Enhancing Human Capital in Children:
A Case Study on Scaling[*]

Francesco Agostinelli

University of Pennsylvania

Ciro Avitabile

The World Bank

Matteo Bobba

Toulouse School of Economics

June 21, 2024

Abstract

This paper provides novel insights into the science of scaling by examining an educational mentoring program in Mexico. The empirical analysis encompasses two independent field experiments, and seizes a unique opportunity to learn from the government's implementation of the same intervention. While the program originally implemented at scale demonstrates limited effectiveness, the introduction of a new modality with enhanced mentor training significantly improves children's outcomes. Mentor-parent interactions are found to stimulate parental engagement at the community-school level. Our findings support the hypothesis that parents can play an important role in facilitating the scalability of educational programs.

1 Introduction

A key challenge in using scientific insights to inform policy decisions arises during the implementation process, where small variations in the protocol of the intervention often 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 a recent debate about the challenges to scale-up education interventions. In particular, we provide an empirical case study involving a mentoring program that deploys recent university graduates to remote and disadvantaged communities in Mexico. Mentors are engaged to a school community for a period of two years. Among other tasks, they help the local instructors, and encourage parental involvement in children's education. The mentoring intervention was initially launched at the national 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 a randomized trial revealed null results of this program modality. The lack of effectiveness of the program served as a catalyst, prompting the need to improve the delivery of mentoring services in the targeted communities.

Our research team collaborated with the government—including accessing the existing infrastructure of the ongoing intervention—to design and implement an experimental evaluation of a new program modality that incorporated an enhanced training protocol for mentors. Among other changes in the training module, mentors attend training workshops and peer-to-peer meetings throughout the school year, during which they share effective practices on how to improve parenting skills and better interact with parents. The program innovations were largely influenced by the 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; Heckman and Mosso, 2014; Attanasio et al., 2022).

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 investment, the government made the decision to adopt the program with the most effective approach. Our comprehensive analysis of the program's implemen-

tation in Chiapas, the poorest state in Mexico, provides strong support for the notion that the effectiveness of the new modality of the mentoring intervention was preserved at scale. Importantly, we find that parental engagement and attitudes toward schooling activities emerge as critical factors for the program's scalability.

The empirical evidence draws from two independent field experiments and the subsequent government scale up of the effective program modality. The first experiment is directly carried out by the government immediately following the national implementation of the original mentoring program. Assignment to the program is randomized across 80 program-eligible primary schools. 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 the mentoring program, the original modality displays relatively small and noisy effects on children's achievement outcomes when compared to the control group with no mentors. The new modality with enhanced training for the mentors delivers sizable and significant gains in children's reading scores (+0.32 standard deviations, p-value = 0.001), math scores (+0.24 standard deviations, p-value = 0.005), and socio-emotional scores (+0.20 standard deviations, p-value = 0.011) as well as a large effect on the probability of enrolling in seventh grade (+12.4 percentage points, p-value = 0.030), out of a basis of 62 percent enrollment in the control group. The effect of the mentoring intervention on educational outcomes is statistically different across program modalities.[1]

The government's decision to transition the program to a more effective modality provides us with a unique opportunity to investigate the factors and mechanisms influencing scaling. We integrate data from several administrative sources and leverage the early assignment of the program at scale across communities in Chiapas as a source of variation in exposure to mentoring activities. For our sample of 1,345 eligible schools, we demonstrate that, after considering the government's official criteria for program assignment, the 356 schools that received the 2-year mentoring program in our sample period exhibit no significant differences in observable characteristics and predetermined educational outcomes when compared to the remaining eligible schools. Our results show that the new variant of the mentoring intervention remained successful when implemented by the government. For the sub-sample of the 1,161 localities outside of the experimental sample, which had never received the *Plus* modality before and encompassed approximately 16,000 children enrolled in eligible

---

[1]All p-values reported in the text are adjusted for multiple hypotheses testing through the step-wise procedure described in Romano and Wolf (2005a,b, 2016). Alternative inference procedures, which are discussed in more detail in Section 3, deliver results that are broadly consistent with those reported here.

schools, the results show a positive effect on secondary school enrollment for schools that received the program, with an average program impact of 5.6 percentage points (p-value = 0.013). We further document positive and significant effects of the program on child literacy, which imply a reduction of illiteracy rates by approximately 20 percent with respect to the sample mean for the overall sample of schools. Regarding the remaining eligible localities that were previously included in the experimental evaluation, we observe comparable results. For instance, the average impact of receiving the new program modality on the fraction of children who enroll in lower-secondary education is +9.1 percentage points (p-value = 0.035).

The effectiveness of the new program when implemented by the government was not guaranteed *a priori*, despite the sizable and precise effect sizes observed in the field experiment. The existing literature underscores the significance of several "non-negotiable" aspects of program designs in the context of such changes in the situation (List, 2022). Neglecting to consider these critical elements during the widespread implementation of the intervention has the potential to not only reduce but also completely eliminate the program effects documented under experimental conditions (Al-Ubaydli et al., 2020; Caron et al., 2021; Al-Ubaydli et al., 2021). While we do observe minor changes in both the quantity and quality of mentors' activities during the scale-up phase, these differences are generally small in magnitude and lack statistical precision. We interpret these results as suggestive evidence that the mentoring program did not experience a substantial disruption during the transition between the two phases.

We argue that a potential source of "voltage drop" of the program results from the experimental design's trade-off between real-world applicability and 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 field experiment from the research team had minimized the extent of this event in the period of the evaluation. The school closure rate is notably high, averaging 11 percent before the experimental evaluation. In contrast, only two schools closed during the randomized trial out of 230 schools. To the extent that the continuity of the schooling services is critical for ensuring the program's effectiveness, this particular difference in the implementation protocol poses a challenge for the ability of the field experiment to inform us about the scalability of the mentoring program.

We zoom into the relationship between exposure to the mentors and school closures in order to study the sources of scalability of the program. We first show evidence that the new program modality, unlike its predecessor, drastically reduces the occurrence of school clo-

sures. 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. 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 (p-value=0.002). 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 findings on school closures and on parental responses provide suggestive evidence that parents can play an important role in the scalability of the program. Under the assumption that the treatments effects of the new modality of the mentoring program operate only through parental engagement, IV estimates document that an increase of 0.1 standard deviation in parental engagement reduces school closures by 2.2 percentage points (p-value = 0.021). The original modality, instead, displays null impacts on school closures in both experiments. This finding underscores the challenge faced by community-based educational programs in situations where parental engagement is lacking, making their success more difficult to achieve.

