Ideas Great and Great Ideas Scale*, Penguin Books, 2022.

_ **, Azeem M. Shaikh, and Yang Xu**, "Multiple Hypothesis Testing in Experimental Economics," *Experimental Economics*, December 2019, *22* (4), 773–793.

_ **, Fatemeh Momeni, and Yves Zenou**, "The Social Side of Early Human Capital Formation: Using a Field Experiment to Estimate the Causal Impact of Neighborhoods," Working Paper 28283, National Bureau of Economic Research December 2020.

Maldonado, Joana E. and Kristof De Witte, "The Effect of School Closures on Standardised Student Test Outcomes," 2020. KU Leuven Discussion Paper Series 20.17.

Maniadis, Zacharias, Fabio Tufano, and John A. List, "One Swallow Doesn't Make a Summer: New Evidence on Anchoring Effects," *American Economic Review*, January 2014, *104* (1), 277–90.

Martínez, Edmundo Ramírez, "Supporting Education in Families and Schools," Unpublished manuscript, Impact Evaluation Report (in spanish) November 2012.

Miguel, Edward and Michael Kremer, "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities," *Econometrica*, January 2004, *72* (1), 159–217.

Mobarak, A. Mushfiq and Austin C. Davis, "A Research Agenda Built for Scale," in "The Scale-up Effect in Early Childhood and Public Policy. Edited by John A. List, Dana Suskind, and Lauren H. Supplee," Routledge, 2021.

Muralidharan, Karthik and Abhijeet Singh, "Improving Public Sector Management at Scale? Experimental Evidence on School Governance India," Working Paper 28129, National Bureau of Economic Research November 2020.

_ **and Paul Niehaus**, "Experimentation at Scale," *Journal of Economic Perspectives*, 2017, *31* (4), 103–124.

O'Brien, Peter C., "Procedures for Comparing Samples with Multiple Endpoints," *Biometrics*, 1984, *40* (4), 1079–1087.

Platas, Linda M., Leanne R. Ketterlin-Geller, and Yasmin Sitabkhan, "Using an Assessment of Early Mathematical Knowledge and Skills to Inform Policy and Practice: Examples from the Early Grade Mathematics Assessment," *International Journal of Education in Mathematics, Science and Technology*, 2016, *4(3)*, 163–173.

Romano, Joseph P. and Michael Wolf, "Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing," *Journal of the American Statistical Association*, March 2005, *100*, 94–108.

_ **and** _ , "Stepwise Multiple Testing as Formalized Data Snooping," *Econometrica*, July 2005, *73* (4), 1237–1282.

_ **and** _ , "Efficient Computation of Adjusted p-Values for Resampling-Based Stepdown Multiple Testing," *Statistics & Probability Letters*, 2016, *113* (C), 38–40.

Westfall, Peter H. and S. Stanley Young, *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment* 1993.

Young, Alwyn, "Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results," *The Quarterly Journal of Economics*, 2019, *134* (2), 557–598.

Zhou, Jin, Alison Baulos, James J. Heckman, and Bei Liu, "The Economics of Child Development with an Application to Home Visiting at Scale," in "The Scale-up Effect in Early Childhood and Public Policy. Edited by John A. List, Dana Suskind, and Lauren H. Supplee," Routledge, 2021.

Appendices

A Data Description

A.1 Administrative Data

School census. The Ministry of Education runs a school census (*Formato 911*) at the beginning and at the end of each school cycle that covers all public schools in Mexico. The census asks the school representative about the number of students enrolled in every grade and whether they are new students or repeaters. Additional information includes the number of instructors and the number of classrooms per school. Information from the 2013 Census is used to construct the baseline school variables that are displayed in Table B-1 and in Panel A of Table B-2. School census data for the years 2015–2020 are used to track the school closures during the government implementation of both the API *Original* and *Plus* modalities, as shown in Table 8 and Figure 5.

Locality-level Population census: The National Institute of Statistics and Geography (INEGI) is in charge of compiling a population count with detailed information on socio-demographics, poverty, and education, among other information every decade. Census data are made available at the individual level for a small random sample of the population, as well as at the locality-level for the universe of localities in Mexico. We use the locality-level information collected in the census rounds of 2010 and 2020 for our analysis. In particular, we use information from the 2010 population census in Table B-1, in Panel B of Table B-2. as well as to construct the indicator variable for the presence of conflicts in the community that is shown in Table 7. We leverage information on schooling outcomes in the 2020 census for all the localities in the state of Chiapas (including those that were part of the experimental sample), which is shown in Figure 4.

Standardized test scores. Between 2007 and 2013, all Mexican students in third grades through ninth grade were required to take a standardized test, the ENLACE (*Evaluación Nacional de Logro Academico en Centros Escolares*). The test was administered by external proctors at the end of each academic year, and it assessed student knowledge in three areas: math, Spanish, and, starting in 2008, a third subject that rotated between science, ethics/civics, history, or geography. We use the school-level average of the Spanish scores in 2012 to construct the strata for the school-level randomization of the second experiment. In

the first experiment, we use individual scores for sixth graders in each pedagogical area in 2013 as our main measures of academic achievement. The *Overall Score* displayed in Table 1 is computed using GLS-weighted score over the three scores (O'Brien, 1984). Last, we use the 2013 ENLACE scores at the school level for the placebo tests displayed in Tables B-12 and B-13.

Transitions to Secondary Schools. We link the enrollment records of the sixth graders in the sample of the second experiment across the population of seventh graders in Chiapas during the following academic year. Individual transitions computed in the school year 2016–2017 (i.e., by the end of the second experiment) are reported in Table 2, while transitions computed in the school year 2017–2018 (i.e., after the first year of the government implementation of the API *Plus* modality) are reported in Figure 3.

Other administrative records. All students in Chiapas schools, irrespective of whether they received the API program, must undergo a diagnostic test at the beginning of each school year. The test covers three subjects: math, Spanish, and natural science. The score for each subject ranges between 5 and 10. We use the individual-level average across the three subjects in the diagnostic tests at the beginning of the 2014–2015 school year to construct the within-school student rankings displayed in Figure B-2 and Table B-10, which proxy for the individual eligibility for the one-on-one remedial education sessions. We also use the official assessments assigned to the students based on those tests (level 1, level 2, and level 3) in Table B-3.

We use student-level longitudinal information for the population of primary schools to construct various measures of school-level changes in student composition reported in Table B-6: whether the student must repeat a grade in school year 2015–2016, attrition from the school system in Chiapas between the school years 2014–2015 and 2015–2016, and whether in 2015-2016 the student attends the same school as in 2014–2015.

