NBER WORKING PAPER SERIES

ENHANCING HUMAN CAPITAL IN CHILDREN: A CASE STUDY ON SCALING

Francesco Agostinelli
Ciro Avitabile
Matteo Bobba

Working Paper 31407
http://www.nber.org/papers/w31407

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2023

Ciro 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 Social (CONEVAL). Matteo 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). The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Enhancing Human Capital in Children: A Case Study on Scaling
Francesco Agostinelli, Ciro Avitabile, and Matteo Bobba
NBER Working Paper No. 31407
June 2023
JEL No. C90,C93,D02,I3,J1

ABSTRACT

This paper provides novel insights into the science of scaling by examining an educational mentoring program in Mexico. Our analysis encompasses two separate field experiments, and takes advantage of a unique opportunity to learn from the government's implementation of the program on a large scale. While the originally implemented program at scale demonstrates limited effectiveness, the introduction of a new modality with enhanced mentor training significantly improves children's outcomes. This improvement is observed in both the field experiment and the subsequent large-scale government adoption. We also find that the new program's enhanced mentor-parent interactions stimulate parental engagement at the community-school level, which emerges as a critical factor in facilitating the program's scalability.

Francesco Agostinelli
University of Pennsylvania
Perelman Center for Political Science
and Economics
133 South 36th Street
Philadelphia, PA 19104
and NBER
fagostin@upenn.edu

Matteo Bobba
Toulouse School of Economics
University of Toulouse Capitole
1, Esplanade de l'Université
Toulouse Cedex 06 31080
France
matteo.bobba@tse-fr.eu

Ciro Avitabile
The World Bank
1818 H St. N.W.
Washington, DC 20433
cavitabile@worldbank.org

1 Introduction

A key challenge in using scientific insights to inform policy decisions arises during the implementation process, where small changes between interventions translate into substantial differences in outcomes. Even when programs display large and significant effect sizes in randomized evaluations, their success in different situations is far from guaranteed (List, 2022). This is particularly evident when transitioning from a controlled research setting to real-world implementation by the government.

This paper contributes to the recent debate about the challenges to scale-up education interventions. In particular, we provide a case study involving a mentoring program that was implemented at scale in Chiapas, the poorest state in Mexico. The program assigns recent university graduates to remote and disadvantaged communities. Among other tasks, mentors help the local instructors, and encourage parental involvement in children's education through home visits. We evaluate the relative effectiveness of two program modalities that differ in terms of the content of the training provided to the front-line mentors, both within two independent field experiments as well as during the government scale up.

The mentoring program was initially launched on a large scale by the government without undergoing a rigorous evaluation. It featured a training module for mentors focused on curricular knowledge and pedagogical practices. However, subsequent evidence gathered through two independent field experiments revealed null results of this program modality. The lack of effectiveness of the program served as a catalyst, prompting the need to improve the delivering of mentoring services in the most disadvantaged communities. Our research team collaborated with the government—including accessing the existing government infrastructure of the ongoing program at scale—to embark on an experimental evaluation of a new program modality that incorporated an enhanced training protocol for mentors.

The new modality of the mentoring program encompasses a significant change in the training module, enhancing mentors' ability to effectively interact with and engage parents. Mentors attend periodic peer-to-peer meetings throughout the school year in which they share, among others, their own experiences in the local communities and design common strategies to better manage their interactions with families. These changes were motivated by the large economic literature showing 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), with ample evidence that

successful home visit and mentoring programs, in both developing and developed countries, share the common outcome of stimulating parental investment (Heckman and Mosso, 2014; Carneiro et al., 2019; Attanasio et al., 2022b). Moreover, recent findings point toward the quality of child/home-visitor interactions and parent/home-visitor interactions as key ingredients for boosting the impact of early childhoods interventions (Heckman and Zhou, 2021; Zhou et al., 2021; García and Heckman, 2023).

Science guides policy. Following the release of compelling evidence regarding the effectiveness of the new modality, which demonstrated significant positive effects on children's outcomes and increased parental investments and engagement, the government made the decision to adopt the program with the most effective approach. The original program had already been implemented on a large scale, encompassing approximately 1,300 schools and 18,000 students in the State of Chiapas alone. Our subsequent comprehensive analysis of the new program's implementation in Chiapas provides robust support for the effectiveness of the new modality at scale. Importantly, we find that parental engagement and attitudes toward schooling activities emerge as critical factors for the program's scalability, highlighting the role played by actively involved parents within the local community in promoting the scalability of educational programs.

Throughout the analysis, our empirical evidence draws from two field experiments and the subsequent government scale up of the effective program modality. The first experiment is directly carried out by the government after the national implementation of the original mentoring program. Assignment to the program is 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 assign both the original and the new 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 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 new 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 government's decision to transition the program to a more effective modality offers a valuable opportunity to investigate the factors and mechanisms influencing scaling. We combine several administrative sources of data, and we exploit the variation in the program assignment across communities in Chiapas. The assignment at scale of the program involved a scheme with a priority-based mechanism, which allows us to explore the determinants and mechanisms of scaling. We demonstrate, through a placebo test, that this variation appears to be uncorrelated with predetermined outcomes after accounting for the eligibility criteria officially employed by the government. Our results show that the mentoring program remained successful at scale. Within the localities of the experimental schools, the average impact of the new modality at scale on the fraction of children who enroll in lower-secondary education is +9.1 percentage points. For the 1,161 localities outside of the experimental sample, which include approximately 16,000 children enrolled in eligible schools, the results show a positive effect on secondary school enrollment, with an average program impact of 5.6 percentage points. There is no statistically significant difference in the estimated effects between the two samples of schools. We further document positive effects of the program on child literacy, which imply a reduction of illiteracy rates by 21 percent with respect to the sample mean for the overall sample of schools.

The effectiveness of the new program at scale was not guaranteed *a priori*, despite the positive and significant treatment effects observed in the field experiment. Existing literature highlights the importance of various "non-negotiable" aspects in the program design. Failure to account for these critical elements during the implementation of the intervention at scale can potentially diminish or even eliminate the size effects observed in the experimental estimates (Al-Ubaydli et al., 2020; Caron et al., 2021; List, 2022). While we do observe some slight changes in both the quantity and quality of mentors' activities during the scale-up phase, these estimates are generally small in magnitude and lack precision. As a result, we cannot conclusively state that the mentoring program underwent substantial changes across the two different situations.

We argue that a potential source of "voltage drop" of the program at scale is due to the fact that the design of the experiment traded off real-world applicability for the purity of the evaluation. While a significant challenge faced by educational programs in this context is the occurrence of frequent school closures, the intense monitoring during the experimental evaluation from the research team has minimized the extent of this negative event in the field experiment. To the extent that the continuity of the schooling services is critical for ensuring the program's effectiveness at scale, this particular aspect of the implementation

protocol poses a particular challenge to the ability of the field experiment to inform about the scalability of the mentoring program. We show evidence that the new program modality at scale, unlike its predecessor, drastically reduces the occurrence of school closures.

We zoom into the relationship between exposure to the mentors and school closures in order to study the sources of scalability of the program. Within the community-based schooling system under investigation, parents emerge as pivotal actors, wielding influence through their decisions and votes within the parent association. Their choices and actions directly impact crucial aspects such as resource allocation, investments, and the ultimate determination of whether the school remains open or not (Gertler et al., 2012). While the original modality of the mentoring program does not significantly affect parental investments, mentors with enhanced training are more effective in boosting parental engagement, both toward the school and directly with the child. Our measure of parenting practices increases by 0.36 standard deviations under the new program modality. After correcting inference for multiple hypotheses testing, we can reject the null hypothesis of equal treatment effects across program modalities on all four parental outcomes considered in the analysis. We further show that mentors with enhanced training significantly increase both the quantity and the quality of their periodic interactions with parents, which in turn shaped parental attitudes and behaviors toward their children's education.

Taken together, the evidence on school closures and on parental responses strongly suggests that parents can play a crucial role in the scalability of the program. We evaluate this hypothesis through an instrumental variables (IV) approach that leverages the changes in community-level parental engagement induced by the random assignment of the mentors with enhanced training. We find that an increase of 0.1 of a standard deviation in the overall parental engagement index is causally associated with a reduction of 2.2 percentage points in the probability that their children experience a school closure. This effect is both statistically and quantitatively significant. The original modality, instead, displays small and noisy effects on school closures in both experiments. This finding further reinforces the idea that community educational programs struggle to succeed in situations marked by a lack of parental engagement.

Qualitative data obtained from in-depth surveys of mentors and local instructors provide additional support for the pivotal role of parents in ensuring the continuity of educational activities within communities, particularly in contexts with inadequate school infrastructure and frequent disruptions in schooling activities. Collectively, the quantitative and qualitative

evidence strongly indicate the crucial role of parents in preventing school closures and consequently enhancing the effectiveness of the mentoring intervention during the government implementation.

In recent years, there has been increasing concern among scholars and policymakers regarding the effectiveness of field experiments in informing policy decisions. This concern stems from the challenges of replicating the effects observed in small-scale randomized trials when interventions are implemented at a larger scale (Bold et al., 2018; Cameron et al., 2019; Muralidharan and Singh, 2020; Bobba et al., 2023). Our empirical analysis builds upon the insights from recent studies that employ at-scale randomized designs (Egger et al., 2022; Banerjee et al., 2023; Muralidharan et al., 2023), allowing us to contribute to the ongoing debate on the challenges of scaling up experimental evaluations. We highlight the informative features of our experimental design, addressing the key threats identified in Al-Ubaydli et al. (2020). Firstly, we leverage the value of replication by conducting two independent field experiments on different and representative samples of schools (Maniadis et al., 2014; Allcott, 2015; Davis et al., 2021). Drawing joint inferences from these experiments enhances the robustness and generalizability of our findings. Secondly, the field experiments were conducted while the original program was already being implemented at scale, in close collaboration with the government agency responsible for the subsequent scale-up of the new modality (Muralidharan and Niehaus, 2017). This collaborative approach guarantees the harmonization of our research study with the practical considerations and implementation realities on a larger scale. In particular, the design of the new modality was a joint effort between the government agency and our research team, taking into account the set financial and human resource constraints specific to the context under study (Banerjee et al., 2017). Lastly, our randomization was implemented at a relatively large unit level, encompassing schools and communities. This research design accounts for 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., 2023).