The implications of our results are pertinent to policy discussions on the design of educational 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. For instance, August et al. (2006) find that as the situation shifted from their original field study to a larger scale implementation, there was a notable decline in family involvement in a program aimed at preventing conduct problems. This reduction in participation on a larger scale could significantly impact the effectiveness of broader implementations. The precise elements of the intervention (such as the training module in our study) and the consequent efficacy of mentor-parent interactions are key factors influencing how parents respond, thereby influencing the potential scalability of educational interventions.

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

recent and growing evidence on 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 field experiment has been designed and implemented during the roll-out of the original program, in close collaboration with the government agency responsible for the subsequent scale-up of the new modality. This collaborative approach guarantees the harmonization of the research design with the practical considerations and implementation realities of the policy under study (Banerjee et al., 2017; Muralidharan and Niehaus, 2017; Duflo et al., 2024; List, 2024).

We analyze two independent field experiments on different and representative samples as well as a larger, non-experimental sample of schools. Drawing joint inferences from these samples harnesses the statistical power of our findings (Maniadis et al., 2014; Allcott, 2015; Al-Ubaydli et al., 2020). Furthermore, the randomization was implemented at a relatively large unit level, encompassing schools and their surrounding communities. This design feature 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 build upon the theory of human capital investments (Becker, 1962) by highlighting that scaling educational interventions is inherently a socially-determined outcome. 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; Zhou et al., 2021; García and Heckman, 2023), and that parental choices are responsive to the environments that families face (Doepke and Zilibotti, 2017; Agostinelli, 2018; Agostinelli et al., 2020). The evidence presented here sheds light on how the scalability of educational programs critically depends upon the local engagement of parents in the schooling activities.

While this unique case study provides us with an opportunity to examine the challenges and determinants of scaling in the context of a 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 a large-scale program without an existing government infrastructure. For instance, in our setting this would entail developing the organizational capital to recruit a significant number of new mentors and personnel responsible for program operations. Our findings are silent on these implementation aspects.

2   Context and Data

In this section, we discuss some relevant features of the mentoring program under study. We leverage two independent randomized experiments in conjunction with the conversion of the mentoring program into the new modality, serving as a compelling case study to uncover novel insights on the science of scaling. We draw on a rich combination of administrative and survey data sources, along with qualitative interviews with instructors and mentors (see Web Appendix A for more details).

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 multi-grade classroom with on average 15 students. For brevity, throughout the text we will refer to 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. 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 contributing to the school's financial means. 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 basic education. 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.

Mentors are assigned to work within a specific school community for the entire duration of the program, spanning two consecutive school years. In the event of a mentor's early departure from the community, the government endeavors to identify a replacement to ensure the uninterrupted continuation of program activities. The supply of available mentors only allows for accommodating a subset of the CONAFE schools in the country. Consequently, the allocation of the mentoring program across schools follows a priority-based mechanism based on four criteria: (i) at least 30 percent of the students are classified as "insufficient" in the national standardized achievement test; (ii) there are high, or very high, levels of poverty in the respective localities, as measured by a composite index of deprivation (CONEVAL, 2018); (iii) the municipality where the school communities are located is the recipient of a national anti-poverty program (the National Crusade Against Hunger); and (iv) the school has not received a mentor in previous academic cycles. A given school can receive the mentoring program over multiple (and even consecutive) two-year spells.

Mentors conduct periodic home visits as well as meetings on the school premises to update parents about their children's progress in school and encourage their active involvement in school activities. Each mentor is responsible for organizing individual remedial education sessions at school, which are held primarily 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 possibly improve 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.[2] Finally, mentors hold periodic meetings with their own supervisors throughout the school year (every two months in two-day sessions), which are focused on enhancing mentors' pedagogical practices with the students. Henceforth, we will refer to this format of the mentoring program as the API *Original*.

Starting from 2016, the government adopted a new modality of the program, which we will refer to as the API *Plus* modality. The API *Plus* encompasses all the features of the

---

[2]Only the remedial tutoring activities are targeted toward the six weakest students in the class. The other tasks of the mentors (home visits and teachers' support) are directed toward all children in the school.

Figure 1: The Mentoring Program in Chiapas and the Different Study Samples
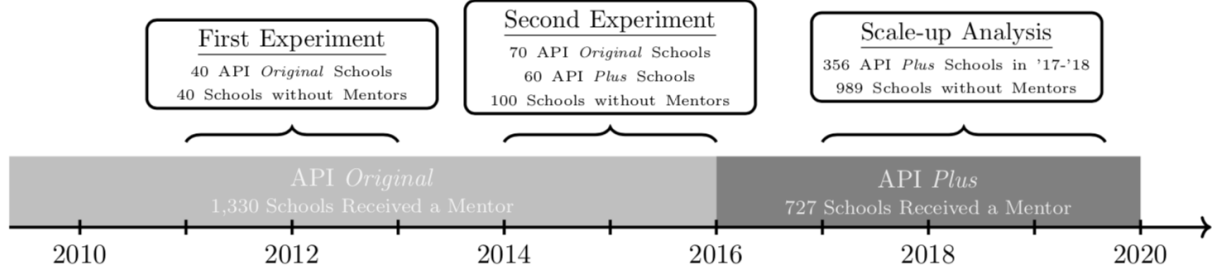
API *Original* but with substantial changes in the training module. First and foremost, it extends the initial training period from one week to two. During this additional week, the focus shifts to practical, hands-on strategies designed to enhance students' reading and math skills. Secondly, the bimonthly meetings have undergone significant modifications. In contrast to the *Original* modality, each *Plus* modality meeting includes a training workshop to improve parenting skills (communication, learning activities, and managing transitions). Additionally, there is an extra day dedicated to peer-to-peer sessions. These sessions are intended to enable mentors to collaboratively devise strategies for more effective interaction and engagement with parents.[3] 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., 2022).

Figure 1 illustrates the timeline of the mentoring program in the State of Chiapas. Between 2009 and 2016, the API *Original* served over 1,300 schools and approximately 18,000 students. Following the transition to the API *Plus* between 2016 and 2020, around 700 schools and 10,000 students received mentoring support. Due to the program's two-year rotation cycle, different schools received the program over time, guided by the timeline and assignment criteria. CONAFE has the capacity to deploy approximately 360 mentors for each program's cycle in Chiapas, and this number has been reasonably stable since 2009. This reflects both the financial constraints of the program and a supply-side constraint related to the availability of highly-educated mentors in this context. For this reason, not all primary schools receive a mentor within a cycle. However, by the spring of 2016, almost the totality

---

[3]The decision to innovate the program's modality along these lines was inspired by extensive economic literature, which suggests that successful mentoring programs in similarly disadvantaged contexts share a design feature of fostering parental engagement (Heckman and Mosso, 2014; Attanasio et al., 2022; García and Heckman, 2023). Moreover, this decision was influenced by the feedback received from the mentors during the initial implementation of the *Original* modality. They reported that the interactions with parents were the most challenging aspect of their tasks in the local communities.