A.2 Survey Data

Data collection took place in the spring of 2016 in the 224 schools and the surrounding communities that form part of the second experiment. The household module of the survey was collected for a random sample of five households within a five kilometer radius from each school. The information is linked at the child-parent level with the student test scores through unique student identifiers. It entails the following array of survey modules and

measurement tools.

Measures of Children's Achievement. The reading scores reported in Tables 2 and B-10 are given by the latent factor of an exploratory factor analysis of the following eight domains: 1) letter name, 2) initial name, 3) initial sound, 4) word recognition, 5) word reading, 6) reading comprehension, 7) listening, 8) dictation. The math scores reported in Tables 2 and B-10 are given by the latent factor of an exploratory analysis of the following seven domains: 1) number identification, 2) number discrimination, 3) missing number, 4) addition, 5) subtraction, 6) problem solving, 7) shape recognition. An orthogonal rotation is applied before standardizing each factor with respect to the mean and the standard deviation in the control group. The individual components of the math and reading scores are reported in Table B-7.

The household survey contains a set of measures of behavioral problems reported by the caregivers of the children in our sample. The socio-emotional scores reported in Tables 2 and B-10 are the sum of the following thirty-two items on how often the child displays a given emotion/behavior: 1) has serendipitous mood changes, 2) feels or complains that nobody loves him/her, 3) is tense or nervous, 4) lies or cheats, 5) is scared or anxious, 6) talks and argues too much, 7) has difficulty focusing on a specific activity for an extended amount of time, 8) gets easily confused, 9) has his/her head is in the clouds, 10) threatens or is mean with other children, 11) tends to challenge parental authority, 12) does not feel guilty after a bad deed, 13) does not get along with other children, 14) is impulsive or acts "fast" without thinking, 15) has inferiority issues, 16) has no friends, 17) has difficulty letting go of certain thoughts, 18) is hyper active, 19) has a bad temper or is irascible, 20) easily loses his/her temper, 21) feels unhappy, sad, or depressed, 22) is shy, does not socialize with others, 23) breaks objects on purpose, 24) is too attached to adults, 25) cries too much, 26) demands a lot of attention, 27) is too much dependent on others, 28) is afraid of other people's judgment, 29) tends to be in bad company; 30) reserved, keeps things for himself/herself, 31) worries about everything, 32) misbehaves at school and does not respect the instructor.

The *Overall Score* of students' achievement displayed in Table 2 is computed using GLS-weighted averages over the two cognitive measures and the socio-emotional score.

Parenting Practices. The household survey collects information on parents' behavior and investment in their children's education. The same information was collected during the mid-line survey of the first experiment. The parental engagement outcomes reported in Table 3 are computed using GLS-weighted averages (Anderson, 2008) over different indicators of

parental behavior. For *Engage at School*: whether or not parents (i) volunteer at the school, (ii) donate money to the school, (iii) donate in kind to the school, and (iv) offer food to the instructor. For *Manage School Resources*: whether or not parents (i) directly manage the school budget, (ii) propose some materials to the school, (iii) decide to use some materials for the school, and (iv) decide on how to allocate money for some school activities, and (v) define the pedagogical targets of the school. For *Engage with Child*: whether (i) parents help with their child's homework, (ii) meet with the instructor, (iii) expect their child to complete secondary education or more, and (iv) children participate in other academically-related activities outside the school hours. The *Engagement Index* is the same GLS-weighted average over each of the individual components described above, which are reported in Table B-9.

Parent-Mentor Interactions. The household module collects several questions on both the quantity and the quality of parents' interactions with the mentors for those households that were assigned to either the API *Original* group or the API *Plus* group. This information is used to construct the four variables reported in Panel A of Table 4. Basic information on both the household module respondent and household characteristics is reported in Panel C of Table B-2.

Parenting Styles. The mentors' questionnaire included a battery of questions on the specific competencies they promote during their interactions with parents. The indicator variables for each competency are used as outcomes variables in Panel B of Table 4. Since the mentors were not located in the communities on a continuous basis, the survey firm interviewed them by an end-of-year evaluation session. Some of their characteristics are reported in Panel D of Table B-2, as well as in Table B-3 for the subset of the mentors who reported working in different schools from those they were initially assigned to.

Teaching Practices. We measure time use and different learning activities of community instructors as well as their ability to keep students engaged using an adapted version of Stallings classroom snapshot, which is a rubric for timed observations that has been used previously in Mexico (Bruns and Luque, 2015). An observer scores the instructor's effective use of 15 different activities over the course of a full one-hour lesson, with snapshots every three minutes. Each activity was scored between 1 and 4. In every snapshot, the external observer reports whether the instructor is present in the classroom. Given the nature of the API intervention and the multi-grade context, the tool was adapted to capture the instructor's ability to use materials and keep the rhythm of the class.

The information included in this survey module is used to construct GLS-weighted averages over the different types of teacher behavior, which are displayed in Table B-11. *Learning Activities* is the sum of the amount of time children spend on (i) reading aloud alone, (ii) reading aloud in a group, (iii) questions and answers, (iv) memorizing, (vi) individual homework, and (viii) verbal tasks. *Engage with Students* is the sum of the amount of time the instructor spends on (i) elaborating on a given concept, (ii) students were not involved, and (iii) keeping discipline. *Manage Time* is the amount of time the instructor spends (i) out of the classroom, (ii) effectively administering some tasks in the classroom, (iii) whether or not the instructor complies with the start and end time of each classroom, (iv) whether or not the instructor keeps the rhythm of the class as well as of the individual students according to their age and their mother-tongue, and (v) whether or not the students were grouped according to their respective academic levels. *Use of Material* is the sum of four indicator variables: (i) whether the instructor uses any book to explain a given topic, (ii) whether the instructor uses any material from the community to explain a given topic, (iii) whether drawings and other students' artworks are displayed in the classroom, and (iv) whether charts and maps are displayed in the classroom. The *Overall Index* is the same GLS-weighted average of the individual components of teacher behavior described above.

Local instructors were also asked standard questions on their socio-demographic characteristics, education, experience and, if they were in the treatment group, their relationship with the mentors. Those are reported in Panel B of Table B-2.

A.3 In-Depth Interviews

In the spring of 2022 we implemented a series of semi-structured phone interviews with a small sample of local instructors and mentors who participated in the program. In total, we were able to locate and contact 104 local instructors and 68 mentors. Of those, 12 instructors and 16 mentors agreed to complete the phone interview. More than half of the survey respondents continued working as mentors after the 2016 government implementation of the *Plus* modality. The characteristics of the survey respondents in comparison with the overall sample are shown in Tables B-4 and B-5.