Our findings align with the perspective that human capital accumulation is inherently a socially determined outcome (Coleman, 1988), emphasizing the significance of the local community in determining the success of education interventions implemented at scale (List et al., 2023). The implications of our results are pertinent to policy discussions and future research aimed at designing mentoring and home-visiting interventions in disadvantaged contexts. While parents within local communities are readily available without supply-side constraints, it is crucial not to overlook their beliefs and attitudes toward schooling activities.

The specific details of the training protocol and the resulting effectiveness of mentor-parent interactions play a pivotal role in shaping parental responses, which have been demonstrated to be critical for the scalability of education interventions. By recognizing the importance of engaging parents, policymakers and practitioners can enhance the design and implementation of educational interventions in underprivileged settings. This acknowledgement opens avenues for further exploration and investigation into optimizing the impact of education interventions by fostering meaningful connections between mentors, parents, and the local community.

2 Context and Data

In this section, we delve into the study's context and present a concise overview of the diverse datasets we have collected for the empirical analysis. Our study focuses on a mentoring program implemented in the Mexican state of Chiapas, serving as a compelling case study to uncover novel insights on the science of scaling. Two independent field experiments were conducted to assess the intervention's effectiveness a few years after its widespread implementation by the government.

Building upon the experimental evidence, the government made a crucial decision to replace the original program with a new modality that incorporates a significant change in the training module provided to the mentors. The government leveraged the existing infrastructure used for the Original program, including the pool of mentors who were already employed and personnel responsible for program operations, and adapted it to the more effective Plus modality.

2.1 The Mentoring Program

The *Consejo Nacional de Fomento Educativo* (CONAFE) is a government agency responsible for providing schooling services in rural and highly marginalized communities of Mexico with a population below 2,500 inhabitants. In 2013, these schools accounted for 10 percent of the roughly 99,000 primary schools across the 31 Mexican states. The largest presence of CONAFE schools is in Chiapas, the Mexican state with the highest incidence of poverty in the country (CONEVAL, 2018). CONAFE primary schools typically have a single multigrade classroom with on average 15 students. Hereafter, we will refer to the population of

CONAFE primary schools as schools.

The local instructors predominantly consist of community residents aged between 15 and 29 years old, who typically have minimal to no formal training as teachers. 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. Furthermore, schools frequently face closure due to similar challenges. In fact, the average yearly rate of school closures in Chiapas stands at 11 percent. 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 vital role in the decision-making process to ensure the continuation of school operations.

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. 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. Prior to start working as mentors, selected applicants receive a week-long training session focused on curricular knowledge and basic notions of pedagogy. Schools receive mentors for a two-year period. The assignment of the mentors follows a priority-based mechanism that depends on 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 conduct periodic home visits to update parents on their children's progress in school and encourage their active involvement in school activities. In addition to addressing behavioral issues directly with the children, mentors are expected to discuss these concerns with parents during home visits. Each mentor is responsible for organizing individual remedial education sessions at school, which are held after regular instructional hours. The tutoring sessions are offered to the six weakest students in the class, identified through a diagnostic evaluation conducted at the beginning of the school year and an additional exam administered by the mentor. During regular school hours, mentors are tasked with observing and

taking notes on the teaching practices of community instructors. They also assist students with learning difficulties and provide support outside the classroom for those unable to attend the afternoon remedial sessions. Mentors hold meetings with their supervisors every two months in two-day sessions throughout the school year. Henceforth, we will refer to this program format as the API *Original*.

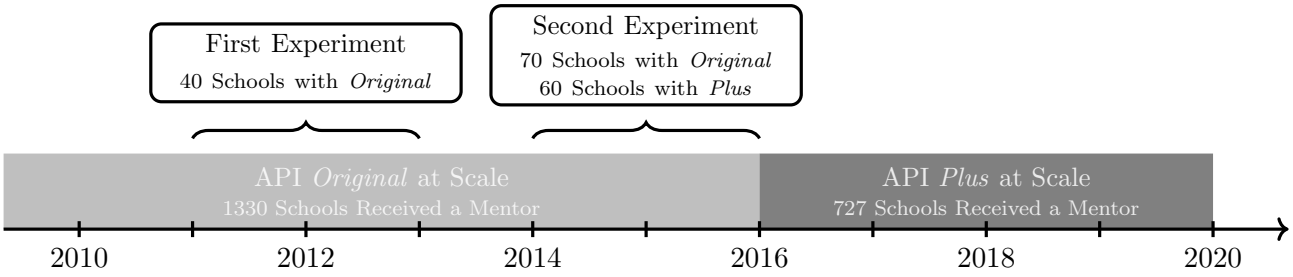
The API *Plus* modality incorporates all the features of the API *Original*, with two significant changes in the training module. Firstly, it includes two weeks of initial training instead of one. The additional week is dedicated to hands-on strategies aimed at improving students' reading and math skills. Secondly, mentors attend an extra day during each bimonthly meeting throughout the school year. This additional day is dedicated to peer-to-peer sessions, where mentors can share experiences and develop common strategies to enhance the quality of their interactions with parents in the local communities. The decision to innovate the program's modality was influenced by extensive economic literature, which suggests that successful mentoring programs in similarly disadvantaged contexts have a shared design feature of fostering parental engagement (Heckman and Mosso, 2014; Attanasio et al., 2022b; García and Heckman, 2023). Notably, the revised training module considers the diverse range of financial and human resources constraints that the government would face at scale. The cost of the API *Plus* is US $332 per child, compared to US $285 per child for the API *Original*. These cost figures align closely with those of another recent government-run program in Colombia, which targets both children and parents (Attanasio et al., 2022a).

2.2 Two Independent Evaluations of the API Program

While the original version of the mentoring program was enacted by the government without prior rigorous evaluations of its effectiveness, two subsequent and independent randomized evaluations took place in the midst of the nation-wide implementation. The first experiment was directly carried out by the government. We design and implement the second experiment in close collaboration with the government, leveraging the existing program's infrastructure. In particular, our research team gains the privilege of utilizing the existing stock of mentors employed by the government under the original modality to draw samples for our experiment. This approach encompasses the adoption of identical mentor recruitment and assignment processes across both experimental and non-experimental schools.

Figure 1 depicts the timeline of the mentoring program in the State of Chiapas, whereby the API *Original* served over 1,300 schools and approximately 18,000 students between 2009

Figure 1: Timeline of the Mentoring Program in Chiapas



and 2016, while approximately 700 schools and 10,000 students received a mentor after the subsequent conversion to the API *Plus*. Below, we discuss the design of each experiment in some detail.

First Experiment. Eighty program-eligible primary schools are selected among those that never received the mentoring program before. Of those, 62 schools are located in the state of Chiapas and the remaining 18 schools are in the three States of Hidalgo, Queretaro, and Veracruz. Assignment to the mentoring program is randomized at the school level using a block design, with the strata represented by the Mexican states where schools are located. Forty schools are assigned to receive the API *Original* starting from the 2011–2012 school year while the remaining half of the schools are assigned to the control group without mentors.

Student outcomes are measured two school-years after the assignment of the API program through the performance in the national standardized test for students in grades three through six. A mid-line survey records parental behaviors and investments for 208 parents in 73 schools (the enumerators were not able to reach the parents in seven schools). Due to the incomplete take-up of the standardized achievement test–mainly due to the opposition from the teachers' unions in some states–we are able to match 70 schools with 599 test score records out of the sub-sample of 73 schools with parental outcomes. Out of the ten schools that were part of the experimental sample and we are unable to match in our final sample, five schools are in the treatment group and five are in the control group. Table B-1 shows balance with respect to the assignment of the mentor for school and community characteristics measured in the year before the start of the first experiment.

Second Experiment. 230 program-eligible primary schools are selected in rural Chiapas among those that never received the mentoring program before. Assignment of the mentors is 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. As a result, 60 schools are assigned to receive API-*Plus* mentors starting from

the 2014-2015 school year, 70 schools are assigned to receive API-*Original* mentors over the same time period, and the remaining 100 schools are in the control group with no mentors.

We draw on a rich combination of administrative and survey data sources, along with qualitative interviews with instructors and mentors (see Appendix A for more details). The data collection took place by the end of the second school year after the inception of the mentoring 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. Out of the six schools that dropped out of the sample, two schools are in the control group, two are in the *Original* group, and two in the *Plus* group. The number of schools part of the second experiment is 224. Table B-2 shows that a large array of pre-determined covariates of schools, teachers, children, households, and mentors is balanced with respect to the assignment of both API *Original* and API *Plus*. The household module of the survey is 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 through unique student identifiers. The final sample consists of 1,045 children.

2.3 The Scale-up of the API *Plus* Program

After learning about the results of the second experiment (see Section 3), the government decided to replace the API *Original* program with the enhanced training modality. All its primary schools, including those that were part of the evaluation samples of the two field experiments, were deemed eligible to receive the API *Plus* program modality. This unique policy change creates two interesting circumstances that are informative for our case study on scaling. The schools that received the API *Plus* within the second experiment experienced a change in the situation–from the research setting to the government implementation–under the same program modality. The rest of the schools, that were not part of the experiment but received the mentoring program under the API *Original* modality at scale, underwent a reform in program design within the same government situation.