of schools in Chiapas received a mentor due to the reassignment of mentors across schools. When in the fall of 2016 the government converted the program in its *Plus* modality, the mentor assignments across schools started anew. This explains why the cumulative number of schools that received the API *Original* modality at least once during the 2009-2016 (seven school years) is approximately double the cumulative number of schools that received the API *Plus* modality at least once during the 2016-2020 (four school years).[4]

2.2 Experimental Design and Scale-up Analysis

Two consecutive and independent randomized evaluations took place following the nation-wide implementation of the *Original* mentoring program. A first experiment was directly carried out by the government. We designed and implemented a second experiment in close collaboration with the government, leveraging the existing program's infrastructure. Figure 1 visually depicts the timelines for each pilot in comparison to the government's implementation at scale of the programs. After learning about the results of the second experiment (see Section 3), the government decided to discontinue the API *Original* program in the summer of 2016. All its primary schools, including those that were part of the evaluation samples of the two randomized trials, were deemed eligible to receive the API *Plus* program modality from the school year 2016-2017.

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.

We use administrative data on student-level test scores and a household survey that was collected by the government. Student outcomes are measured two school-years after the

---

[4]Out of the 1,523 schools in the program-eligible communities of Chiapas, 1,330 received mentoring at least once between 2009 and 2016. The final deployment of API-*Original* mentors took place in the 2015-2016 school year, coinciding with the second year of the program assignment for the second experiment (see Figure 1). At that point, the vast majority of the schools that did not receive mentoring were part of the experimental sample (100 schools in the control group and 60 schools in the API *Plus* group). As a result, the fourth criterion of the priority-based scheme as described in Section 2.1, which assessed whether communities had previously received a mentor, lost its significance in determining assignment priorities for the expansion of the *Plus* program.

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 in the Web Appendix 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 the 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, another group of 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.

The data collection occurred at the end of the second experimental school year. By that time, two of the original 230 schools in the sample had closed, and another four schools couldn't implement the program 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 in the Web Appendix 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.

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 seventh-grade enrollment data in Chiapas in the Fall of 2016 with the students in our sample

who were in sixth grade during the spring of 2016 and were therefore making the decision of whether to enroll in lower-secondary school. The sample reduces to 468 sixth graders in 182 schools due to the variation in student composition across schools in later grades. 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 survey.[5]

API *Plus* Scale-up Analysis. As indicated in Figure 1, we focus the analysis on the 356 schools that received a mentor during the 2017-2018 school year out of a total of 1,345 eligible schools (see Section 2.3). This sample selection allows for two complete school years between the arrival of the mentors in the communities and the data collection. At the same time, it enables us to assess the program's effectiveness under the *Plus* modality after sufficient time for program operations to fully adapt to the enhanced training module for mentors. Among the 356 schools that were assigned API-*Plus* mentors during the 2017-2018 school year, 270 were not part of our second experiment and 86 were part of the experimental sample. Within the experimental sample, the probability of receiving a mentor during the national scale-up of the *Plus* modality is constant with respect to the treatment arms of the second experiment after controlling for the program eligibility criteria—p-value($Original$) = 0.367, p-value($Plus$) = 0.660.

We employ administrative records detailing the program implementation in Chiapas under the API *Plus* modality and match this information with the quasi-universe of schools located in villages surveyed in the 2020 population census. We utilize two village-level educational outcomes from the Census data, which refer to the school year that started in the Fall of 2019: (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. Unlike other school-survey-based or 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 API *Plus* program when implemented by the government (see Section 4.1).

---

[5]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-3 of the Web Appendix document no program effects on grade repetition and attrition, which suggest that conditioning on grade attainment is not problematic in our context.

2.3 Sample Representativeness

We focus our empirical analysis in 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.[6]

In Table 1 we compare means in observable characteristics between the overall population 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. Instead, as shown in the fifth column of Table 1, 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).

This result underscores the representativeness of the study sample in the second experiment with respect to the population of interest. The fact that the sample of schools in the initial government-led experiment may not offer a comprehensive picture of the intervention's target population in Chiapas provides a further rationale for conducting a second field experiment to assess both mentoring program modalities, namely API *Original* and API *Plus*, within a larger and more representative sample of schools. In the scale-up analysis, we separate the experimental schools from the rest of the non-experimental schools in Chiapas in order to isolate any difference in the impact of the program across situations (field experiment Vs. government implementation of the API *Plus* program).

There are 1,523 schools in Chiapas that are eligible to receive the mentoring program.[7] Of those, we are able to match 1,345 schools (88 percent) with the 2020 population census containing village-level educational outcomes for the quasi-universe of the schools and

---

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

[7]Only schools with six or more enrolled students are eligible for the program.

Table 1: Differences Across Populations

| | Panel A: School Characteristics | | | | |
	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]
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]
Number of Schools	1,523	80	230	1,603	1,753

| | Panel B: Locality Characteristics | | | | |
	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]
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 localities	1,523	80	230	1,603	1,753

Notes: Panel A shows school-level variables from the school census (2010) whereas Panel B displays community-level characteristics from the population census (2010). 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. 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. See Web Appendix A.1 for more details on the data sources.

the localities in Mexico.[8] The probability of being unmatched is orthogonal with respect to the probability of receiving an API-*Plus* mentor during the government's program implementation in Chiapas (p-value=0.634). The 1,345 schools in the matched sample serve approximately 19,000 students, with a total of 165,000 people living in the surrounding communities. Out of these 1,345 schools, 356 received a government mentor by the 2017-2018 school year while the remaining 989 localities did not receive mentors and are used as a comparison group, based on the assignment's priority criteria (see Figure 1). This variation enables us to estimate the program's impact at scale by comparing educational outcomes

---

[8]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. 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 October, 2023.

from the 2019 school year, as recorded in the census, coinciding with the termination of the 2-year program cycle for those schools. The matched-census sample comprises 1,161 non-experimental schools and 184 schools previously engaged in the second experiment. These two sub-samples maintain their representativeness in terms of observable characteristics in relation to the overall targeted populations in Chiapas (see Web Appendix Table B-4).[9]

Program assignment based on specific priority-based rules often fails to adequately represent the entire population of potential beneficiaries of the intervention. The assignment of the mentors across schools in our second experiment is randomized independently of the criteria that will later dictate program assignment priorities at a larger scale. The estimates of program impacts on student outcomes reported in Web Appendix Table B-5 do not exhibit patterns of heterogeneity based on the information underlying these official criteria. This evidence is consistent with the hypothesis that the estimated effect of the *Plus* intervention as implemented by the government, which are derived from the 2017-2018 assignment of the mentors across eligible communities, is representative of the impact under the full scale implementation of the program.