The survey contains a series of open questions related to the experiences of the mentors/local instructors with the parents in the communities. Below, we report the original quotes in Spanish that we refer to in the main body of the paper (authors' translation from Spanish). In particular, these quotes from the mentors about the peer-to-peer sessions of the training

are reported in Section 3.3:

> "*Fue un momento de la capacitación en donde me dijeron que debía adaptarme al contexto de su centro del trabajo, de comprender las necesidades y de entender situaciones que se vivían en la misma comunidad, para poder dialogar con los padres y atender a los niños sin afectar o modificar lo que ellos conciben como su medio.*"

> "*Recomendaban hacer las visitas domiciliarias con frecuencia y ayudarle en algo a los papás o salían con ellos a visitas y les daba más confianza.*"

> "*[Las sesiones de orientacion me permitieron] escuchar las diferentes estrategias que ellos tenían para poder probarlas e implementarlas.*"

These quotes from the local instructors about the role of parents in the day-by-day routine of the school are reported Section 5.2.

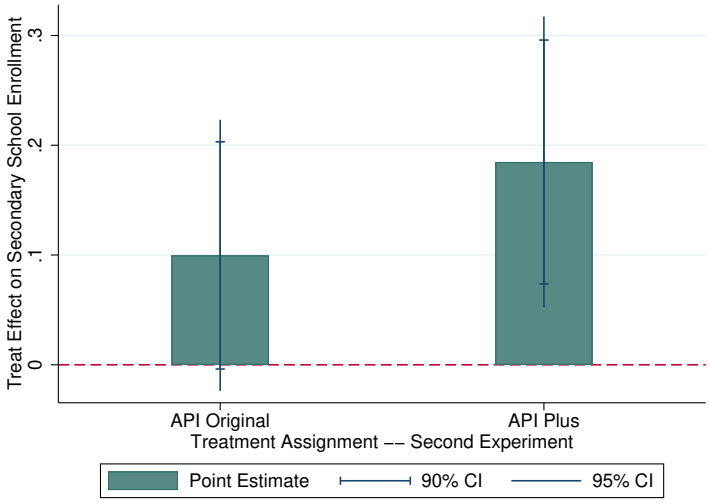
> "*La gestión dela escuela y se le hicieron mejoras de cercado, pintaron la escuela arreglaron los baños y se compraron materiales.*"

> "*Eran participativos, estaban pendientes del bienestar de la escuela por ejemplo la construcción, de materiales e incluso de los desayunos y alimentación del instructor.*"

> "*Los padres apoyaban en el seguimiento al bloc de tareas y trabajaban en equipo cuando los API que no podían estar presentes por apoyar a otra comunidad, los mantenían al corriente o, incluso un poco más avanzados, por lo que cuando los APIs regresaban podían dar continuidad a sus clases sin ningún atraso.*"

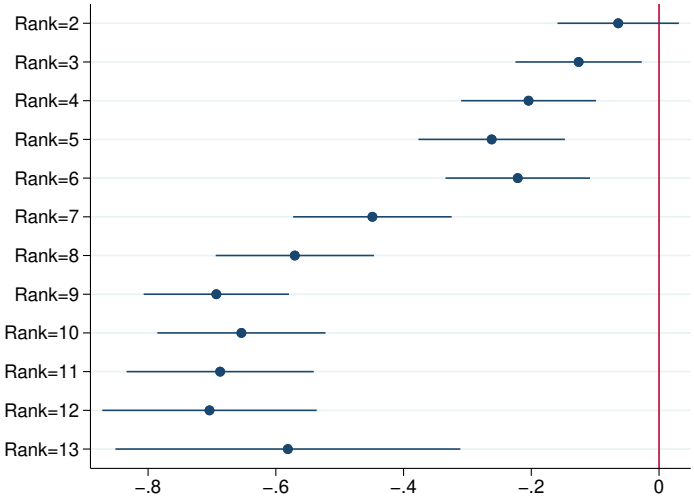
B Additional Figures and Tables

Figure B-1: Treatment Effects on Secondary School Enrollment During the Transition Between the Second Experiment and the Government Implementation



Notes: The bars depicted in this figure show the OLS estimates of the original treatment assignments in our experiment on the probability of enrolling in seventh grade in the year after the end of the second experiment (2017). The vertical lines overlaid on the bars represent asymptotic confidence intervals at the 90 percent and the 95 percent confidence levels. Confidence intervals are based on asymptotic inference. The sample includes 207 schools of the 224 that were part of the experiment. Beyond a school that permanently closed, the sample attrition is caused by schools not having sixth graders during that school year. This fact is consistent with the multi-grade nature of the CONAFE system. Attrition is balanced among schools that were part of the two treatment arms (p-values = 0.914, and 0.768)

Figure B-2: Probability of Being in Remedial Sessions by Inverted Achievement Rank



Notes: The dots in this figure are estimated marginal effects from Probit regression models of indicator variables for the inverted within-school student rank based on the average score on the diagnostic tests in math, Spanish, and natural science on the probability of participating in the one-on-one remedial education sessions with the mentors. The indicator variable for whether the student is ranked first (i.e., the worst-performing student in the class) is the omitted category. The horizontal lines around each dot represent 90 percent confidence intervals. Confidence intervals are based on asymptotic inference.