We conducted our empirical analysis specifically on the State of Chiapas, which hosts the majority of the schools that participated in the first randomized experiment, as well as all the schools involved in the second experiment. 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. In Table 1 we compare means in observable characteristics between the overall population

Table 1: Differences Across Populations

	All Chiapas Mean (SD)	First Experiment Mean (SD)	Second Experiment Mean (SD)	Chiapas vs. Experiment 1 Mean Difference [p-value]	Chiapas vs. Experiment 2 Mean Difference [p-value]
			Panel A: School Characteristics		
Average Test Score (Spanish)	424.503	399.116	431.340	-25.387	6.837
	(56.466)	(32.631)	(60.810)	[0.000]	[0.139]
Average Test Score (Math)	414.921	379.165	421.333	-35.756	6.412
	(75.300)	(45.339)	(80.895)	[0.000]	[0.297]
Number of Students	14.049	15.507	15.009	1.458	0.960
	(8.468)	(8.781)	(6.053)	[0.175]	[0.037]
Number of Teachers	1.231	1.333	1.217	0.102	-0.014
	(0.467)	(0.505)	(0.413)	[0.099]	[0.638]
Share Over-aged Students	0.349	0.230	0.324	-0.119	-0.025
	(0.797)	(0.552)	(0.659)	[0.088]	[0.610]
			Panel B: Locality Characteristics		
Total Population	118.758	247.280	121.389	128.522	2.630
	(221.648)	(549.923)	(240.562)	[0.043]	[0.879]
Rate of Extreme Poverty	0.490	0.486	0.473	-0.004	-0.017
	(0.500)	(0.503)	(0.500)	[0.949]	[0.644]
Incidence of Social Conflicts	0.190	0.150	0.187	-0.040	-0.003
	(0.392)	(0.359)	(0.391)	[0.335]	[0.919]
Rate of Illiteracy	0.313	0.321	0.295	0.008	-0.018
	(0.160)	(0.157)	(0.153)	[0.662]	[0.127]
Labor Force Participation	0.297	0.289	0.303	-0.008	0.006
	(0.076)	(0.071)	(0.070)	[0.352]	[0.259]
Locality Access without Road	0.216	0.203	0.179	-0.013	-0.037
	(0.411)	(0.404)	(0.384)	[0.777]	[0.181]
Water Network (Y/N)	0.028	0.050	0.022	0.022	-0.006
	(0.164)	(0.219)	(0.146)	[0.365]	[0.578]
Sewage System (Y/N)	0.011	0.038	0.009	0.026	-0.002
	(0.105)	(0.191)	(0.093)	[0.219]	[0.712]
Garbage Collection (Y/N)	0.022	0.038	0.022	0.016	0.000
	(0.146)	(0.191)	(0.146)	[0.463]	[0.994]
Number of Schools/localities	1,523	80	230	1,603	1,753

Notes: The first three columns show means and standard deviations in parentheses for various characteristics collected before the introduction of the API program. The last two columns show asymptotic p-values for mean differences between the overall population and the experimental samples after adjusting for Strata fixed effects. Panel A shows school-level variables from the school census (2010) whereas Panel B displays community-level characteristics from the population census (2010). See Appendix A.1 for more details on the data sources.

of schools in the state of Chiapas and both experimental samples. The students enrolled in the schools of the first experiment tend to perform worse in the national standardized tests (Spanish and Math) when compared to the overall population of students. Also, schools in the first experiment are located in larger localities in terms of population size.[1] As shown in the fifth column, instead, we cannot reject equal means across the several variables assessed between the sample of schools of the second experiment and the overall population of schools in Chiapas. There is only a small imbalance in the number of enrolled students (see Panel A in Table 1).

Upon reviewing the evidence presented in Table 1, it becomes evident that the sample of schools in the initial government-led experiment may not offer a comprehensive representation of the intervention's target population in Chiapas. This finding emphasizes the impor-

[1]Mean differences and the corresponding p-values presented in Table 1 have been adjusted for Strata fixed effects. This adjustment accounts for the presence of 18 schools in the first experiment out of a total of 80 schools that are situated in different Mexican States other than Chiapas.

tance of conducting a second field experiment to assess both mentoring program modalities, namely API *Original* and API *Plus*, within a sample of schools that accurately represents the targeted population of the program at scale.[2]

As shown in the last row of Table 1, there are 1,523 schools in Chiapas that are potentially eligible to receive the mentoring program. Of those, we are able to match 1,345 schools (88 percent) with the population census (2020) containing village-level educational outcomes for the quasi-universe of the schools and the localities in Mexico. 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.[3] We cannot reject the hypothesis that the probability of being unmatched is balanced with respect to the assignment of the API *Plus* at scale (p-value=0.634). Furthermore, it is worth emphasizing that the sample of localities (and their corresponding schools) that are matched with the population Census maintain their representativeness in terms of observable characteristics in relation to the overall targeted population in Chiapas. This implies that the matched sample accurately reflects the broader population of interest in Chiapas, ensuring the validity of the findings and their applicability to the scaling analysis (Table B-3). The schools in the matched sample serve approximately 19,000 students, with a total of 165,000 people living in the surrounding communities.

3 The Impacts of the Mentoring Programs on Children

In this section, we assess the impact of two different mentoring program modalities on various measures of children's outcomes. We provide empirical evidence supporting the ineffectiveness of the API *Original* by analyzing the results of two independent field experiments together. Subsequently, we quantify the positive effects of the API *Plus* within the experimental setting and for a broader sample of program-eligible schools in Chiapas during the government's implementation at scale. This larger sample of schools includes the experimental schools that underwent a change in situation between the field experiment and the government's program implementation due to its conversion.

[2] Heckman (1992) discusses selection into field experiments and finds that the characteristics of subjects who participate in a job training program in the US can be distinctly different from those of subjects who do not participate. Allcott (2015); Davis et al. (2021) document evidence of positive selection of eligible participants in experimental evaluations.

[3] 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 May, 2023.

3.1 Empirical Strategies

We analyze the two experiments through separate regression models on the treatment assignment indicators for the API *Original* and the API *Plus* modality after two years of exposure to the mentoring program. An indicator for whether or not the child speaks an indigenous language is the only covariate that is not balanced across treatment arms in the second experiment (see Panel B in Table B-2). For this reason, we include the indicator for indigenous language in the regression analysis of the second experiment. All models further include the strata control variables that account for the block randomization designs, as well as student's age and gender, which are predictive of education outcomes. During the data collection in the second experiment, a few schools had to be surveyed on a second or third visit due to adverse weather conditions or high political instability. The inclusion of survey weeks and survey routes indicators is meant to control for the different timing of the survey in these communities. The error terms are clustered at the school level, which represents the unit of randomization in both field experiments.

To expand our analysis, we extend our focus to encompass the entire population of program-eligible schools within the state of Chiapas. Our objective is to investigate whether the API *Plus* modality of the mentoring program, implemented on a larger scale by the government, has effectively enhanced educational opportunities for children in these disadvantaged communities. We analyze the impact of the API *Plus* at scale using the following linear regression model:

$$Y_j = \alpha_0 + \alpha_1 Plus_j + \boldsymbol{\delta'}\, \boldsymbol{X}_j + \epsilon_j \,, \tag{1}$$

where Y_j is a locality-level outcome on children's education attainment for locality j, while $Plus_j$ takes a value of one if the school in locality j receives a mentor during the government implementation of the *Plus* modality, and zero otherwise. The vector $\boldsymbol{X}_j$ consists of the four criteria used to determine the differential priority across eligible localities/schools to receive the mentors (see Section 2.1). Furthermore, we control for the number of hostile events related to land property, religion, elections, crime, or drug addiction as reported at the locality level in the population census (2010) as well as an indicator variable for prior exposure to the API *Original* modality as additional determinants of the assignment of the mentors across localities. The parameter of interest, α_1, represents the effect of the program during the government implementation on the outcome of interest.

We operate under the underlying assumption that the assignment of the program at scale across localities is *conditionally* random, once we control for the criteria determining the priority of program assignments. In other words, after conditioning on the assignment criteria and the other covariates in equation (1), schools/localities that receive and do not receive the API *Plus* program at scale are assumed to be similar in terms of unobserved characteristics. We run some placebo tests to bolster the credibility of this assumption using the school-level standardized achievement test scores collected before the conversion of the mentoring program under the *Plus* modality. Table B-4 shows the results. The assignment of the mentoring program at scale is not unconditionally random (odd columns of the table), as priority is given to more disadvantaged communities. Instead, when we control for the vector $\boldsymbol{X}_j$, the estimated coefficients displayed in the even columns of Table B-4 are very small and statistically insignificant.

In our analysis, we go beyond the conventional asymptotic inference by employing three additional procedures. Firstly, we present p-values based on randomization inference, which offer accurate results even when dealing with a limited number of clusters. This approach is particularly relevant for the first experiment, where the number of schools per treatment arm was smaller compared to the second experiment. Secondly, given the extensive range of hypotheses explored throughout our analysis, we also provide adjusted p-values that account for multiple hypothesis testing across various outcome families (List et al., 2019). Thirdly, building upon the insights in Maniadis et al. (2014), we leverage the value of conducting two independent evaluations within the same program environment. To test hypotheses across both experiments, we employ Fisher's combined probability test, akin to the joint statistical significance test commonly used in meta-analyses.[4]

3.2 Experimental Evidence on API *Original*

Table 2 and the first row of Table 3 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 2 are based on administrative records of third to sixth graders in a national standardized test. For the second experiment, we collect our own

[4]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.

Table 2: 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 Schools	70	70	70	70
Number of 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 standard errors. 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 table through the step-wise procedure described in Romano and Wolf (2005a,b, 2016). All p-values account for clustering at the school level.

measures of cognitive and socio-emotional skills (first to fourth columns of Table 3), as the national standardized test was terminated in 2014.

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 2)—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.[5] 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 standard deviations—a non-negligible effect size that is nonetheless not statistically different from zero (p-value=0.23, after adjusting for multiple hypotheses testing). The effect of the *Original* modality of the mentoring program on the transition rates to lower secondary school

[5]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 3: Children's Achievement and Attainment—Second Experiment

	Survey-Based Test Scores				Admin Records
	Reading	Math	Socio-emotional	Overall Index	Enroll Secondary
API *Original*	0.126	0.056	0.071	0.124	0.073
	[0.104]	[0.455]	[0.418]	[0.187]	[0.255]
	{0.138}	{0.483}	{0.440}	{0.218}	{0.283}
	(0.147)	(0.554)	(0.554)	(0.234)	(0.312)
API *Plus*	0.315	0.237	0.199	0.366	0.124
	[0.001]	[0.008]	[0.022]	[0.001]	[0.074]
	{0.001}	{0.012}	{0.030}	{0.001}	{0.084}
	(0.001)	(0.005)	(0.011)	(0.001)	(0.032)
Original = *Plus*	[0.043]	[0.043]	[0.178]	[0.020]	[0.469]
	{0.086}	{0.115}	{0.225}	{0.024}	{0.570}
	(0.045)	(0.045)	(0.098)	(0.023)	(0.376)
Number of Schools	224	224	224	224	182
Number of Observations	1044	1044	1045	1045	468

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 standard errors. 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.

are shown in the last column of Table 3. The estimated effect size is noisy, with an increase of seven percentage points out of a basis of 62 percent enrollment rate in seventh grade in the control group.

The evidence consistently reveals that a lack of statistical significance for the effect of the API *Original* modality may be indicative of a null result. The test statistic of the joint hypothesis of no effect across both experiments for the overall indices of student achievement has a p-value=0.460. This specific mentoring approach has not demonstrated substantial improvements in children's educational outcomes. Taken together, these findings give rise to concerns about the potential impact and effectiveness of the *Original* mentoring program, which had already been implemented on a larger scale by the government.