3 The Impact of the Mentoring Program on Children

In this section, we assess the impact of two different mentoring program modalities on various measures of children's academic outcomes. We provide empirical evidence supporting the ineffectiveness of the API *Original* by analyzing the results of two independent field experiments. Subsequently, we quantify the positive effects of the API *Plus* in the randomized evaluation as well as during the government's implementation of the mentoring intervention.

3.1 Empirical Model

We analyze the two experiments through linear 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 of the Web Appendix). For this reason, we include the indicator for indigenous language in the regression analysis of the second experiment. All

---

[9]In our scale-up sample, forty-four schools part of the control group in the second experiment also did not receive a mentor during the national scale-up of the *Plus* modality.

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 encompass the vast majority of program-eligible schools within the state of Chiapas (see Section 2.2). 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 leverage the variability in program assignments across communities, determined by the priority-based mechanism described in Section 2.1, to assess the impact of the API *Plus* on a large scale. To do this, we estimate the following linear regression model:

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

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 in the school year 2017-2018, and zero otherwise. The vector $\boldsymbol{X}_j$ consists of indicator functions for the four criteria used to determine the differential priority across eligible localities/schools to receive the mentors (see Section 2.1). We also 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 additional determinants of the assignment of the mentors across localities. Finally, we include in the vector $\boldsymbol{X}_j$ an indicator variable for prior exposure to the API *Original* modality during the period 2009-2015. The parameter of interest, α_1, represents the effect of exposure to the mentoring program during the government implementation on the outcome of interest.

The underlying identification assumption for unbiased and consistent estimation by OLS of α_1 in equation (1) is that the assignment of the program outside of the experiments 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 are assumed to be similar in terms of unobserved characteristics. We provide

16

two pieces of evidence that should bolster the credibility of this conditional independence assumption in our setting. First, we cannot reject the joint null hypothesis of no differences in observable characteristics at school and locality-level based on the school assignment of the mentoring program during the year 2017-2018, after conditioning on the priority criteria (see Web Appendix Table B-6).[10] Second, we run some placebo tests using the school-level standardized achievement test scores collected before the conversion of the mentoring program under the *Plus* modality. Web Appendix Table B-7 displays the results. The assignment of the mentoring program outside of the experiments 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-7 are very small and statistically insignificant.

To curb the possibility of detecting false positives, we go beyond the conventional asymptotic inference by employing three additional procedures. Firstly, we present *p*-values based on randomization inference, which are accurate irrespective on the number of sampling units or 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 and the scale-up sample. Secondly, given the extensive range of hypotheses explored throughout our analysis, we provide adjusted *p*-values that account for multiple hypothesis testing across various outcome families (List et al., 2019).[11] Thirdly, building upon the insights in Maniadis et al. (2014); Al-Ubaydli et al. (2020), we leverage the value of replication by conducting two independent randomized trials within the same program environment (API *Original*) as well as by contrasting evidence on the impact of the API *Plus* modality under different situations (field experiment and government roll-out). For each program modality, we calculate *p*-values for joint null hypotheses across the different study samples used throughout the analysis.[12]

---

[10]The only covariate that shows a significant difference at 5% level is of whether the locality can be accessed with a road ("Locality Access without Road"). The inclusion of this extra covariate in our regression model (1) does not affect the estimated effect of the program on children's outcomes.

[11]The Romano–Wolf correction (Romano and Wolf, 2005a,b, 2016) asymptotically controls the family-wise error rate, that is, the probability of rejecting at least one true null hypothesis among a family of hypotheses under test. This correction is considerably more powerful than earlier multiple-testing procedures, given that it takes into account the dependence structure of the test statistics by re-sampling from the original data.

[12]To test hypotheses across study samples, we employ Fisher's combined probability test: $-2 \sum_{i=1}^{k} log(p_i) \sim \chi_{2k}^2$, where $p_i \sim U[0,1]$ is the *p*-value for the ith hypothesis test and k is the number of independent study replications being combined. This is akin to the joint statistical significance test commonly used in meta-analyses.

3.2   Evidence on API *Original*

Tables 2 and 3 display the impacts of the *Original* modality on children's schooling outcomes 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 of academic achievement. For the second experiment, we collect our own measures of cognitive and socio-emotional skills (first to fourth columns of Table 3), as well as a measure of educational attainment (fifth column of Table 3).[13]

In spite of the differences in the measurement of children's academic achievement, 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 on children in the first experiment ranges from positive to negative and is not statistically different from zero. The size of the estimated treatment effect on the overall index for academic 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.[14]

Effect sizes are positive and slightly more precise in the second experiment (see the first row of Table 3), 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 0.13 standard deviations—a non-negligible effect size that is nonetheless not statistically different from zero (p-value=0.24, after adjusting for multiple hypotheses testing). The effect of the *Original* modality of the mentoring program on the transition rates to lower secondary school are shown in the last two columns of Table 3. The estimated effect sizes are noisy, with an increase of 7-8 per-

---

[13]The national test that we employ in the first experiment (ENLACE) was administered to all Mexican students in grades three through six through the year 2013 (see Web Appendix A.1). The test was terminated in 2014, and so we cannot use it as a source of measurement for the academic achievement of the children that participated in the second experiment. Another national standardized test was administered by the National Institute for the Evaluation of Education (INEE) starting in 2015, the PLANEA National Plan for Learning Evaluation, although it was collected only in selected grades and in a random sample of students within schools.

[14]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 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 Web Appendix A.1. The dependent variables are standardized with respect to their means and the standard deviations in the control group. The *p*-values reported in brackets refer to the conventional asymptotic standard errors. The *p*-values reported in braces are computed using randomization inference (randomization-*t*). The *p*-values reported in parentheses are adjusted for testing the null impact of API Original across the four 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.

centage points out of a basis of 62 percent enrollment rate in seventh grade in the control group.