Table B-1: Baseline Characteristics and Covariate Balance – First Experiment

	Panel A: Original Sample of Schools				
	Treatment (40)		Control (40)		Diff
	Mean	Std. Dev.	Mean	Std. Dev.	P-Value
Number of households	43.050	117.302	48.100	89.883	0.832
Total Population	205.250	575.294	227.275	480.694	0.855
Share Economically Active Pop	0.287	0.066	0.287	0.077	0.970
Water connection (Y/N)	0.025	0.158	0.050	0.221	0.567
Sewer system (Y/N)	0.025	0.158	0.025	0.158	1.000
Share of analphabet population	0.320	0.180	0.319	0.171	0.984
Share of dwellings with dirt floor	0.334	0.356	0.334	0.254	0.999
Garbage collection (Y/N)	0.025	0.158	0.050	0.221	0.564
ENLACE Spanish 2010	399.476	38.545	398.560	27.853	0.901
ENLACE Math 2010	375.065	42.763	386.407	50.357	0.279
Enrollment	16.231	9.192	15.550	8.246	0.712
Number of Teachers	1.385	0.544	1.450	0.597	0.606
Share over-aged students	3.135	9.832	1.884	3.941	0.464
	Panel B: Schools in Mid-Line 2012 Survey				
	Treatment (37)		Control (36)		Diff
	Mean	Std. Dev.	Mean	Std. Dev.	P-Value
Number of households	45.297	121.712	49.667	94.636	0.863
Total Population	217.054	597.061	234.778	506.694	0.888
Share Economically Active Pop	0.286	0.064	0.276	0.069	0.553
Water connection (Y/N)	0.027	0.164	0.056	0.232	0.547
Sewer system (Y/N)	0.027	0.164	0.028	0.167	0.990
Share of analphabet population	0.321	0.170	0.333	0.173	0.745
Share of dwellings with dirt floor	0.307	0.330	0.349	0.261	0.544
Garbage collection (Y/N)	0.027	0.164	0.056	0.232	0.551
ENLACE Spanish 2010	401.971	38.973	399.036	28.974	0.703
ENLACE Math 2010	377.916	43.159	388.422	51.038	0.351
Enrollment	15.917	8.334	14.917	7.987	0.597
Number of Teachers	1.389	0.549	1.417	0.604	0.827
Share over-aged students	2.134	7.225	1.961	4.094	0.900
	Panel C: Schools with Test Score 2013 Data				
	Treatment (35)		Control (35)		Diff
	Mean	Std. Dev.	Mean	Std. Dev.	P-Value
Number of households	46.971	124.974	48.686	95.832	0.950
Total Population	225.857	612.996	227.543	512.201	0.990
Share Economically Active Pop	0.287	0.065	0.278	0.069	0.579
Water connection (Y/N)	0.029	0.169	0.057	0.236	0.568
Sewer system (Y/N)	0.029	0.169	0.029	0.169	1.000
Share of analphabet population	0.327	0.165	0.335	0.175	0.823
Share of dwellings with dirt floor	0.321	0.334	0.345	0.264	0.733
Garbage collection (Y/N)	0.029	0.169	0.057	0.236	0.566
ENLACE Spanish 2010	401.869	40.034	399.206	29.378	0.748
ENLACE Math 2010	377.168	44.284	390.561	50.120	0.242
Enrollment	15.971	8.449	14.743	8.034	0.527
Number of Teachers	1.400	0.553	1.400	0.604	1.000
Share over-aged students	2.195	7.321	2.017	4.140	0.900

Notes: This table shows means and standard deviations for community and school characteristics collected in the population census (2010) and the school census (2010). See Appendix A.1 for more details on these data sources. The fifth column reports the associated p-values of the differences in means between the treatment and the control group.

Table B-2: Baseline Characteristics and Covariate Balance – Second Experiment

Sample	Control	API Original	API Plus	All Evaluation	
(Number of Schools)	(100)	(60)	(70)	(230)	
Statistic	Mean	Mean	Mean	Original-Control	Plus-Control
	(SD)	(SD)	(SD)	(SE)	(SE)
	Panel A: School Characteristics				
Average Test score Spanish	431.88	431.65	431.36	-0.223	-0.516
	(64.43)	(66.87)	(66.60)	(2.581)	(2.783)
Average Test score Math	455.75	454.85	451.68	-0.902	-4.076
	(80.71)	(83.50)	(81.06)	(5.790)	(6.911)
Average Test score Science	440.15	441.24	441.27	1.095	1.120
	(52.52)	(48.66)	(50.89)	(4.273)	(4.784)
Community Instructors	1.220	1.300	1.200	0.080	-0.020
	(0.416)	(0.462)	(0.403)	(0.066)	(0.067)
Number of Enrolled Students	15.160	15.314	14.233	0.154	-0.927
	(5.839)	(5.714)	(5.782)	(0.901)	(0.946)
	Panel B: Community Instructors Characteristics				
Lower than upper second.	0.067	0.062	0.066	-0.002	0.009
	(0.251)	(0.242)	(0.250)	(0.035)	(0.033)
Lower than higher ed.	0.918	0.901	0.908	-0.000	0.002
	(0.276)	(0.300)	(0.291)	(0.044)	(0.040)
Training weeks at baseline	4.515	4.704	4.500	0.128	-0.042
	(1.322)	(1.259)	(1.426)	(0.196)	(0.253)
3rd and 4th grade students	3.655	3.986	3.716	0.346	0.137
	(2.434)	(2.286)	(2.230)	(0.349)	(0.356)
5th and 6th grade students	3.517	3.838	3.507	0.325	0.054
	(2.408)	(2.507)	(2.298)	(0.354)	(0.352)
	Panel C: Household Characteristics				
Indigenous Language	0.326	0.366	0.476	0.049	0.142
	(0.469)	(0.483)	(0.500)	(0.065)	(0.077)
Read	0.715	0.686	0.734	-0.031	0.022
	(0.452)	(0.465)	(0.443)	(0.041)	(0.042)
Less than Primary	0.615	0.587	0.584	-0.028	-0.030
	(0.487)	(0.493)	(0.494)	(0.043)	(0.041)
Upper Sec. or Higher	0.015	0.016	0.019	-0.001	0.003
	(0.123)	(0.124)	(0.135)	(0.009)	(0.009)
Oportunidades	0.813	0.807	0.829	-0.003	0.015
	(0.391)	(0.395)	(0.377)	(0.033)	(0.031)
Refrigerator	0.397	0.387	0.373	-0.010	-0.019
	(0.490)	(0.488)	(0.485)	(0.047)	(0.055)
Television	0.692	0.738	0.651	0.048	-0.040
	(0.462)	(0.440)	(0.478)	(0.047)	(0.051)
Car	0.084	0.081	0.063	-0.003	-0.019
	(0.277)	(0.273)	(0.244)	(0.027)	(0.024)
Sewage	0.254	0.253	0.320	-0.003	0.068
	(0.436)	(0.435)	(0.467)	(0.042)	(0.052)
Phone	0.220	0.233	0.204	0.014	-0.014
	(0.414)	(0.423)	(0.404)	(0.037)	(0.038)
Light	0.863	0.916	0.873	0.054	0.006
	(0.344)	(0.278)	(0.333)	(0.040)	(0.040)

	Panel D: Mentors' Characteristics		
	API Original	API Plus	Difference
Variable	Mean	Mean	Plus-Std
	(SD)	(SD)	(SE)
Age	28.491	28.543	-0.135
	(3.760)	(3.075)	(0.650)
Male	0.566	0.587	-0.064
	(0.500)	(0.498)	(0.097)
High Edu Complete	0.887	0.891	0.014
	(0.320)	(0.315)	(0.066)
Previously Instructor	0.792	0.848	-0.079
	(0.409)	(0.363)	(0.072)
Previously Education Assistant	0.075	0.065	0.014
	(0.267)	(0.250)	(0.049)

Notes: The first three columns of the table report mean and standard deviations in parentheses for various characteristics collected before the assignment of the API program in the evaluation sample. The school variables in Panel A are computed from the 2013 national standardized tests and from the 2013 school census. The other characteristics reported in Panels B-D are collected in the survey data. The differences reported in the last two columns of the table are based on OLS estimates of the regression models that control for stratification dummies. Standard errors of the mean differences for the student characteristics are reported in parentheses in the last two columns and they are clustered at school level. See Appendix A for more details on the data sources.