3.3 Experimental Evidence on API *Plus*

We next turn to discuss the evidence on the effectiveness of the API *Plus* modality of the mentoring program. The second row of Table 3 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 receive the *Plus* modality increase their reading scores by 0.32 standard deviations (p-values $\leq$ 0.001). Quantitatively, the API *Plus* effect is approximately 2.5 times higher than the effect of the API *Original*. We find similar patterns when we look at math scores (second column), which show a sizable and highly significant effect of the *Plus* modality with an estimated treatment effect of 0.24 standard deviations. The API *Plus* program also generates a sizable improvement in the socio-emotional score of 0.2 standard deviations (third column). 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 displayed in the fourth column of Table 3 is very large, 0.37 standard deviations—precisely estimated (p-values $\leq$ 0.001), and statistically different at the 95 percent level from the effect of the *Original* modality.[6]

The last column in Table 3 reports the estimated effects on the average transition rate to secondary school. We use separate administrative data on students' records to construct an indicator for enrollment in seventh grade, which is the first grade in lower secondary school. 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. The sample reduces to 468 sixth graders in 182 schools, which is due to the multi-grade aspect of the schooling system where student composition among grades in each school is not homogeneous in size. The choice of this cohort of students is meant to maintain the same length of exposure to the mentoring program of the sample of children in the first four columns of Table 3.[7] 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 API *Plus* modality increases the probability of a child's enrolling in seventh grade by 12 percentage

[6]In Table B-5 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). For math scores, the *Plus* modality seems particularly effective on numbers' identification and discrimination as well as additions. Similarly, in Table B-6 we report the effects of the two program modalities for each individual component of the socio-emotional score.

[7]The distribution of missing schools in the analysis of transition to secondary school is 18 schools in the control group, 14 in the API *Original* and 16 in the API *Plus*. Due to the different individual identifiers, we are not able to match this dataset to the survey data. The estimates reported in Table B-7 document no program effects on grade repetition and attrition, which suggest that conditioning on grade attainment is not problematic in our context.

points. Although a bit noisier than the test score estimates (p-value=0.032, after adjusting for multiple hypotheses testing), this effect on education attainment 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.

The inference drawn from both field experiments seems to convincingly point toward the relative effectiveness of the API *Plus* modality of the mentoring program when compared to both the API *Original* modality and the control group with no mentors. The test statistic for the joint hypothesis of no difference between the *Plus* and the *Original* modalities across the two experiments for the overall indices of student achievement has a p-value=0.010.

3.4 API *Plus* at Scale

We finally investigate the extent to which the positive effects of the API *Plus* modality of the mentoring program on children's outcomes can be sustained at a larger scale. As discussed in Section 2, the government converted the mentoring program from the API *Original* to the API *Plus* modality. We leverage administrative records detailing the government's program conversion in Chiapas under the API *Plus* modality and match this information with the quasi-universe of schools observed in the 2020 population census (data collection in the Fall of 2019). We specifically focus on the 2017-2018 school year in our analysis to align with the duration of the API program's implementation during the second experiment, spanning two complete school years. Moreover, this approach enables us to assess the program's efficacy under the *Plus* modality after sufficient time for program operations to fully adapt, considering that the transition to this new modality commenced shortly before the start of the 2016-2017 school year. During the 2017-2018 school year, 351 schools received the mentoring program out of a total of 1,345 eligible schools (see Section 2.3). This sample includes 184 schools that were previously involved in the second experiment, of which 86 were assigned the API-*Plus* mentors during the government implementation of the program. The remaining 1,161 are defined as non-experimental schools.

We leverage two village-level educational outcomes from the Census data: (i) the rate of lower-secondary enrollment among children between twelve and fourteen years old and (ii) the rate of child literacy for children between eight and fourteen years old. Secondary school is a critical period for the educational outcomes of the disadvantaged population under study, as more then a quarter of the children aged 12 to 14 in Chiapas are out of school. Likewise, 13 percent of school-aged children are still illiterate. Unlike other school-survey-based or

Table 4: Children's Attainment—API *Plus* Scale-up

	Non-Experimental Schools		Experimental Schools	
	Enroll Secondary	Child Literacy	Enroll Secondary	Child Literacy
API *Plus*	0.056	0.028	0.091	0.035
	[0.010]	[0.012]	[0.022]	[0.078]
	{0.013}	{0.013}	{0.022}	{0.068}
	(0.013)	(0.013)	(0.035)	(0.054)
Number of Schools	1161	1161	184	184

Notes: This table shows OLS estimates and the associated robust p-values on locality-level outcomes measured after two years of exposure to the API *Plus* modality of the mentoring program under the government implementation. For detailed descriptions of the outcome variables used in this table, see Appendix A.1. Control variables include indicators for the whether or not the locality satisfy the program assignment criteria, an indicator variable for prior exposure to the API *Original* modality, and the number of hostile event related to land property, religion, elections, crime, or drug addiction as reported at the locality level in the population census (2010). p-values reported in brackets refer to the conventional asymptotic standard errors. 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 Plus for the two different sub-samples of schools (non-experimental and experimental) through the step-wise procedure described in Romano and Wolf (2005a,b, 2016).

administrative test scores, these outcomes are not subject to any censoring due to school closures. This allows us to avoid the concerns about sample selection and survivorship bias, due to differential school closures induced by the program at scale (see Section 4.2).

Table 4 shows the estimation results. For the sample of schools that did not previously participate in the second experiment (Non-Experimental Schools), we find that the program increases the fraction of children who enroll in secondary education by 5.6 percentage points (p-value = 0.013, after adjusting for multiple hypotheses testing), which represents an increase of 7.6 percent with respect to the sample mean. For the schools that were previously part of the experiment (Experimental Schools), the impact of receiving the program during the government implementation is larger (+9.1p.p., p-value = 0.035, see third column in Table 4), 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 3).[8] 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. The cumulative effect of three consecutive years of exposure to the API *Plus* implies that the secondary school enrollment rates in these disadvantaged and rural areas would catch up with the enrollment rates in urban Mexico (see Figure B-2).

[8]The school enrollment variable reported in Table 4 is not immediately comparable with our previous measure of enrollment in seventh grade (see Table 3), which draws from administrative schooling records. 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.

The estimates of the impact of API *Plus* at scale on child literacy are displayed in the even columns of Table 4. After two years of exposure, we find that villages that received mentors under the *Plus* modality at scale display a 2.8 percentage points (p-value = 0.013) increase in child literacy rates when compared to villages without mentors. The magnitude of this effect implies a reduction of illiteracy rates by 21 percent with respect to the sample average. The estimated program effect for the subsample of experimental schools is quantitatively similar, although a bit noisier (+3.5 percentage points, p-value = 0.054). Overall, our results support the notion that the API *Plus* modality of the mentoring program as implemented at scale by the government has effectively enhanced the education attainment for children in these disadvantaged communities.

4 The Threat of Voltage Drop in the New Situation

Despite the significant impact of the mentoring program on supporting students and improving their educational outcomes, there are potential risks associated with the government's conversion of infrastructure for the large-scale implementation of the *Plus* modality. The literature (Al-Ubaydli et al., 2020) discusses various mechanisms that can cause a voltage drop. In this section, we outline specific "non-negotiable" aspects in the implementation protocol of the mentoring program that may have led to contrasting outcomes between the experimental phase and the subsequent government implementation at scale.[9]

While our unique case study provides us with an opportunity to examine the challenges and determinants of scaling in the context of the change in situation (List, 2022), our analysis does not address the "vertical" aspects of scaling. Specifically, we do not address the challenges that arise when implementing the program at scale without an existing intervention that is already in place. In such cases, the program's implementation requires the creation of a large-scale infrastructure from scratch. In our context, this would entail recruiting a significant number of new mentors and personnel responsible for program operations, as

[9]There may be other "negotiable" differences in the program implementation across the experimental and the scale-up regimes 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.

well as developing organizational capital. Our findings do not delve into these challenges but rather focus on the specific issues arising from the change in situation between field experiments and government operations.

We focus on two main mechanisms that can cause deterioration in the quality of service provided by mentors during the government implementation. First, the fidelity of the training and supervision might fall at scale even when scaling-up does not require hiring and training an increased number of service providers. In our case, some of the scalability concerns had been addressed at the research design stage. For instance, in the second experiment, the training intensity of the Plus modality was carefully tailored, taking into consideration the incentives and constraints that would be relevant for a larger-scale implementation. Likewise, adjustments made to the supervision intensity in the second experiment were aligned with the program's capacity when implemented on a broader scale. We use available information to study the quantity and quality of mentoring at scale, which can ultimately impact the overall effectiveness of the program.

Next, we explore the potential bottleneck for the program of widespread school closures, a frequently encountered issue in the local provision of education services in Mexico.[10] A functional school environment plays a crucial role in the effective implementation of educational mentoring programs within communities. Therefore, when schools close, it poses a significant challenge to the program's success on a larger scale. This potential bottleneck stands in contrast to the experimental setting, where the research team's continuous monitoring helped alleviate such issues.

4.1 Quantity and Quality of the Mentoring Service at Scale

The effectiveness of the mentoring program at scale, when the government is in charge for the implementation, can influence the quantity and the quality of the service provided. The technology of the implementation during the government conversion of the *Plus* modality can deviate from the experimental conditions due to differences in the screening and the training of the mentors, as well as in the level of support and supervision they receive.

We begin by examining the extent to which the population of mentors is similar between the experiment and the scale-up. The second experiment was conducted within the existing

[10]This issue is not unique to our context; it is also prevalent in other settings, including public-school systems in the US (Engberg et al., 2012) and Europe (Haan et al., 2016). The challenge of widespread school closures extends beyond our program and is experienced by educational systems in various regions.

government infrastructure of the program, including but not limited to the large pool of available mentors that were recruited by the government. However, there were two minor differences in terms of how mentors were recruited and assigned to communities when compared to the API *Original* at scale. First, the most important criterion for the assignment of the mentors was the ability to speak the main indigenous language in the community. Second, supervisors of the mentors received a salary increase in exchange for an obligatory increase in the frequency of their visits to the targeted communities. These changes were supposed to be part of the new protocol of the government scale-up of the API *Plus* modality.