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

3.3 Evidence on API *Plus*

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 schools that receive the *Plus* mentors increase their reading scores by 0.32 standard deviations (*p*-values ≤ 0.001). Quantitatively, the API *Plus* effect is approximately 2.5 times higher than the effect of the API *Original*. The difference between the two program effects is statistically different from zero after adjusting for multiple hypotheses testing (*p*-value=0.045). 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

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	0.126	0.073	0.081
	[0.104]	[0.455]	[0.418]	[0.182]	[0.255]	[0.519]
	{0.134}	{0.486}	{0.446}	{0.222}	{0.281}	{0.567}
	(0.150)	(0.558)	(0.558)	(0.240)	(0.311)	(0.478)
API *Plus*	0.315	0.237	0.199	0.368	0.124	0.298
	[0.001]	[0.008]	[0.022]	[0.001]	[0.074]	[0.030]
	{0.001}	{0.014}	{0.032}	{0.001}	{0.089}	{0.052}
	(0.001)	(0.005)	(0.011)	(0.001)	(0.030)	(0.030)
Original = Plus	[0.043]	[0.043]	[0.178]	[0.021]	[0.469]	[0.134]
	{0.077}	{0.112}	{0.221}	{0.024}	{0.568}	{0.230}
	(0.045)	(0.045)	(0.100)	(0.024)	(0.372)	(0.156)
Number of Schools	224	224	224	224	182	76
Number of Observations	1044	1044	1045	1045	468	106

Notes: This table shows OLS estimates and the associated p-values on student outcomes measured after two academic years of exposure to the API program under the second experiment designed and implemented by the authors in collaboration with the government. For detailed descriptions of the test scores used in this table, see Web 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 Web Appendix A.1). The p-values reported in brackets refer to the conventional asymptotic standard errors. The p-values reported in braces are computed using randomization inference (randomization-t). The p-values reported in parentheses are adjusted for testing each null hypothesis (null impact of API *Original*, API *Plus*, and the comparison) for the each family of outcomes (survey-based and admin records) through the step-wise procedure described in Romano and Wolf (2005a,b, 2016). All p-values account for clustering at the school level.

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 at the margin of statistical significance (p-value = 0.100), the larger effect of the *Plus* modality is consistent with qualitative evidence documenting that mentors with enhanced training acquired more effective skills to best deal with children's emotions during the bimonthly sessions.[15] 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 ≤ 0.001), and statistically different at the 5 percent

---

[15]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. Web Appendix A.3 reports more details about these interviews. Tables A-1 and A-2 show that the characteristics of these survey respondents are broadly comparable to those of the mentors and the local instructors in the second experiment.

level from the effect of the *Original* modality.[16]

The last two columns in Table 3 reports the estimated effects on the average transition rate to secondary school. Less than two-thirds of the sixth graders in the control group enroll in seventh grade, while the corresponding national average is 95 percent. The API *Plus* modality increases the probability of a child's enrolling in seventh grade by 12 percentage points. This effect on education attainment is precisely estimated (p-value=0.030, after accounting for multiple hypothesis testing) and 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 size of the effect more than doubles when we focus on the sub-sample of over-aged sixth graders (13 years old or more, sixth column). Given recent longitudinal evidence on the labor market returns associated to the primary-to-secondary schooling transition in Mexico (Araujo and Macours, 2021), our estimated effect sizes on schooling attainment are particularly important in terms of life-cycle opportunities.

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. 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. The first column of Table 4 shows that the program increases the fraction of children who enroll in secondary education by 5.6 percentage points (p-value = 0.020) for the sample of schools that did not previously participate in the second experiment. This represents an increase of 7.6 percent with respect to the sample mean. For the schools that were previously part of the experiment, the impact of receiving the program during the government implementation is larger (+9.1p.p., p-value = 0.041, see third column in Table 4). These effects of the mentoring intervention during the government's program implementation are statistically similar across the two school sub-samples and they are in line with the experimental findings of the API *Plus* modality on the enrollment in seventh grade documented in the fifth column of Table 3.[17]

---

[16]In Web Appendix Table B-8 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 Web Appendix Table B-9 we report the effects of the two program modalities for each individual component of the socio-emotional score.

[17]The census-based information reported in Table 4 represents the locality-level stock (rates) of children enrolled in secondary school in a given year, while our measure of enrollment in seventh grade (see Table 3) represents the flow of new students enrolling in any secondary schools.

Table 4: Children's Achievement and 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]
	{0.011}	{0.014}	{0.021}	{0.075}
	(0.020)	(0.020)	(0.041)	(0.075)
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 Web Appendix A.1. The p-values reported in brackets refer to the conventional asymptotic standard errors. The p-values reported in braces are computed using randomization inference (randomization-t). The 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). All p-values allow for heteroskedasticity of unknown form.

The estimates of the impact of the government-run API *Plus* mentoring modality on child literacy are displayed in the even columns of Table 4. This is another relevant education outcome for the disadvantaged communities that are targeted by the intervention, which is akin to the achievement test scores reported in Tables 2 and 3. In our sample, 13 percent of school-aged children are still illiterate. After two years of exposure, we find that villages that received mentors display a 2.8 percentage points (p-value = 0.020) 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 sub-sample of experimental schools is quantitatively similar, although a bit noisier (+3.5 percentage points, p-value = 0.075).

Both across different samples of schools under the same program modality and across different situations for the same sample, our findings overwhelmingly support the notion that the API *Plus* program has improved the education outcomes for children in these disadvantaged communities. This interpretation is corroborated by the test statistics for the joint hypothesis of no effect for this modality of the mentoring program on schooling achievement and attainment, which are highly significant (p-values = 0.0001 and 0.001, respectively). Furthermore, the program's impacts endure beyond the conclusion of the two-year intervention, enhancing its potential for scalability. Leveraging the initial randomization from the second experiment, Web Appendix Figure B-1 shows that the effect of the *Plus* program on secondary school enrollment continues beyond the two-year time frame of the program's cycle. Web Appendix Figure B-2 instead shows that extended exposure to the program beyond the

initial two-year cycle further enhances the program's impact.

4   Challenges and Pathways to Scale

Despite the substantial 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 discusses various mechanisms that can cause a voltage drop in the new situation (Al-Ubaydli et al., 2020). In this section, we first outline specific 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. We next study the possible mechanisms behind the success of the *Plus* modality of the mentoring intervention using an array of survey modules collected during the two field experiments. This analysis aims to shed light on the pathways that likely facilitated the program's scalability.

4.1   Program Implementation Fidelity

The fidelity of the training and supervision might fall during the national program implementation even when scaling-up does not require hiring and training an increased number of service providers. There were two differences in terms of how mentors were recruited and assigned to communities within the randomized trial when compared to the on-going government intervention (API *Original*). 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. The extent to which these implementation changes were later adopted by government during the scale-up of the API *Plus* modality can potentially influence the effectiveness of the mentoring service.

We begin by examining the extent to which the population of mentors is similar between the experiment and the scale-up. To do so, we integrate the survey data on mentors from the second experiment with the program-roster data of mentors during the scale up.[18] De-

---

[18]The survey data comprises a total of 139 mentors, while the program-register data includes 441 mentors. 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.

spite limited set of common variables across these two datasets, Web Appendix Table B-10 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 lends at least some empirical support to the notion 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 modules 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 data from 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 Web Appendix A.2 for further details on the surveys) in order to test the hypothesis that mentoring practices underwent significant changes during the government's scale-up.

The estimates displayed in Figure 2 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 indicates 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.[19] We observe 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 one minute less during their last visit

---

[19]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.