Table B-3: Characteristics of Dropout Mentors

	Original	Plus	Plus - Original
Former CONAFE facilitator	0.689	0.703	0.012
	(0.468)	(0.463)	(0.102)
At least 5 days of training	0.467	0.514	0.061
	(0.505)	(0.507)	(0.111)
Sleeps in community (y/n)	0.711	0.757	0.052
	(0.458)	(0.435)	(0.097)
Number of nights in community last week	3.022	2.757	-0.301
	(2.061)	(1.978)	(0.442)
Number of students with personalized attention	6.049	5.767	-0.284
	(0.835)	(1.104)	(0.264)
Days spent in community during last month	10.220	10.200	0.063
	(4.613)	(4.715)	(1.148)
Number of students below Level 2	3.450	3.560	0.079
	(1.679)	(1.660)	(0.440)
Number of students below Level 3	2.727	2.731	-0.020
	(1.773)	(1.845)	(0.488)

Notes: This table reports means and standard deviations for the characteristics of the mentors who dropped out from the schools where they were originally assigned across API *Original* and API *Plus* modalities. The differences reported in the last column of the table are based on OLS estimates of the regression models that control for stratification dummies. Standard errors of the mean differences for the student characteristics are reported in parentheses in the last column and they are clustered at the school level. For detailed descriptions of the survey variables used in this table, see Appendix A.2.

Table B-4: Characteristics of Mentors—Sample vs Phone Survey

	Original Sample	2022 Survey	Difference
Age	28.443	27.556	0.888
	(3.260)	(3.941)	(1.150)
Male	0.585	0.778	-0.193
	(0.495)	(0.441)	(0.171)
High School Completed	0.868	1.000	-0.132
	(0.340)	(0.000)	(0.114)
Training Weeks	2.858	2.667	0.192
	(2.035)	(1.871)	(0.703)
Experience as Api	21.274	13.444	7.829
	(10.058)	(6.803)	(3.425)
Previously Local Instructor	0.840	0.778	0.062
	(0.369)	(0.441)	(0.130)
Previously Education Assistant	0.085	0.000	0.085
	(0.280)	(0.000)	(0.094)
Days Spent in the Community	13.528	13.556	-0.027
	(5.331)	(4.876)	(1.840)
Students Lagging Behind	5.698	5.889	-0.191
	(1.657)	(3.018)	(0.621)

Notes: This table reports means and standard deviations for the characteristics of the mentors in the main sample of the analysis and those of the mentors who participated in the in-depth phone interviews (2022). The differences reported in the last column of the table are based on OLS estimates of the regression models that control for stratification dummies. Standard errors of the mean differences for the student characteristics are reported in parentheses in the last column and they are clustered at school level. For detailed descriptions of the survey variables used in this table, see Appendix A.2.

Table B-5: Characteristics of Local Instructors—Sample vs. Phone Survey

	Original Sample	2022 Survey	Difference
Age	21.284	21.157	0.127
	(2.585)	(2.034)	(0.702)
Male	0.560	0.786	-0.226
	(0.497)	(0.426)	(0.135)
Lower than Upper Second	0.062	0.071	-0.010
	(0.241)	(0.267)	(0.066)
Upper Second Complete	0.800	0.643	0.157
	(0.401)	(0.497)	(0.111)
Above Upper Second	0.138	0.286	-0.148
	(0.346)	(0.469)	(0.097)
Experience in Months	13.545	13.429	0.117
	(9.408)	(9.362)	(2.577)
Training Weeks at Baseline	4.768	5.500	-0.732
	(4.114)	(5.019)	(1.140)
Time spent in the School	9.509	9.071	0.438
	(4.220)	(3.269)	(1.146)
Sleeps in the Community	0.651	0.857	-0.206
	(0.478)	(0.363)	(0.130)
Nights spent in the Community	3.204	3.071	0.132
	(2.065)	(2.093)	(0.566)

Notes: This table reports means and standard deviations for the characteristics of the mentors in the main sample of the analysis and those of the mentors who participated in the in-depth phone interviews (2022). The differences reported in the last column of the table are based on OLS estimates of the regression models that control for stratification dummies. Standard errors of the mean differences for the student characteristics are reported in parentheses in the last column and they are clustered at the school level. For detailed descriptions of the survey variables used in this table, see Appendix A.2.

Table B-6: Treatment Assignment and School-Level Student Composition

	Repeat	Attrition	Outside CONAFE in $t-1$	Same school in $t-1$
API Original	-0.011	-0.018	-0.002	0.019
	[0.116]	[0.322]	[0.895]	[0.295]
API Plus	-0.010	-0.006	-0.003	0.011
	[0.153]	[0.751]	[0.861]	[0.574]
H0: API Original = API Plus	[0.834]	[0.491]	[0.911]	[0.620]
Observations	1019	1019	1019	1019
Number of Clusters	224	224	224	224

Notes: This table shows the estimates of the two API modalities on various measures of school-level changes in student composition. The number of observations drops from 1045 to 1019 due to incomplete school identifiers (CURP) for 26 students. All p-values account for clustering at the school level. Asymptotic p-values reported in brackets are clustered at school level. For a detailed descriptions of the survey variables used in this table, see Appendix A.1.