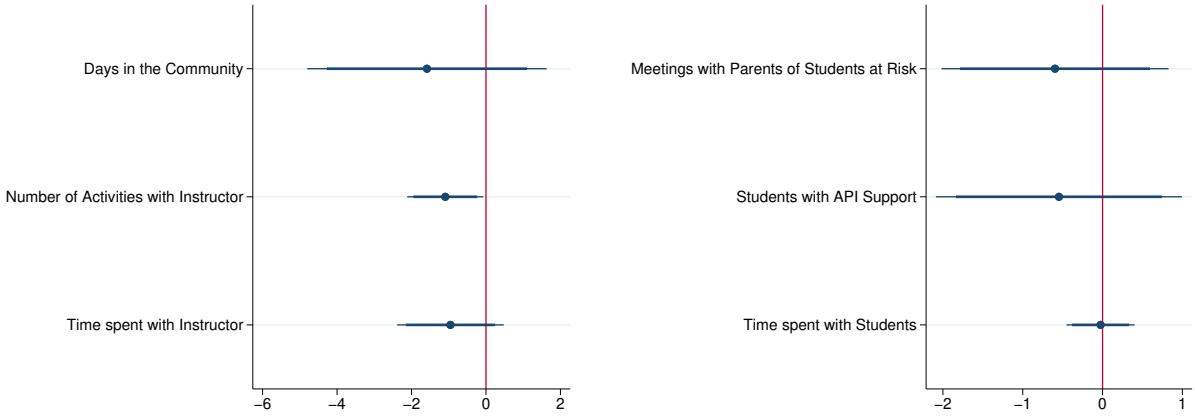
In order to directly test for differences in observed mentor characteristics across situations, we integrated the 2016 survey data on mentors from the second experiment with the administrative data of mentors during the scale up. The 2016 survey data comprises a total of 139 mentors, while the administrative data of mentors at scale includes 441 mentors.[11] Despite limited set of common variables across these two datasets, Table B-8 demonstrates that the observable traits of mentors in our experiment are similar to those of mentors in the program's scale-up. Gender, age, and the percentage of mentors who speak an indigenous language are evenly distributed across settings, which confirms our hypothesis that the recruitment practices used during the program's scale-up were consistent with those used in the experiment.

The presence of similar populations of mentors across different situations does not necessarily imply consistency in mentoring practices. Differences in incentive structures and training protocols between the government implementation and the field experiment could potentially impact both the quantity and quality of the mentoring service. To examine this, we leverage survey data in our experimental schools on various topics related to the schooling environment, with a specific focus on the activities of mentors. We use two survey rounds that record instructor-reported measures of mentoring practices from 56 and 58 schools, respectively, that were part of the *Plus* program (see Appendix A.2 for further details on the surveys). With this data, we sought to test the hypothesis that mentoring practices underwent significant changes during the government's scale-up.

The information presented in Figure 2 provides a comparison of the results from two surveys conducted in 2016 and 2018. During this period, the government fully converted the

[11]The number of mentors exceeds the number of schools because the survey included both mentors that were assigned to schools, as well as those who were awaiting a role. In the 2016 survey, for instance, the 139 mentors were either assigned to 107 unique schools included in the survey, or they were currently awaiting a role within the program.

Figure 2: Mentors Community Engagement



Panel A: Quantity

Panel B: Quality

Notes: The figure shows the comparison in the quantity and quality of API *Plus* program between the second experiment and the government implementation. This information is collected during the surveys of the local instructors, in the school years 2015-2016 and 2018-2019. Each dot in the figure represents an OLS estimate for the difference in the mentoring services across the two situations, whereas the horizontal bars are the associated 90% and 95% confidence intervals. The associated table with the OLS estimates, p-values, and number of observations are also reported in Table B-9. All the regressions include the same set of controls as in Table 4.

mentoring program into its *Plus* modality. The displayed estimates represent the difference in means between the two survey periods, and relative inference, whereby the first period denotes the experimental setting and the second period denotes the scale-up regime. Panel A of the figure examines the quantity aspect of the mentoring service in more detail. Overall, the point estimates are negative, but generally small and noisy. The first variable shown in this panel is the number of days that mentors spent in the community during their last visit. The coefficient for this variable is -1.58, which suggests that, on average, during the government implementation mentors spent 1.5 fewer days in the communities (of the 14-day visit) compared to the experimental setting. The second variable of Panel A is the number of activities (ranging from zero to five) that the mentor carries out with the local instructor in the current school year.[12] We observed that mentors, in comparison to the field experiment, decrease the number of pedagogical training activities provided to teachers by approximately one in the current school year. The third variable indicates a decrease in the amount of time mentors spend with local instructors across the two scenarios. Specifically, mentors spend

[12]This measure represents the total number of activities that are completed by the mentor out of the following five: (i) talking with students about the school and their families; (ii) going over the diagnostic tests to students; (iii) explaining the pedagogical practices to the teachers; (iv) explaining to the teachers what to do to improve the performance of their classroom; and (v), supporting the teacher in the creation of the classroom materials.

one minute less during their last visit to the community. In two out of three cases we cannot reject the null hypothesis of zero effect at conventional levels of significance.

In terms of the quality of the mentoring programs, our results also show a small and statistically insignificant reduction in our observed measures between the field experiment and the government setting. The estimates of the mean differences across situations are shown in Panel B of Figure 2. Both the number of meetings with parents of under-performing students (-0.60) and the number of students benefiting from the mentor support (-0.55) decreased during the mentor's most recent visit to the community. Finally, when considering the time that mentors spent with children during the last visit, our results suggest no change in mentoring practices. Mentors spend the same amount of time (minutes) with students both in the field experiment compared with the scale-up regime.

Overall, the conversion of the program from a field experiment to government implementation has the potential to create significant disruptions in both the quantity and quality of the mentoring services. When comparing the mentoring practices between the experimental setting and the government scale-up, we find modest negative correlations that are not statistically distinguishable from zero. Our data does not support the notion of a significant and drastic decline in service provision at scale.

4.2 School Closures

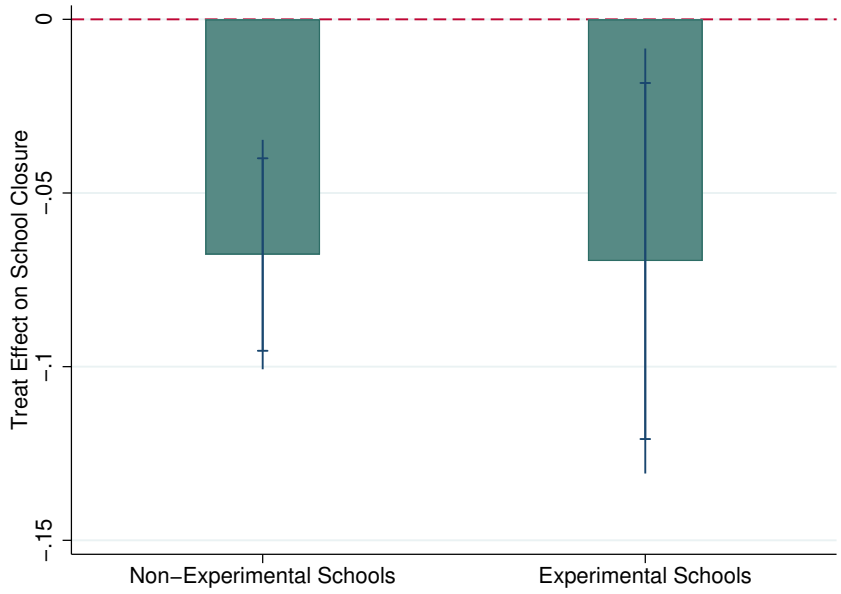
One of the key distinctions between the implementation protocol of the second experiment and the government implementation lies in the potential occurrence of school closures. This was a prevalent issue faced by the government during the large-scale implementation of the API *Original*, and there is no guarantee that the *Plus* program is immune to this threat during the government implementation. The continuity of school services is vital for maintaining the program's effectiveness at scale, as schools serve as the conduit for delivering the mentoring program. Consequently, the occurrence of school closures can significantly disrupt the program, thereby jeopardizing its potential effectiveness on a large scale. On one hand, it is conceivable that the program could fail at scale due to the high frequency of school closures in the new context. On the other hand, if the *Plus* program successfully prevents the adverse event of school closures during the government implementation, it presents us with a valuable opportunity to gain insights into the mechanisms that enhance the scalability of this program modality when compared to the previous modality.

Some institutional details make school closures more salient during the government implementation. For example, the decision of closing schools is determined by the parent association with a vote. In particular, whenever the number of students enrolled drops below six the school ceases to operate by default, unless the majority of parents oppose by vote. This procedure can determine a notable difference in situation between the field experiment and the government scale-up. 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. As a result, only two schools closed in the sample of 230 schools in the second experiment, compared with a an average 11 percent school closing rate in the rest of Chiapas for the three years before the experiment, and with a 19 percent probability of school closures for schools with few than 10 children enrolled (median school size).

Did the *Plus* program prevent the adverse effects of school closures on students? To answer this question, we adopt the same regression model (1) and the same sample of schools previously used to evaluate the program at scale (see Table 4). The outcome of interest is whether a school results permanently closed from the administrative school census during the fall of 2019. Figure 3 shows that the government implementation of the *Plus* modality induces a significant and substantial effect on school closures. Both experimental schools and non-experimental schools in Chiapas exhibit similar patterns of school closures. When focusing on the schools outside of the experimental sample in Chiapas (N=1,161), we observe a 6.8 percentage point reduction in the probability of school closures due to the program (p-value < 0.001). Schools that were previously part of the experimental sample (N=184) also experience a notable decrease in school closures during the government implementation of the API *Plus*, with an average impact of the mentoring program of -7.0 percentage points (p-value $= 0.026$).

The impact of the program at scale for the schools in the experimental sample closely aligns with the corresponding impact of the experimental API *Plus* intervention two years after the experiment's conclusion (-8.3 percentage points, see Table 7). Our previous findings on the impact of the *Plus* program on educational outcomes, combined with this additional piece of evidence, suggest that the program's underlying effectiveness endures during the government implementation. This further corroborates the mixed evidence regarding a possible reduction in the quality of the government-provided mentoring service discussed in the previous subsection. In the next section, we will study the mechanisms behind the success of this modality of the program at scale.

Figure 3: The Impact of the API Plus Program at Scale 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 (same as in Equation 1) 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. The OLS estimates, p-values, and number of observations for the two subsamples of schools are also reported in Table B-10.