Figure 2: Differences in the Mentoring Practices Between Experiment and Scale-up

Panel A: Quantity of Mentoring   Panel B: Quality of Mentoring
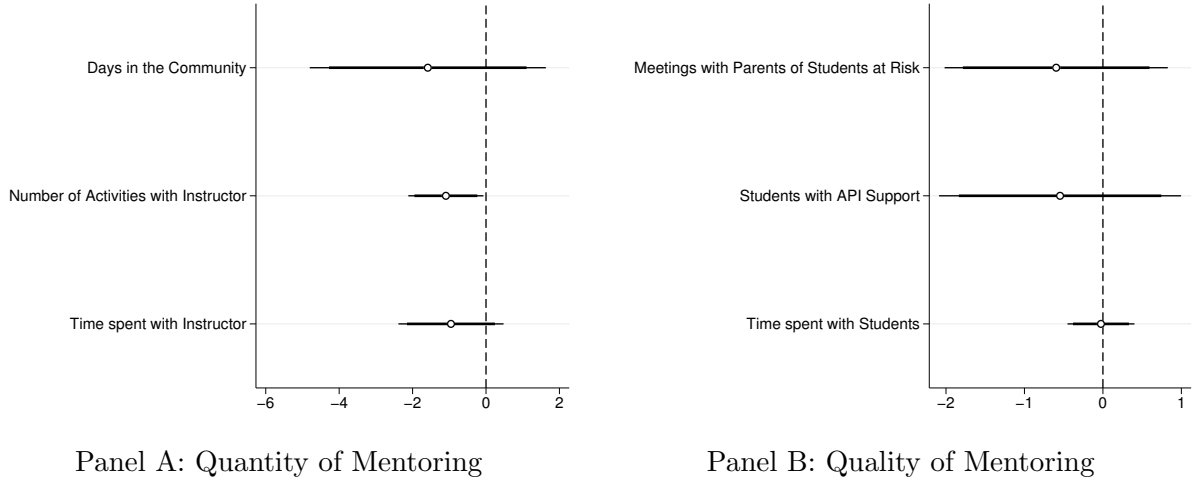
Notes: The figure shows the comparison in the quantity and quality of API-*Plus* mentors between the second experiment and the government implementation. This information is collected during the surveys of the local instructors, in the school years 2015-2016 and 2018-2019. Each dot in the figure represents 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 Web Appendix Table B-11. All the regressions include the same set of controls as in Table 4.

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 government scale-up of the API *Plus* intervention. Finally, when considering the time that mentors spent with children during their last visit in the communities, our results suggest no change in mentoring practices. Mentors spend the same amount of time (minutes) with students in the field experiment compared with the scale-up regime.

The continuity of school services is vital for maintaining the program's effectiveness, as schools serve as the conduit for delivering the mentoring service. Consequently, the occurrence of school closures can significantly disrupt the program. Some institutional details make school closures more salient during the government implementation of the mentoring intervention when compared to the experimental situation. For example, the decision of closing schools is determined by the parent association with a vote. Whenever the number of

students enrolled drops below six the school ceases to operate by default, unless the majority of parents oppose by vote. 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 of the study period. As a result, only two schools closed in the experimental sample of 230 schools, 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 a size below the median.

On one hand, it is conceivable that the program could fail when implemented on a larger scale due to the potential of school closures. 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. To gain insights into this potential threat to scalability, we adopt the same regression model (1) and the same sample of schools previously used to evaluate the mentoring intervention on children's outcomes at scale (see Table 4). In this analysis, 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 large effect, in percent terms, on school closures.[20] 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 conversion of the program from a field experiment to government implementation has the potential to create significant disruptions. The evidence does not support the notion of a severe decline in the mentoring practices in the scale-up phase. Our previous findings on the impact of the *Plus* program on educational outcomes documented in Table 4 and the evidence on school closures discussed in this section are consistent with the hypothesis that

---

[20]For the overall sample of schools in the scale-up analysis (both experimental and non-experimental schools), only 1.6% of the schools with API-*Plus* mentors are permanently closed after two years of exposure to the government-run program against 9.1% of closures among schools without mentors.

Figure 3: The Impact of the API *Plus* Program on School Closures
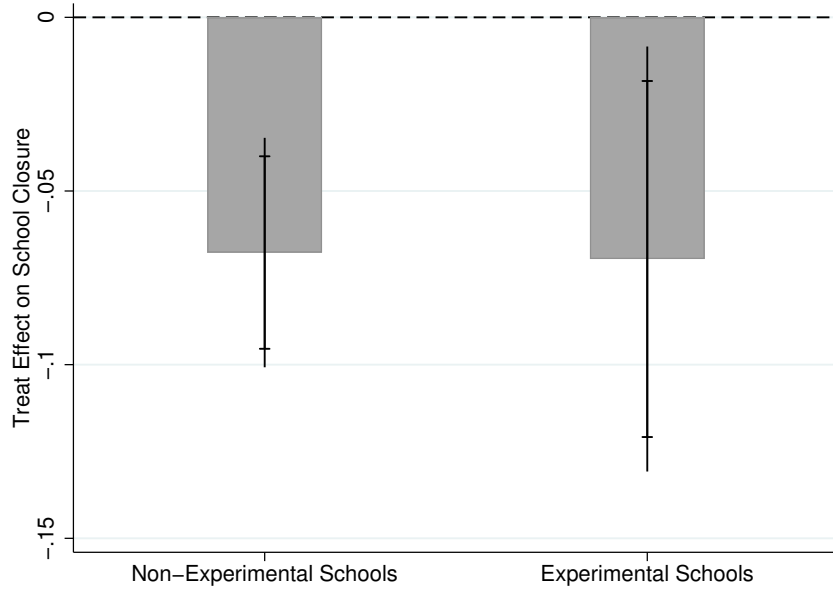
Notes: The bars in the figure represents the OLS estimates of the assignment to the API program during the government implementation of the *Plus* modality (same as in Equation 1) on the rate of school closures as measured over the subsequent two years. Vertical lines overlaid on each bar display the 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 Web Appendix Table B-12.

the program's underlying effectiveness endures during the government implementation.

4.2 *Plus* vs. *Original*: Channels

In this sub-section, we study the possible mechanisms behind the success of the *Plus* modality of the mentoring intervention using an array of survey modules collected during the two field experiments.[21] Table 5 presents the average impact of the program on GLS-weighted indices of parental investment in their children's education (see Web Appendix A.2).[22] Panel A

---

[21]As discussed in Section 2.3, the sample of schools of the second experiment is largely representative of the broader population of schools in the State of Chiapas in terms of observable characteristics (see Tables 1 and B-4) as well as in terms of program impacts (see Table 4 and Figure 3).

[22]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. Web Appendix Table B-13 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.