Table B-7: Average Program Impacts by Subdomains of the Reading and the Math Scores

	Panel A: Share of Correct Reading Answers by Subdomain							
	Letter Name	Initial Name	Initial Sound	Word Recogn.	Word Reading	Read Comprehen.	Listening	Dictation
API Original	0.103	0.006	0.122	0.129	0.075	0.118	-0.004	0.129
	[0.232]	[0.941]	[0.156]	[0.091]	[0.300]	[0.107]	[0.963]	[0.120]
	{0.285}	{0.949}	{0.194}	{0.124}	{0.341}	{0.138}	{0.968}	{0.173}
	(0.449)	(0.996)	(0.365)	(0.255)	(0.510)	(0.290)	(0.996)	(0.314)
API Plus	0.240	-0.019	0.042	0.318	0.197	0.321	0.123	0.378
	[0.005]	[0.816]	[0.565]	[0.000]	[0.014]	[0.000]	[0.145]	[0.000]
	{0.010}	{0.824}	{0.584}	{0.000}	{0.026}	{0.001}	{0.185}	{0.000}
	(0.005)	(0.789)	(0.728)	(0.000)	(0.021)	(0.000)	(0.226)	(0.000)
API Original = API Plus	[0.180]	[0.771]	[0.343]	[0.039]	[0.183]	[0.023]	[0.094]	[0.005]
	{0.174}	{0.799}	{0.479}	{0.062}	{0.229}	{0.059}	{0.220}	{0.003}
	(0.328)	(0.727)	(0.421)	(0.077)	(0.328)	(0.045)	(0.194)	(0.010)
Observations	1044	1044	1044	1044	1044	1044	1044	1044
Clusters	224	224	224	224	224	224	224	224
	Panel B: Share of Correct Math Answers by Sub-Domain							
	Number Identif.	Number Discrim.	Missing Number	Add	Subtract	Problem Solving	Shape Recogn.	
API Original	0.094	0.036	0.099	0.011	0.061	-0.051	0.022	
	[0.252]	[0.661]	[0.192]	[0.874]	[0.402]	[0.481]	[0.789]	
	{0.301}	{0.681}	{0.226}	{0.882}	{0.447}	{0.511}	{0.800}	
	(0.576)	(0.919)	(0.483)	(0.923)	(0.789)	(0.817)	(0.923)	
API Plus	0.259	0.201	0.204	0.215	0.111	0.116	0.099	
	[0.005]	[0.026]	[0.022]	[0.003]	[0.103]	[0.156]	[0.316]	
	{0.011}	{0.036}	{0.035}	{0.008}	{0.130}	{0.200}	{0.365}	
	(0.007)	(0.033)	(0.033)	(0.007)	(0.137)	(0.163)	(0.247)	
API Original = API Plus	[0.095]	[0.103]	[0.218]	[0.008]	[0.500]	[0.046]	[0.396]	
	{0.163}	{0.129}	{0.420}	{0.020}	{0.514}	{0.080}	{0.550}	
	(0.191)	(0.191)	(0.361)	(0.008)	(0.516)	(0.090)	(0.516)	
Observations	1044	1044	1044	1044	1044	1044	1044	
Clusters	224	224	224	224	224	224	224	

Notes: This table shows OLS estimates and the associated p-values of the two API modalities: API *Original* and API *Plus* for 1,044 students enrolled in third to sixth grade by the end of the second school year since treatment assignment. For detailed descriptions of the sub-components of the reading and math scores used in this table, see Appendix A.2. The outcome variables are standardized with respect to their means and the standard deviations in the control group. The inference procedures take into account clustering of the error terms at the school level and the block randomization design at the strata level. p-values reported in brackets refer to the conventional asymptotic inference. p-values reported in braces are computed using randomization inference (randomization-t). All p-values account for clustering at the school level. p-values reported in parentheses are adjusted for testing each null hypothesis (null impact of API *Original*, API *Plus*, and the comparison) on multiple outcomes through the step-wise procedure described in Romano and Wolf (2005a,b, 2016).

Table B-8: Average Program Impacts by the Individual Components of the Socio-Emotional Score

	Panel A: First 16 Components															
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
API Original	0.040	-0.068	0.074	0.003	-0.008	0.026	0.072	-0.009	0.006	0.015	0.017	0.042	-0.013	-0.024	0.030	-0.020
	[0.293]	[0.041]	[0.049]	[0.943]	[0.835]	[0.477]	[0.047]	[0.818]	[0.863]	[0.679]	[0.646]	[0.205]	[0.737]	[0.410]	[0.348]	[0.563]
	{0.340}	{0.052}	{0.065}	{0.945}	{0.849}	{0.507}	{0.062}	{0.826}	{0.868}	{0.700}	{0.654}	{0.246}	{0.748}	{0.447}	{0.386}	{0.588}
	(0.989)	(0.370)	(0.409)	(1.000)	(1.000)	(0.999)	(0.393)	(1.000)	(1.000)	(1.000)	(1.000)	(0.934)	(1.000)	(0.997)	(0.994)	(0.999)
API Plus	0.125	0.058	0.057	-0.012	-0.014	0.038	0.096	-0.023	0.021	-0.007	0.055	0.056	0.047	0.061	0.040	0.003
	[0.001]	[0.136]	[0.158]	[0.773]	[0.720]	[0.317]	[0.019]	[0.584]	[0.510]	[0.870]	[0.150]	[0.113]	[0.205]	[0.057]	[0.216]	[0.937]
	{0.002}	{0.168}	{0.204}	{0.798}	{0.748}	{0.352}	{0.035}	{0.607}	{0.533}	{0.889}	{0.173}	{0.149}	{0.249}	{0.078}	{0.251}	{0.939}
	(0.010)	(0.775)	(0.813)	(0.999)	(0.999)	(0.972)	(0.157)	(0.997)	(0.995)	(0.999)	(0.809)	(0.710)	(0.901)	(0.421)	(0.908)	(0.999)
API Original = API Plus	[0.044]	[0.002]	[0.690]	[0.721]	[0.863]	[0.777]	[0.560]	[0.739]	[0.696]	[0.595]	[0.380]	[0.706]	[0.141]	[0.014]	[0.759]	[0.532]
	{0.073}	{0.003}	{0.641}	{0.758}	{0.894}	{0.812}	{0.772}	{0.795}	{0.680}	{0.637}	{0.413}	{0.796}	{0.174}	{0.024}	{0.789}	{0.580}
	(0.367)	(0.013)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(0.998)	(1.000)	(0.843)	(0.119)	(1.000)	(0.999)
Observations	1045	1045	1045	1045	1045	1045	1045	1045	1045	1045	1045	1045	1045	1045	1045	1045
Clusters	224	224	224	224	224	224	224	224	224	224	224	224	224	224	224	224
	Panel B: Second 16 Components															
	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)	(25)	(26)	(27)	(28)	(29)	(30)	(31)	(32)
API Original	-0.005	-0.050	0.015	-0.030	0.044	-0.034	0.085	-0.026	0.040	0.026	0.060	0.010	0.075	0.002	0.024	0.033
	[0.882]	[0.138]	[0.677]	[0.405]	[0.178]	[0.116]	[0.020]	[0.450]	[0.328]	[0.519]	[0.054]	[0.720]	[0.044]	[0.956]	[0.553]	[0.301]
	{0.894}	{0.159}	{0.707}	{0.448}	{0.192}	{0.143}	{0.038}	{0.491}	{0.370}	{0.564}	{0.076}	{0.730}	{0.067}	{0.967}	{0.564}	{0.345}
	(1.000)	(0.823)	(1.000)	(0.997)	(0.905)	(0.757)	(0.189)	(0.998)	(0.991)	(0.999)	(0.436)	(1.000)	(0.381)	(1.000)	(0.999)	(0.989)
API Plus	0.073	-0.009	0.091	0.021	0.040	-0.013	0.077	0.071	0.045	0.037	0.100	0.053	0.020	0.036	0.037	0.007
	[0.018]	[0.807]	[0.014]	[0.559]	[0.214]	[0.547]	[0.031]	[0.048]	[0.305]	[0.336]	[0.005]	[0.049]	[0.613]	[0.344]	[0.327]	[0.838]
	{0.028}	{0.817}	{0.028}	{0.586}	{0.245}	{0.608}	{0.045}	{0.065}	{0.353}	{0.379}	{0.009}	{0.071}	{0.647}	{0.366}	{0.383}	{0.846}
	(0.154)	(0.999)	(0.117)	(0.997)	(0.908)	(0.997)	(0.258)	(0.371)	(0.972)	(0.972)	(0.037)	(0.379)	(0.997)	(0.972)	(0.972)	(0.999)
API Original = API Plus	[0.018]	[0.246]	[0.055]	[0.191]	[0.923]	[0.350]	[0.848]	[0.012]	[0.925]	[0.796]	[0.301]	[0.193]	[0.203]	[0.422]	[0.735]	[0.494]
	{0.037}	{0.298}	{0.092}	{0.233}	{0.933}	{0.408}	{0.896}	{0.027}	{0.960}	{0.775}	{0.444}	{0.175}	{0.210}	{0.463}	{0.742}	{0.493}
	(0.146)	(0.966)	(0.432)	(0.935)	(1.000)	(0.996)	(1.000)	(0.102)	(1.000)	(1.000)	(0.989)	(0.935)	(0.937)	(0.998)	(1.000)	(0.999)
Observations	1045	1045	1045	1045	1045	1045	1045	1045	1045	1045	1045	1045	1045	1045	1045	1044
Clusters	224	224	224	224	224	224	224	224	224	224	224	224	224	224	224	224