5 Pathways to Scale

In this section, we investigate the possible mechanisms that promoted the scalability of the *Plus* program. We make use of an array of survey modules collected during the two field experiments. For this part of the analysis, we will thus focus on the experimental samples of schools. The sampling design is explained in Section 2.2. The schools of the second experiment are largely representative of the broader population of schools in the State of Chiapas, in term of observable characteristics (see Tables 1 and B-3) as well as in terms of program impacts at scale (see Table 4 and Figure 3).

5.1 Parental Investment and Behavior

Table 5 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 cor-

Table 5: 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 Schools	73	73	73	73
Number of 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)
Original = *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 Schools	224	224	224	224
Number of 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.

responding 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. After correcting inference for multiple hypotheses testing, we can reject the null hypothesis of equal treatment

effects on all four parental outcomes shown in Panel B of Table 5.[13]

Overall, the results show that the API intervention had differential impacts on parental investments, 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 parental engagement toward their children's education.

Home visits are a key component of the mentoring intervention under study. The goal of these visits as well as other encounters between mentors and parents in the school's premises is to increase parental awareness about their children's educational trajectories through periodic interactions. We study the role of the these interactions as a potential mechanism behind the large and positive effect of the *Plus* modality on parental outcomes documented in Table 5. Panel A in Table 6 displays the estimated differences across the two API modalities on selected survey variables when parents were asked about the frequency and content of their interactions with the mentors over a period of two months prior to the survey.[14] In spite of quite noisy estimates due to missing observations 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. The last two columns 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

[13]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 5. Table B-11 reports the 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.

[14]The number of observations varies across the columns in Panel A due to some of the 591 interviewed parents not responding to the survey questions. Missing values for each outcome are balanced with respect to the assignment of the API *Plus* (p-values = 0.746, 0.183, 0.442, 0.517, 0.539, and 0.575).

Table 6: 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)
Number of Observations	482	491	504	354	353	357

	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)
Number of 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). 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).

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 6 shows the effect of the API *Plus* on different competencies, or "parenting styles," that the mentors report to have promoted during their encounters with parents.[15] 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.

[15]Of a total of 126 schools that received mentors between the *Original* and *Plus* modalities, our survey enumerators were able to collect information for 107 schools. The set of schools considered in this section is the same as the one examined in Section 4.1 for comparing the mentor population between the field experiment and the government implementation. The attrition of survey participation is unrelated to the treatment assignment (p-value $= 0.514$). For further details on the survey of mentors, please refer to Appendix A.2.

These findings point 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 parental investment and behavior as well as in children's outcomes. Although we are unable to precisely quantify the individual impact of each training module, it is probable that these effects can be attributed to the peer-to-peer sessions facilitated by mentors. The sessions provided a platform for participants to exchange valuable information on effectively engaging parents in their children's learning. Instead, the extra week of initial training is 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):[16]

- *"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."*

We evaluate the role of other possible channels related to the mentoring service that might partially account for the effectiveness of the *Plus* program compared to the *original* modality. In particular, we focus on the additional two activities part of the mentoring program: (i) remedial education sessions with students lagging behind; and (ii), pedagogical support to the local instructors. Although the design of the second experiment does not allow us to isolate the direct effect of the remedial education sessions within each API modality, we exploit the discontinuity in the eligibility of children for the remedial sessions (see Section 2.1 for details on the eligibility). The estimates displayed in Table B-12 suggest that there is no differential effect across children's outcomes in the relative impact of the two training modalities between children who are more or less likely to be eligible for the remedial sessions

[16] We conducted a series of in-depth interviews in the spring of 2022 for a small and representative sub-sample of 16 mentors and 12 community instructors who were part of our study. Appendix A.3 reports more details about these interviews. Tables B-14 and B-15 show that the characteristics of these survey respondents are broadly comparable to those of the mentors and the local instructors in our main sample.

(see also Figure B-3).[17]

We next consider the role of the pedagogical practices of the community instructors. Because mentors provide help in improving their teaching habits, we test the hypothesis of whether this factor may partly explain the differential effect of the *Plus* modality on children's outcomes. Table B-13 reports estimates of the effect of the API *Original* and API *Plus* using data at the school-level 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).[18] The results show erratic patterns of positive and negative signs with no statistically significant effects of either API modality.

In summary, cross-modality differences in the effectiveness of the remedial education sessions or in the pedagogical support for instructors are unlikely to explain the success of the *Plus* program. Both our quantitative and qualitative evidence establish the key role of a more active parental involvement, which was likely triggered by enhanced parent-mentor interactions. These interactions are likely influenced by the additional component of peer-to-peer sessions in the training sessions of the *Plus* modality, during which mentors share their experiences regarding the home visits and their interactions with families.

5.2 Evidence on Parents as Means of Scalability

As discussed in Section 2.1, the functioning of these community-based schools is heavily reliant on the active involvement of parents through the local parental association. In particular, the association rules over the decision of whether or not to close the school, a situation that is automatically considered when the number of students enrolled in the school drops below six. Because school closures can undermine the success of the API *Plus* at scale (see Section 4.2), this effectively implies that parents can play a crucial role in the scalability of the mentoring program.

We evaluate this hypothesis by examining variations both across and within two experiments. Specifically, we investigate whether the contrasting responses in parental investment and engagement at the local school across the two mentoring interventions (refer to Table 5) are

[17]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. Because the diagnostic score is not perfectly correlated with the test score outcomes, the threshold rule provides us with variation to rule out the mediation role of remedial sessions on the treatment effect determinants.

[18]The sample average number of instructors per school is 1.2 in the school year prior to the start of the second experiment.

reflected in differential rates of school closures between the two program modalities. The first two columns of Table 7 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 scale-up 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 finding supports the notion that situations characterized by a lack of parental engagement—as indicated by our previous results—are not conducive to promoting community-based educational programs, with school closure rates that resemble the ones of schools and communities with no mentors, as well as the high rates of school closures during the scale-up of the *Original* program.

The second column of Table 7 shows that the *Plus* modality, which substantially boosts parental engagement (see Table 5), has a significant impact on school closures. Schools are 8.3 percentage points less likely to close two years after the *Plus* modality was adopted by the government. An effect that 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 3.[19]

The IV estimates shown in the third column of Table 7 go a step further and quantify the extent to which parental engagement affects the probability of school closures. Because of the contextual information on the role of the parent 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 null impacts of the *Original* program across different experiments on both parental investments and school closures are consistent with this exclusion restriction. We find that an increase of 0.1 of a standard deviation in the overall parental engagement index is causally associated with a reduction of 2.2 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 local instructors,

[19]The probability of subsequently receiving a mentor during the government implementation was found to be unrelated to the randomized mentor assignment during the second experiment. 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 scale-up 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 7: School Closures and Parental Engagement

	Outcome: School Closures		
	First Experiment	Second Experiment	Second Experiment, IV
API Original	0.063	-0.031	-0.031
	[0.225]	[0.396]	[0.410]
API Plus		-0.083	
		[0.030]	
Overall Parental Engagement			-0.217
			[0.021]
Observations	73	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 5. *p*-values reported in brackets refer to the robust asymptotic inference.

engaged parents may have more at stake in keeping the schools open as they invest more 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."*

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.*"

Previous literature has highlighted the role of parental investments and parent-mentor/home visitor interactions in boosting treatment effects of home visiting programs (Heckman and

Zhou, 2021), and that parental choices 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 at scale of educational programs depends upon the local engagement of parents in the schooling activities.[20]

6 Discussion and Conclusion

We study a school mentoring program in the state of Chiapas, Mexico. By exploiting two independently run field experiments, as well as the government implementation at scale of the program, we show that variations in the training content of mentors can lead to significant variations in the final outcomes. The government's original implementation of the program proves to be largely ineffective. However, an alternative approach that prioritizes mentors' training in enhancing their ability to effectively interact with and engage parents has proven successful in enhancing test scores and improving educational attainment for the students in our sample. Within this new program modality, 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. Parental responses are shown to prevent schools to close, an otherwise threat for the scalability of the program, 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*).

This paper seizes a unique opportunity to investigate the challenges and determinants of scaling when transitioning an educational intervention from a field experiment to government implementation. The case study prominently highlights the aspects of our experimental design that contribute to an informative evaluation of the impacts at scale. However, we

[20]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?*"

acknowledge the limitations of our study in addressing the process of "vertical" scaling, which involves constructing the infrastructure for program implementation at scale from scratch following the evaluation in the field experiment. While this limitation restricts the direct implications of our study for the supply-side considerations of scaling, it provides us with valuable insights into the key challenges that arise from the changes in situation. Another limitation of our study is that it relies on university graduates as mentors, which may hinder the program's scalability in contexts where such resources are scarce.

Beyond the specific context of our analysis, we believe that our case study offers valuable insights for scholars interested in designing and evaluating scalable interventions. We highlight that scalability is an outcome that is influenced by social factors, and we underscore the pivotal role that local communities and individuals play in promoting the success of community-based interventions. We recognize that each parent is inherently unique in their approach to parenting, making it challenging to replicate individual strategies on a large scale. However, we highlight the significance of engaged communities of parents as a non-scarce asset that can be harnessed to promote and sustain positive outcomes in educational initiatives. By leveraging the power of communities with engaged parents, we can pave the way for scalable interventions that address the educational needs of children in disadvantaged circumstances.

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.

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.

Allcott, Hunt, "Site Selection Bias in Program Evaluation," *The Quarterly Journal of Economics*, 2015, *130* (3), 1117–1165.

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).

Banerjee, Abhijit, Rema Hanna, Benjamin A. Olken, Elan Satriawan, and Sudarno Sumarto, "Electronic Food Vouchers: Evidence from an At-Scale Experiment in Indonesia," *American Economic Review*, February 2023, *113* (2), 514–547.

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.

_ , Veronica Frisancho, and Marco Pariguana, "Perceived Ability and School Choices: Experimental Evidence and Scale-up Effects," IZA Discussion Papers 16168, Institute of Labor Economics (IZA) May 2023.

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.

Caron, EB, Kristin Bernard, and Allison Metz, "Fidelity and Properties of the Situation, Challenges and Recommendations," in "The Scale-up Effect in Early Childhood and Public Policy. Edited by John A. List, Dana Suskind, and Lauren H. Supplee," Routledge, 2021.

Coleman, James S., "Social Capital in the Creation of Human Capital," *American Journal of Sociology*, 1988, *94*, S95–S120.