Table 5: Parental Investment

| | Panel A: First Experiment | | | |
	Engage at School	Manage School Resources	Engage With Child	Overall Index
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			
	Engage at School	Manage School Resources	Engage With Child	Overall Index
API *Original*	-0.188	-0.124	0.167	-0.034
	[0.049]	[0.176]	[0.015]	[0.684]
	{0.070}	{0.216}	{0.030}	{0.709}
	(0.058)	(0.197)	(0.015)	(0.630)
API *Plus*	0.217	0.087	0.353	0.359
	[0.034]	[0.344]	[0.001]	[0.001]
	{0.047}	{0.393}	{0.001}	{0.001}
	(0.037)	(0.247)	(0.000)	(0.001)
Original = Plus	[0.001]	[0.056]	[0.029]	[0.001]
	{0.000}	{0.058}	{0.171}	{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 Web Appendix A.2. The p-values reported in brackets refer to the conventional asymptotic inference. The p-values reported in braces are computed using randomization inference (randomization-t). The 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.

displays the estimates of the *Original* modality in the first experiment, while Panel B shows the corresponding figures for both the *Original* and *Plus* modality in the second experiment. Under the API *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.

Parents appear to be systematically more invested in their children's education activities under the *Plus* modality of the mentoring program. The estimates reported in Panel B of Table 5 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. After correcting inference for multiple hypotheses testing, three out of four coefficients are statistically significant at the 5 percent level, with a very large effect size for the overall index of parenting practices of 0.36 standard deviations. We can reject the null hypothesis of equal treatment effects on all four parental outcomes.

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 observed effect of the *Plus* modality on parental investment.

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—parents in the control group cannot be part of this analysis by design.[23] The evidence shows a clear pattern in spite of quite noisy estimates due to missing observations and the reduced sample size. 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 third and fourth 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 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 5–10 percent level depending on the inference procedure (fifth column).

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

---

[23]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]
	{0.049}	{0.092}	{0.024}	{0.637}	{0.749}	{0.118}	{0.315}
	(0.070)	(0.070)	(0.040)	(0.846)	(0.846)	(0.295)	(0.542)
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 Web 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).

(see Web Appendix A.2).[24] 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 displayed in the third column of Panel B shows a sizable and significant effect (*p*-value = 0.040) 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. The estimates in the last four columns of Panel B cover different aspects of the parent-child relationship, particularly emotional practices. Our results show positive effects in this area of parent-mentor interactions, although these effects are statistically insignificant.

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 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 parenting skill training modules and the peer-to-peer sessions facilitated by mentors. Instead, the extra week of initial training is focused on pedagogical practices targeted to children at school. Qualitative evidence seems indeed to corroborate

---

[24]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 attrition of survey participation is unrelated to the treatment assignment (*p*-value = 0.514).

this hypothesis. We report here a few quotes from mentors who have participated to the training sessions of the *Plus* modality (see Web Appendix A.3 for more details):

- *"During the workshops I was told that I should be able to adapt to the context of the community and understand the local living arrangements in order to establish a dialog with the parents without modifying what they conceive as their environment."*

- *"It was recommended that we pay frequent home visits so as to establish a relationship with the parents and gain their trust."*

- *"[The workshops] exposed us to effective strategies of other mentors [for dealing with parents] that we could try and implement in our community."*

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 main tasks of the mentors in the school communities beyond parental involvement: (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 Web Appendix Table B-14 suggest that there is no differential effect across achievement 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 (see also Web Appendix Figure B-3).

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. Web Appendix Table B-15 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 Web Appendix A.2).[25] The results show erratic patterns of positive and negative signs with no statistically significant effects of either API modality.

---

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

In summary, differences in effectiveness across modalities of remedial education sessions or variations in pedagogical support for instructors are unlikely to account for the success of the *Plus* program. The available evidence suggests that a more active parental involvement, which was likely triggered by enhanced parent-mentor interactions, played a central role.

4.3 Parents as Potential Means of Scalability

Given the results documented in Sections 4.1 and 4.2, we next explore the link between school closures and the engagement of parents with the school community under the randomized program assignment. The functioning of the community-based schools under study 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 (see Section 2.1). Because school closures can undermine the success of the API *Plus* mentoring modality outside of the experimental conditions, this effectively implies that parents can play a crucial role in the scalability of the mentoring program.

We explore this possibility by examining whether or not the contrasting responses in parental investment across the two mentoring interventions, as shown in Table 5, are 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 two years after the program intervention. 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 the effectiveness of community-based educational programs.

The second column of Table 7 shows that the *Plus* modality, which substantially boosts parental engagement during the experiment, has a large and significant impact on school closures two years after the *Plus* modality was adopted by the government. Schools are 8.3 percentage points less likely to close (*p*-value=0.030). 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.

The IV estimates shown in the third column of Table 7 go a step further and quantify the

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. Instrument)			13.833

Notes: The first two columns of this Table report the estimates at the school level for the reduced-form effects of the API modalities during the two experiments on the probability of school closures two years after the program intervention. The third column reports the instrumental variable estimates at the individual level of the impact of parental engagement on school closures, where 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 variable "Overall Parental Engagement" is the same variable used in the last column of Table 5. The asymptotic p-values reported in brackets allow for heteroskedasticity of unknown form in the first two columns, while they are clustered at the school level in the third column.

causal effect of parental engagement on the probability of school closures. An increase of 0.1 of a standard deviation in the overall parental engagement index is associated with a reduction of 2.2 percentage points in the probability that their children experience a school closure (p-value=0.021). We propose three main reasons why it seems plausible to assume that parents are the primary channel through which the *Plus* modality of the API program influences school closures. First, contextual information points to the role of the parental association in deciding school closures. The role of parents in ensuring continuity in schooling activities clearly emerges in the qualitative evidence.[26] Second, our findings reported in Section 4.2 reject the hypothesis that other behavioral responses by teachers and students may mediate the effect of the API-*Plus* program on school closures. Third, the absence of any impact from the *Original* mentoring program in both independent field experiments on parental investments and school closures (see Tables 5 and 7) serves as further corroboration

---

[26]As reported by the local instructors, engaged parents may have more at stake in keeping the schools open as they invest in durable goods for the local school: *"[Parents] help manage the school and contribute by improving the fencing, painting the walls, fixing the toilets, as well as buying school materials." "[Parents] serve the needs of the school with construction works and they provide food to the local instructor."* 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 the children are able to make progress in the schooling activities without setbacks."*

that when parental investments are not boosted by the intervention, the underlying impact on school closures is muted.

Taken together, the evidence presented in this section is consistent with the hypothesis that the effectiveness of the mentoring intervention during the large-scale implementation of the program likely depends on the active involvement of parents in educational activities. Within the new program modality, parents not only increased their interactions and investment with children—a shared result among past successful interventions (Heckman and Mosso, 2014; Zhou et al., 2021; García and Heckman, 2023)—but also they intensified their engagement at the school and community level. Parental responses are shown to prevent schools from closing, which would otherwise pose a threat to the scalability of the program during government implementation.