Notes: This table shows OLS estimates and the associated p-values of the two API modalities: API *Original* and API *Plus* for 1,044 students enrolled in third to sixth grade by the end of the second school year since treatment assignment. The individual components of the socio-emotional score are indicator variables for whether the child displays one of the following emotions/behaviors: 1) has serendipitous mood changes, 2) feels or complains that nobody loves him/her, 3) is tense or nervous, 4) lies or cheats, 5) is scared or anxious, 6) talks and argues too much, 7) has difficulty in focusing on a specific activity for an extended amount of time, 8) gets easily confused, 9) it seems that his/her head is in the clouds, 10) threatens or is mean with other children, 11) tends to challenge parental authority, 12) does not feel guilty after a bad deed, 13) does not get along with other children, 14) is impulsive or acts "fast" without thinking, 15) feels has inferiority issues, 16) has no friends, 17) has difficulty letting go certain thoughts, 18) is hyper-active, 19) has a bad temper, or is irascible, 20) looses easily his/her temper, 21) feels unhappy, sad, or depressed, 22) is shy, does not socialize with others, 23) breaks objects on purpose, 24) is too attached to the adults, 25) cries too much, 26) demands a lot of attention, 27) is too much dependent on others, 28) is afraid of other people's judgement, 29) Tends to be in bad company; 30) is reserved, keeps things for himself/herself, 31) worries about every thing, 32) misbehaves at school and does not respect the instructor (see Appendix A.2). The inference procedures take into account clustering of the error terms at the school level and the block randomization design at the strata level. p-values reported in brackets refer to the conventional asymptotic inference. p-values reported in braces are computed using randomization inference (randomization-t). All p-values account for clustering at the school level. p-values reported in parentheses are adjusted for testing each null hypothesis (null impact of API *Original*, API *Plus*, and the comparison) on multiple outcomes through the stepwise procedure described in Romano and Wolf (2005a,b, 2016).

Table B-9: Average Program Impacts by the Individual Components of Parental Investments

	Engage with School				Manage School Resources					Engage with Child			
	Volunteering	Donate Cash	Donate In-Kind	Food Instructor	Manage School Resources	Propose School Material	Decide School Material	Decide Money Allocation	Evaluate School Targets	Help With Homework	Extra-Academic Activities	Meeting Teachers	Expect Upper Secondary
					Panel A: First Experiment								
API Original	0.042	0.118	0.063	0.046	-0.042	0.026	-0.009	0.002	-0.040	0.210	0.055	0.203	0.025
	[0.417]	[0.126]	[0.478]	[0.560]	[0.579]	[0.726]	[0.912]	[0.974]	[0.487]	[0.358]	[0.528]	[0.291]	[0.608]
	{0.435}	{0.147}	{0.494}	{0.566}	{0.597}	{0.734}	{0.916}	{0.971}	{0.512}	{0.382}	{0.524}	{0.322}	{0.626}
	(0.955)	(0.475)	(0.969)	(0.969)	(0.969)	(0.969)	(0.983)	(0.983)	(0.969)	(0.928)	(0.969)	(0.872)	(0.969)
Number of clusters	73	73	73	73	73	73	73	73	73	73	73	73	73
Observations	208	208	207	208	208	208	208	208	208	208	207	208	199
					Panel B: Second Experiment								
API Original	-0.031	-0.004	-0.058	-0.058	-0.029	-0.070	-0.062	-0.010	-0.027	0.222	0.074	0.043	0.010
	[0.356]	[0.894]	[0.130]	[0.042]	[0.471]	[0.095]	[0.122]	[0.772]	[0.389]	[0.027]	[0.082]	[0.568]	[0.781]
	{0.884}	{0.981}	{0.452}	{0.194}	{0.917}	{0.369}	{0.452}	{0.981}	{0.888}	{0.137}	{0.350}	{0.942}	{0.981}
	(0.377)	(0.902)	(0.155)	(0.057)	(0.488)	(0.123)	(0.153)	(0.783)	(0.422)	(0.048)	(0.117)	(0.598)	(0.791)
API Plus	0.036	0.018	0.044	0.071	0.069	0.001	0.006	0.010	0.018	0.221	0.108	0.192	0.094
	[0.289]	[0.625]	[0.329]	[0.013]	[0.095]	[0.978]	[0.890]	[0.776]	[0.570]	[0.066]	[0.015]	[0.020]	[0.019]
	{0.765}	{0.953}	{0.778}	{0.062}	{0.323}	{0.977}	{0.977}	{0.977}	{0.953}	{0.245}	{0.063}	{0.072}	{0.072}
	(0.341)	(0.666)	(0.364)	(0.024)	(0.128)	(0.977)	(0.901)	(0.791)	(0.598)	(0.105)	(0.025)	(0.037)	(0.034)
Clusters	224	224	224	224	224	224	224	223	224	224	224	223	224
Observations	1042	1042	1039	1042	1033	1036	1027	1031	1029	1044	1033	974	1017