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, Jonathan Guryan, Kelly Hallberg, and Jens Ludwig, "Studying Properties of the Population: Designing Studies that Mirror Real World Scenarios," in "The Scale-up Effect in Early Childhood and Public Policy. Edited by John A. List, Dana Suskind, and Lauren H. Supplee," Routledge, 2021.

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.

Egger, Dennis, Johannes Haushofer, Edward Miguel, Paul Niehaus, and Michael W Walker, "General Equilibrium Effects of Cash Transfers: Experimental Evidence from Kenya," *Econometrica*, November 2022, *90* (6), 2603–2643.

Engberg, John, Brian Gill, Gema Zamarro, and Ron Zimmer, "Closing schools in a shrinking district: Do student outcomes depend on which schools are closed?," *Journal of Urban Economics*, 2012, *71* (2), 189–203.

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.

García, Jorge Luis and James J. Heckman, "Parenting Promotes Social Mobility Within and Across Generations," *Annual Review of Economics*, 2023, *15* (1), null.

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.

Haan, Monique De, Edwin Leuven, and Hessel Oosterbeek, "School Consolidation and Student Achievement," *The Journal of Law, Economics, and Organization*, 2016, *32* (4), 816–839.

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.

List, John, Fatemeh Momeni, and Michael Vlassopoulos, "Neighborhood Spillover Effects of Early Childhood Interventions," CEPR Discussion Papers 18134, C.E.P.R. Discussion Papers May 2023.

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.

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.

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.

_ , _ , **and Sandip Sukhtankar**, "General equilibrium effects of (improving) public employment programs," *Econometrica*, 2023, (Forthcoming).

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.

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 7 and Figure 3.

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 Tables 1, B-1 and B-3. We leverage information on schooling outcomes in the 2020 population census for all the localities in the state of Chiapas (including those that were part of the experimental sample), which is shown in Table 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 in each pedagogical area in 2013 as our main

measures of academic achievement. The *Overall Score* displayed in Table 2 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 Table B-4.

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 3, 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 B-2.

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-3 and Table B-12, which proxy for the individual eligibility for the one-on-one remedial education sessions.

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-7: 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

Measures of Children's Achievement. 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.[21] The reading scores reported in Tables 3 and B-12 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 3 and B-12 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-5.

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 deviation observed in the control group. The socio-emotional scores reported in Tables 3 and B-12 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)

[21] 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).

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 3 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 5 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-11.

Mentor Characteristics. As part of the data collection activities, we have collected basic socio-demographic information on the mentors who served in the schools of the second experiment. Those are reported in Panel C of Table B-2 and in the second column Table B-8. For the other schools in Chiapas that were not part of the experimental sample, we rely on administrative rosters about mentors' characteristics from the program. Those are reported in the first column of Table B-8.

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 6. Basic information on both the household module respondent and household characteristics is reported in Panel B 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 6.

Teaching Practices. Local instructors were asked standard questions on their socio-demographic characteristics, education and experience. Those are reported in Panel A of Table B-2. 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-13. *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.

Quantity and Quality of Mentoring Services. Local instructors were asked about mentors' practices and activities within the local communities at two specific points in time: during the end-line survey of the experiment (Spring 2016) and in an additional follow-up survey module conducted in the fall of 2018 among the schools that were previously involved in the second experiment. The end-line survey was conducted in 57 out of a total of 58 schools that received the API *Plus* during the experiment. The follow-up survey was conducted in 93 out of a total of 103 schools that implemented the API *Plus* program at scale. We obtained information about mentors from the responses collected by local instructors for 56 schools in the end-line survey and 58 schools in the follow-up survey. The corresponding measures are presented in Figure 2 and Table B-9.

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-14 and B-15.

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 5.1:

> "*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.

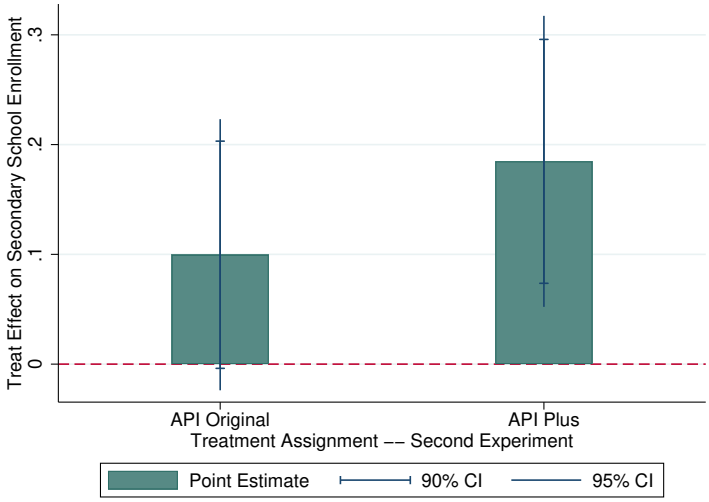
> "*La gestión dela escuela y se le hicieron mejoras de cercado, pintaron la escuela arreglaron los baños y se compraron materiales.*"

> "*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.*"

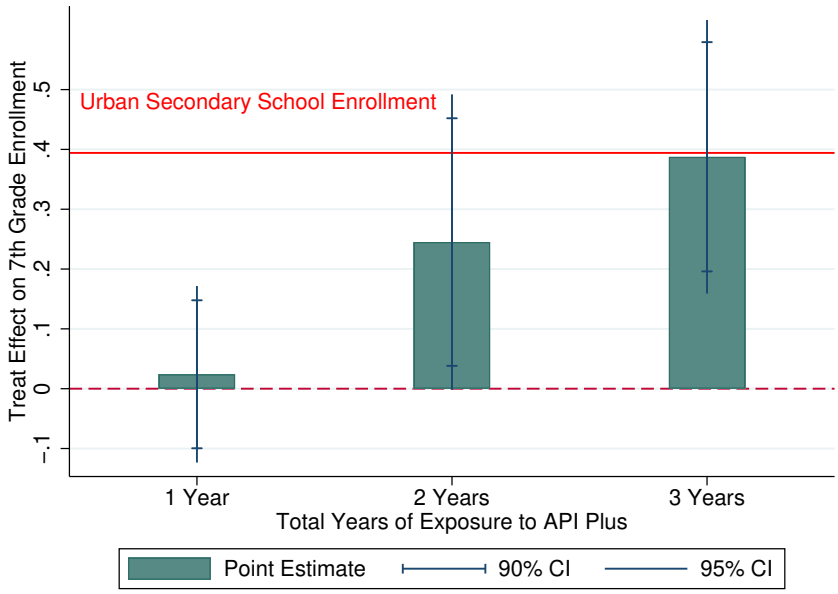
B Additional Figures and Tables

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



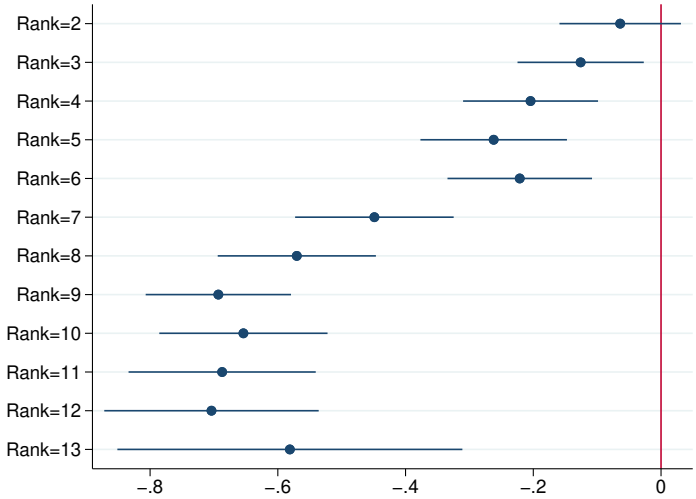
Notes: The bars depicted in this figure show the OLS estimates of the original treatment assignments in our experiment on the probability of enrolling in seventh grade in the year after the end of the second experiment (2017). The vertical lines overlaid on the bars represent asymptotic confidence intervals 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. Attrition is balanced among schools that were part of the two treatment arms (p-values = 0.914, and 0.768).

Figure B-2: The Cumulative Effect of API *Plus* 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 API *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. The sample includes 207 schools of the 224 that were part of the second experiment. Beyond a school that permanently closed, the sample attrition is caused by schools not having sixth graders during that school year. Attrition is balanced with respect to the indicator variables for the years of exposure to API *Plus* (p-value[1 year]=0.467, p-value[2 years]=0.812, and p-value[3 years]=0.568, the reference category is zero years of exposure).

Figure B-3: 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

	API Original		Control		Diff
	Mean	Std. Dev.	Mean	Std. Dev.	p-value
	Panel A: Schools in Mid-Line 2012 Survey of Parents				
Average Test Score (Spanish)	401.971	38.973	399.036	28.974	0.703
Average Test Score (Math)	377.916	43.159	388.422	51.038	0.351
Number of Students	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
Total Population	217.054	597.061	234.778	506.694	0.888
Labor Force Participation	0.286	0.064	0.276	0.069	0.553
Water Network (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
Rate of Illiteracy	0.321	0.170	0.333	0.173	0.745
Garbage Collection (Y/N)	0.027	0.164	0.056	0.232	0.551
Number of Schools/Localities	37		36		
	Panel B: Schools with Individual Test Score 2013 Data				
Average Test Score (Spanish)	401.869	40.034	399.206	29.378	0.748
Average Test Score (Math)	377.168	44.284	390.561	50.120	0.242
Number of Students	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
Total Population	225.857	612.996	227.543	512.201	0.990
Labor Force Participation	0.287	0.065	0.278	0.069	0.579
Water Network (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
Rate of Illiteracy	0.327	0.165	0.335	0.175	0.823
Garbage Collection (Y/N)	0.029	0.169	0.057	0.236	0.566
Number of Schools/Localities	35		35		