5   Conclusion

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. In the context of a school mentoring program in the state of Chiapas, Mexico, we show that relatively minor variations in the training content of mentors can lead to large changes in schooling outcomes. While the government's original implementation of the program proves to be largely ineffective, an alternative approach that prioritizes mentors' ability to effectively interact with and engage parents was successful in enhancing test scores and improving educational attainment for the students in our sample. The magnitudes of the estimated impacts are comparable across situations (field experiment versus government implementation) as well as across study samples (experimental schools versus the rest of the schools in Chiapas).

We acknowledge the limitations of the empirical analysis in addressing the "vertical" aspect of scaling, as outlined in List (2022), which would involve supply-side considerations for the implementation of the program at a larger scale. One possible contextual shortcoming is that the intervention under study relies on university graduates as mentors. This feature may hinder the extent to which the program can be scaled up in settings where human resources with relatively high levels of human capital are scarce. Finally, while we underscore the pivotal role that local communities and parents play in promoting the success of education interventions, our evidence is merely suggestive on the channels through which this particular

program remains successful when implemented at scale. More research is needed to further explore the complex social dynamics triggered by large-scale interventions.

References

Agostinelli, F. (2018). Investing in children's skills: An equilibrium analysis of social interactions and parental investments. Manuscript, Department of Economics, University of Pennsylvania.

Agostinelli, F., M. Doepke, G. Sorrenti, and F. Zilibotti (2020, April). It takes a village: The economics of parenting with neighborhood and peer effects. Working Paper 27050, National Bureau of Economic Research.

Al-Ubaydli, O., M. S. Lee, J. A. List, C. Mackevicius, and D. Suskind (2021). How can experiments play a greater role in public policy? twelve proposals from an economic model of scaling. *Behavioural Public Policy 5*(1), 2–49.

Al-Ubaydli, O., J. A. List, and D. Suskind (2020). 2017 klein lecture: The science of using science: Toward an understanding of the threats to scalability. *International Economic Review 61*(4), 1387–1409.

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

Araujo, M. C. and K. Macours (2021, October). Education, Income and Mobility: Experimental Impacts of Childhood Exposure to Progresa after 20 Years. PSE Working Papers halshs-03364972, HAL.

Attanasio, O., H. Baker-Henningham, R. Bernal, C. Meghir, D. Pineda, and M. Rubio-Codina (2022, 01). Early Stimulation and Nutrition: The Impacts of a Scalable Intervention. *Journal of the European Economic Association 20*(4), 1395–1432.

Attanasio, O., S. Cattan, and C. Meghir (2022). Early childhood development, human capital, and poverty. *Annual Review of Economics 14*(Volume 14, 2022), 853–892.

August, G., M. Bloomquist, S. Lee, G. Realmuto, and J. Hektner (2006, 07). Can evidence-based prevention programs be sustained in community practice settings? the early risers? advanced-stage effectiveness trial. *Prevention science : the official journal of the Society for Prevention Research 7*, 151–65.

Banerjee, A. V., R. Banerji, J. Berry, E. Duflo, H. Kannan, S. Mukerji, M. Shotland, and M. Walton (2017, November). From proof of concept to scalable policies: Challenges and solutions, with an application. *Journal of Economic Perspectives 31*(4), 73–102.

Becker, G. S. (1962). Investment in human capital: A theoretical analysis. *Journal of Political Economy 70*(5), 9–49.

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

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

Bold, T., M. Kimenyi, G. Mwabu, A. Ng'ang'a, and J. Sandefur (2018). Experimental Evidence on Scaling Up Education Reforms in Kenya. *Journal of Public Economics 168*(C), 1–20.

Cameron, L., S. Olivia, and M. Shah (2019). Scaling Up Sanitation: Evidence from an RCT in Indonesia. *Journal of Development Economics 138*, 1–16.

Caron, E., K. Bernard, and A. Metz (2021). 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.

CONEVAL (2018). Medición de pobreza 2008-2018, estados unidos mexicanos.

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

Davis, J., J. Guryan, K. Hallberg, and J. Ludwig (2021). 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.

Doepke, M. and F. Zilibotti (2017, September). Parenting With Style: Altruism and Paternalism in Intergenerational Preference Transmission. *Econometrica 85*, 1331–1371.

Duflo, A., J. Kiessel, and A. M. Lucas (2024, 02). Experimental Evidence on Four Policies to Increase Learning at Scale. *The Economic Journal*, ueae003.

García, J. L. and J. J. Heckman (2023). Parenting promotes social mobility within and across generations. *Annual Review of Economics 15*(1), null.

Heckman, J. (1992). Randomization and social policy evaluation. In *Evaluating Welfare and Training Programs. Edited by C. F. Manski and I. Garfinkel*. Harvard University Press.

Heckman, J. and J. Zhou (2021). Interactions as investments: The microdynamics and measurement of early childhood learning. Working paper, Center for the Economics of Human Development, University of Chicago.

Heckman, J. J. and S. Mosso (2014). The Economics of Human Development and Social Mobility. *Annual Review of Economics 6*, 689–733.

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

List, J. A. (2024). Optimally generate policy-based evidence before scaling. *Nature 626*(7999), 491–499.

List, J. A., F. Momeni, M. Vlassopoulos, and Y. Zenou (2023, May). Neighborhood Spillover Effects of Early Childhood Interventions. CEPR Discussion Papers 18134, C.E.P.R. Discussion Papers.

List, J. A., A. M. Shaikh, and Y. Xu (2019, December). Multiple Hypothesis Testing in Experimental Economics. *Experimental Economics 22*(4), 773–793.

Maniadis, Z., F. Tufano, and J. A. List (2014, January). One swallow doesn't make a summer: New evidence on anchoring effects. *American Economic Review 104*(1), 277–90.

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

Muralidharan, K. and P. Niehaus (2017). Experimentation at Scale. *Journal of Economic Perspectives 31*(4), 103–124.

Muralidharan, K. and A. Singh (2020, November). Improving public sector management at scale? experimental evidence on school governance india. Working Paper 28129, National Bureau of Economic Research.

O'Brien, P. C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics 40*(4), 1079–1087.

Romano, J. P. and M. Wolf (2005a, March). Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing. *Journal of the American Statistical Association 100*, 94–108.

Romano, J. P. and M. Wolf (2005b, July). Stepwise Multiple Testing as Formalized Data Snooping. *Econometrica 73*(4), 1237–1282.

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

Zhou, J., A. Baulos, J. J. Heckman, and B. Liu (2021). 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.