Notes: This table shows OLS estimates and the associated p-values of the two API modalities: API *Original* and API *Plus* for 1,044 students enrolled in third to sixth grade by the end of the second school year since treatment assignment. For a detailed descriptions of the sub-components of the reading and math scores used in this table, see Appendix A.2. The outcome variables are standardized with respect to their means and the standard deviations in the control group. The inference procedures take into account clustering of the error terms at the school level and the block randomization design at the strata level. p-values reported in brackets refer to the conventional asymptotic inference. p-values reported in braces are computed using randomization inference (randomization-t). All p-values account for clustering at the school level. p-values reported in parentheses are adjusted for testing each null hypothesis (null impact of API *Original*, API *Plus*, and the comparison) on multiple outcomes through the stepwise procedure described in Romano and Wolf (2005a,b, 2016).

Table B-10: Remedial Education Sessions and Student Test Scores

	Reading Score	Math Score	Socio-Emotional Score	Overall Index
API Original× Rank≥7	0.193	0.023	0.147	0.192
	[0.105]	[0.844]	[0.313]	[0.177]
API Plus× Rank≥7	0.423	0.274	0.206	0.430
	[0.001]	[0.055]	[0.140]	[0.003]
API Original× Rank<7	0.078	0.045	0.034	0.074
	[0.431]	[0.641]	[0.728]	[0.487]
API Plus× Rank<7	0.261	0.224	0.183	0.327
	[0.011]	[0.042]	[0.082]	[0.003]
H0: Standard=Plus (<7)	[0.104]	[0.095]	[0.192]	[0.039]
H0: Original=Plus (≥7)	[0.072]	[0.081]	[0.721]	[0.144]
H0: [Original-Plus (<7)]=[Original-Plus (≥7)]	[0.766]	[0.675]	[0.639]	[0.937]
Observations	1044	1044	1045	1045
Clusters	224	224	224	224

Notes: This table shows the estimates for the API program once we interact the treatment assignment dummies with indicators of whether a child is among the six lowest-performing children in the class on the diagnostic test (Rank Below 7 and Rank Above 7), which is one of the main determinants for participation in the one-on-one remedial sessions with the mentors (see Appendix Figure B-2). Reading, math, and socio-emotional scores are standardized with respect to the mean and the standard deviation of the control group. See Appendix A.2 for a detailed description of the outcome variables. Asymptotic p-values reported in brackets are clustered at the school level.

Table B-11: Teacher Pedagogical Practices

	Learning Activities	Engage With Students	Manage Time	Use of Material	Overall Index
API Original	0.006	-0.019	0.178	-0.142	-0.040
	[0.960]	[0.903]	[0.264]	[0.388]	[0.755]
	{0.962}	{0.911}	{0.292}	{0.399}	{0.765}
	(0.982)	(0.982)	(0.556)	(0.726)	(0.969)
API Plus	-0.081	0.064	-0.030	-0.029	-0.180
	[0.555]	[0.651]	[0.843]	[0.845]	[0.169]
	{0.569}	{0.654}	{0.848}	{0.858}	{0.168}
	(0.919)	(0.919)	(0.960)	(0.960)	(0.357)
API Original = API Plus	[0.566]	[0.622]	[0.206]	[0.528]	[0.318]
	{0.583}	{0.600}	{0.248}	{0.567}	{0.348}
	(0.847)	(0.847)	(0.470)	(0.847)	(0.616)
Observations	265	265	265	265	265
Clusters	209	209	209	209	209

Notes: This table shows OLS estimates and the associated p-values of the API *Original* and the API *Plus* modalities on teachers' pedagogical practices (Stallings Classroom Snapshot). The outcome variables are standardized with respect to their means and the standard deviations in the control group. The inference procedures take into account clustering of the error terms at the school level and the block randomization design at the strata level. p-values reported in brackets refer to the conventional asymptotic inference. p-values reported in braces are computed using randomization inference (randomization-t). All p-values account for clustering at the school level. p-values reported in parentheses are adjusted for testing each null hypothesis (null impact of API *Original*, API *Plus*, and the comparison) on multiple outcomes through the stepwise procedure described in Romano and Wolf (2005a,b, 2016).

Table B-12: Placebo Test for Years of API Plus Exposure Within Experimental Schools

	Spanish		Math		Science	
1 Year	-0.049	-0.019	0.074	0.025	0.098	0.155
	[0.773]	[0.920]	[0.665]	[0.890]	[0.584]	[0.444]
2 Years	-0.020	-0.007	-0.021	-0.029	0.003	-0.073
	[0.913]	[0.971]	[0.920]	[0.896]	[0.985]	[0.715]
3 Years	-0.235	-0.131	-0.218	-0.123	-0.371	-0.328
	[0.339]	[0.618]	[0.368]	[0.626]	[0.092]	[0.176]
Observations	207	207	207	207	207	207
Controls for Criteria	No	Yes	No	Yes	No	Yes

Notes: This table shows OLS estimates and the associated p-values of the years of exposure to the API program during the transition between the second experiment and the government implementation of the *Plus* modality. For detailed descriptions of the 2013 school-average test scores used in this table as outcome variables, see Appendix A.1. All p-values account for clustering at the school level. Asymptotic p-values reported in brackets are clustered at the school level.

Table B-13: Placebo Test for API Plus Assignment During Policy Implementation

	Spanish		Math		Science	
API Plus	-0.246	-0.045	-0.231	-0.053	-0.205	-0.004
	[0.000]	[0.409]	[0.000]	[0.330]	[0.000]	[0.945]
Observations	1702	1702	1702	1702	1702	1702
Controls for Criteria	No	Yes	No	Yes	No	Yes

Notes: This table shows OLS estimates and the associated p-values of the assignment API *Plus* in the fall of 2017. For detailed descriptions of the 2013 school-average test scores used in this table as outcome variables, see Appendix A.1. All p-values account for clustering at the school level. Asymptotic p-values reported in brackets are clustered at the school level.