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 Statistic	Control Mean (SD)	API Original Mean (SD)	API Plus Mean (SD)	Original-Control Mean Difference (p-value)	Plus-Control Mean Difference (p-value)
Panel A: School and Teacher Characteristics					
Average Test Score (Spanish)	429.389	432.326	430.573	0.846	0.743
	(60.477)	(67.579)	(67.463)	(0.738)	(0.792)
Average Test Score (Math)	453.090	455.820	451.627	0.156	-2.057
	(78.436)	(84.546)	(82.461)	(0.978)	(0.778)
Average Test Score (Science)	438.349	441.259	442.856	1.435	3.866
	(50.264)	(49.323)	(50.492)	(0.735)	(0.390)
Number of Teachers	1.224	1.309	1.207	0.086	-0.016
	(0.419)	(0.465)	(0.409)	(0.213)	(0.820)
Number of Students	15.296	15.441	14.379	0.161	-0.953
	(5.819)	(5.655)	(5.824)	(0.857)	(0.320)
Teacher with Secundary Education	0.763	0.794	0.833	0.031	0.072
	(0.389)	(0.398)	(0.358)	(0.628)	(0.241)
Years of Experience as Teacher	0.737	0.706	0.693	-0.034	-0.042
	(0.872)	(0.802)	(1.085)	(0.802)	(0.797)
Months of Teacher Working in the School	9.531	9.309	9.281	-0.229	-0.249
	(3.947)	(4.925)	(3.266)	(0.751)	(0.676)
Observations	98	68	58	166	156
Panel B: Child and Household Characteristics					
Age in Months at Baseline (September 2014)	104.993	104.289	105.539	-0.818	0.647
	(16.384)	(17.532)	(14.924)	(0.485)	(0.605)
Male (Y/N)	0.532	0.519	0.543	-0.011	0.013
	(0.500)	(0.500)	(0.499)	(0.734)	(0.772)
Indigenous Language (Y/N)	0.302	0.307	0.461	0.012	0.155
	(0.460)	(0.462)	(0.499)	(0.855)	(0.032)
Scholarship (Y/N)	0.746	0.733	0.747	-0.013	0.005
	(0.436)	(0.443)	(0.435)	(0.763)	(0.903)
Parent Can Read	0.715	0.686	0.734	-0.030	0.023
	(0.452)	(0.465)	(0.443)	(0.465)	(0.590)
Parent with Less than Primary	0.614	0.587	0.584	-0.027	-0.029
	(0.487)	(0.493)	(0.494)	(0.526)	(0.483)
Household Receives Oportunidades CCT	0.812	0.807	0.829	-0.003	0.016
	(0.391)	(0.395)	(0.377)	(0.929)	(0.614)
Observations	453	322	269	775	722
Panel C: Mentor Characteristics					
Age in Years		28.386	28.400		0.242
		(3.678)	(3.057)		(0.705)
Male		0.579	0.620		0.051
		(0.498)	(0.490)		(0.597)
High Edu Complete		0.877	0.880		0.006
		(0.331)	(0.328)		(0.926)
Months of Experience as Mentor		22.298	20.040		-2.218
		(10.997)	(8.755)		(0.260)
Observations		57	50		107

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. p-values for the null hypothesis of equal mean differences are reported in parentheses in the last two columns. See Appendix A for more details on the data sources.

Table B-3: Summary Statistics across Samples

	All Chiapas			Second Experiment		
	All Sample Mean (SD)	Census Sample Mean (SD)	Mean Difference (p-value)	All Sample Mean (SD)	Census Sample Mean (SD)	Mean Difference (p-value)
			Panel A: School Characteristics			
Average test score (Spanish)	424.503	422.903	1.600	431.340	433.855	-2.515
	(56.466)	(54.786)	(0.522)	(60.810)	(63.370)	(0.705)
Average test score (Math)	414.921	413.736	1.184	421.333	424.043	-2.710
	(75.300)	(74.699)	(0.725)	(80.895)	(84.848)	(0.760)
Number of students	14.049	13.974	0.075	15.009	15.158	-0.149
	(8.468)	(8.865)	(0.834)	(6.053)	(5.794)	(0.799)
Number of Teachers	1.231	1.240	-0.008	1.217	1.217	-0.000
	(0.467)	(0.480)	(0.671)	(0.413)	(0.414)	(1.000)
Share Over-aged Students	0.349	0.348	0.001	0.324	0.290	0.034
	(0.797)	(0.818)	(0.971)	(0.659)	(0.615)	(0.589)
			Panel B: Locality Characteristics			
Total Population	118.758	121.170	-2.412	121.389	158.276	-36.887
	(221.648)	(208.666)	(0.775)	(240.562)	(337.620)	(0.219)
Share of High-Poverty Villages	0.490	0.489	0.001	0.473	0.453	0.020
	(0.500)	(0.500)	(0.945)	(0.500)	(0.499)	(0.702)
Incidence of Social Conflict (Y/N)	0.190	0.204	-0.014	0.187	0.201	-0.014
	(0.392)	(0.403)	(0.355)	(0.391)	(0.402)	(0.719)
Share of Illiterate Adults	0.313	0.315	-0.002	0.295	0.292	0.003
	(0.160)	(0.159)	(0.703)	(0.153)	(0.150)	(0.860)
Share of Adults in the Labor Force	0.297	0.296	0.002	0.303	0.301	0.002
	(0.076)	(0.077)	(0.575)	(0.070)	(0.067)	(0.765)
Locality Access without Road	0.216	0.224	-0.008	0.179	0.149	0.029
	(0.411)	(0.417)	(0.609)	(0.384)	(0.357)	(0.426)
Water Network (Y/N)	0.028	0.028	0.000	0.022	0.038	-0.016
	(0.164)	(0.164)	(0.998)	(0.146)	(0.192)	(0.341)
Sewage System (Y/N)	0.011	0.012	-0.001	0.009	0.016	-0.008
	(0.105)	(0.109)	(0.830)	(0.093)	(0.127)	(0.497)
Garbage Collection (Y/N)	0.022	0.023	-0.002	0.022	0.027	-0.005
	(0.146)	(0.151)	(0.784)	(0.146)	(0.163)	(0.724)
Observations	1,523	1,161	3,046	230	184	414

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.

Table B-4: Placebo Test for API Plus Assignment During Program Scale-up

	Spanish		Math		Science	
API Plus	-0.104	0.003	-0.093	-0.001	-0.062	0.027
	[0.062]	[0.954]	[0.099]	[0.989]	[0.268]	[0.621]
	{0.062}	{0.949}	{0.112}	{0.993}	{0.277}	{0.643}
	(0.107)	(0.996)	(0.146)	(0.996)	(0.260)	(0.866)
Controls for Criteria	No	Yes	No	Yes	No	Yes
Observations	1183	1183	1183	1183	1183	1183

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. Control variables include indicator functions for the four criteria used to determine the differential priority across eligible schools to receive the mentors (see Section 2.1) as well as an indicator function for prior exposure to the mentoring program and the number of hostile event related to land property, religion, elections, crime, or drug addiction as reported at the locality level in the population census (2010). p-values reported in brackets refer to the conventional asymptotic standard errors. 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 Plus across the two specifications considered (without and with controls) through the step-wise procedure described in Romano and Wolf (2005a,b, 2016).

Table B-5: 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-6: 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-7: 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: *Original* = *Plus*	[0.834]	[0.491]	[0.911]	[0.620]
Number of Schools	224	224	224	224
Number of Observations	1019	1019	1019	1019

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. Asymptotic p-values reported in brackets are clustered at school level. For a detailed descriptions of the schooling records used in this table, see Appendix A.1.

Table B-8: Comparison of Mentors' Characteristics Across Situations

Variable	(1) Chiapas sample	(2) Experiment 2	(3) Chiapas vs Experiment 2
Male	0.571	0.604	0.033
	(0.495)	(0.491)	(0.492)
Age	28.460	28.266	-0.194
	(3.780)	(3.287)	(0.558)
Speaks Indigenous Language	0.295	0.374	0.079
	(0.457)	(0.486)	(0.101)
Observations	441	139	580

Notes: This table shows the comparison of mentors' characteristics between the second experiment and the scale-up of the *Plus* program. The first two columns show mean and standard deviations for both samples. The third column shows the difference and the associated p-values of the null hypothesis of no difference across samples.

Table B-9: Change in Situation and Impacts on Quality and Quantity of Mentoring Program

	Quantity			Quality		
	Days in Community	Number Activities with Instructor	Time Spent with Instructor	Meetings with Parents of Students at Risk	Students with API Support	Time spent with Students
Change in Situation	-1.585	-1.093	-0.954	-0.596	-0.546	-0.025
	[0.330]	[0.037]	[0.189]	[0.407]	[0.483]	[0.908]
Observations	114	113	114	109	96	110
Observations Survey 2016	56	55	56	51	54	52
Observations Survey 2018	58	58	58	58	42	58

Notes: This table shows the comparison in the quantity and quality of API *Plus* program between the second experiment and the government implementation. This information is collected during the surveys of the local instructors, in the school years 2015-2016 and 2018-2019. Each estimate in each column represents an OLS estimate for the difference in the mentoring services across the two situations. The asymptotic p-values are reported in square brackets. All the regressions include the same set of controls as in Table 4.

Table B-10: The Impact of the API Plus Program on School Closures

	Non-Experimental Schools	Experimental Schools
API Plus	-0.068	-0.070
	[0.000]	[0.026]
Observations	1161	184

Notes: This table shows 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. p-values reported in brackets are based on asymptotic inference. All the regressions include the same set of controls as in Table 4.

Table B-11: 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-12: Remedial Education Sessions

	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: *Original* = *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]
Number of Schools	224	224	224	224
Number of Observations	1044	1044	1045	1045

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 Figure B-3). 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-13: Teacher Pedagogical Practices

	Learning Activities	Engage With Students	Manage Time	Use of Material	Overall Index
API *Original*	0.048	-0.066	0.093	-0.125	-0.037
	[0.718]	[0.678]	[0.645]	[0.473]	[0.795]
	{0.711}	{0.692}	{0.658}	{0.458}	{0.797}
	(0.973)	(0.973)	(0.973)	(0.926)	(0.973)
API *Plus*	-0.072	0.050	-0.086	0.025	-0.203
	[0.619]	[0.738]	[0.620]	[0.871]	[0.150]
	{0.620}	{0.752}	{0.598}	{0.884}	{0.141}
	(0.956)	(0.956)	(0.956)	(0.956)	(0.322)
Original = *Plus*	[0.462]	[0.488]	[0.358]	[0.428]	[0.274]
	{0.432}	{0.497}	{0.385}	{0.448}	{0.281}
	(0.709)	(0.709)	(0.699)	(0.709)	(0.580)
Number of Observations	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 the block randomization design at the strata level. p-values reported in brackets refer to the conventional (robust) 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) on multiple outcomes through the stepwise procedure described in Romano and Wolf (2005a,b, 2016).

Table B-14: 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-15: 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.