Enhancing Human Capital in Children: A Case Study on Scaling

Francesco Agostinelli

University of Pennsylvania

Ciro Avitabile

World Bank

Matteo Bobba

Toulouse School of Economics

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.

We are grateful to the Consejo Nacional de Fomento Educativo for the collaboration throughout this project, and in particular to Carmen Gladys Barrios Veloso and Janet Venancio. We thank the editor, John List, and the anonymous referees for their comments and feedback that greatly improved the paper. We also thank Jere Behrman, Jim Heckman, Giuseppe Sorrenti, and Stephane Straub for helpful comments and discussions. We are indebted to Alonso Sanchez for the contribution in the early stages of this project and to Miguel Angel Monroy for excellent research assistance. Ciro Avitabile acknowledges financial support for data collection from the Strategic Impact Evaluation Fund of the World Bank and the Consejo Nacional de Evaluación de la Política de Desarollo Social. Matteo Bobba

Electronically published December 6, 2024

Journal of Political Economy, volume 133, number 2, February 2025.

https://doi.org/10.1086/732886

I. 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 involved in scaling 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 2 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 program's ineffectiveness served as a catalyst, prompting an effort 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 changes to the training module, mentors began to 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 achievement among different socioeconomic groups (Cunha, Heckman, and Schennach 2010; Heckman and Mosso 2014; Attanasio, Cattan, and Meghir 2022).

Science guides policy. After the release of compelling evidence regarding the effectiveness of the new modality in delivering significant positive

acknowledges financial support from the Agence française de développement, the Horizon 2020 Marie Skłodowska-Curie Actions Research and Innovation Staff Exchange project GEMCLIME 2020 GA no. 681228, and the Agence nationale de la recherche under grant ANR-17-EURE-0010 (Investissements d'Avenir program). This study is registered in the American Economic Association Randomized Controlled Trial Registry, and the unique identifying number is AEARCTR-0001645.

effects on children's outcomes and parental investment, the government made the decision to adopt the improved program modality. Our comprehensive analysis of the program's implementation 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 was drawn from two independent field experiments and the subsequent government scale up of the effective program modality. The first experiment was directly carried out by the government, immediately following the national implementation of the original mentoring program. Assignment to the program was randomized across 80 program-eligible primary schools. In the second experiment, we randomly assigned both the original and the new modality, as well as a control group with no mentoring program, across 230 primary schools. After 2 years of the mentoring program, the original modality displayed 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 delivered sizable and significant gains in children's reading scores (+0.32 SD, p-value = 0.001), math scores (+0.24 SD, p-value = 0.005), and socio-emotional scores (+0.20 SD, 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), from a baseline of 62% 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 subsample of the 1,161 localities outside of the

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

experimental sample—which had never received the new program modality before and which encompassed approximately 16,000 children enrolled in eligible 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% with respect to the sample mean for the overall sample of schools. Turning to 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 "nonnegotiable" aspects of program designs in the context of such changes in the situation (List 2022). Failure to consider these critical elements during the widespread implementation of the intervention has the potential to not only reduce but even completely eliminate the program effects documented under experimental conditions (Al-Ubaydli, List, and Suskind 2020; Al-Ubaydli et al. 2021; Caron, Bernard, and Metz 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 what List (2022) called "voltage drop" of the program results from the experimental design's trade-off between real-world applicability and the purity of the evaluation. A significant challenge faced by educational programs in this context is the occurrence of frequent school closures, but the intense monitoring during the field experiment from the research team had minimized the number of such closures in the period of the evaluation. The school-closure rate is notably high, averaging 11% before the experimental evaluation. In contrast, only two schools out of 230 closed during the randomized trial. To the extent that the continuity of 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 closures. 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 affect crucial decisions concerning resource allocation, investments, and even the ultimate determination of whether a school is to remain open. 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 with both the school and directly with the child. Our measure of parenting practices increases by 0.36 SD 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 shape 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 treatment effects of the new modality of the mentoring program operate only through parental engagement, instrumental variables (IV) estimates document that an increase of 0.1 SD 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 the absence of parental engagement and illustrates the importance of such engagement to program success.

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 to not 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. Recent and growing evidence suggests that effects observed in small-scale randomized trials can be challenging to replicate when interventions are implemented at a larger scale (Bold et al. 2018; Cameron, Olivia, and Shah 2019; Muralidharan and Singh 2020;

Bobba, Frisancho, and Pariguana 2023). Our field experiment was designed and implemented during the rollout 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 practical considerations and implementation realities of the policy under study (Banerjee et al. 2017; Muralidharan and Niehaus 2017; Duflo, Kiessel, and Lucas 2024; List 2024).

We analyze two independent field experiments on different and representative samples, as well as a larger, nonexperimental sample of schools. Drawing joint inferences from these samples harnesses the statistical power of our findings (Maniadis, Tufano, and List 2014; Allcott 2015; Al-Ubaydli, List, and Suskind 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 interactions between parents and mentors/home visitors in boosting treatment effects of home-visiting programs (Heckman and Zhou 2021; Zhou et al. 2021; García and Heckman 2023) and has indicated 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, our analysis does not address the "vertical" aspects of scaling (List 2022). 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.

II. 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; taken together, these serve 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 app. A for more details; apps. A and B are available online).

A. *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% of the roughly 99,000 primary schools across the 31 Mexican states. The largest presence of CONAFE schools is in Chiapas, the Mexican state with the highest incidence of poverty in the country (CONEVAL 2018). CONAFE primary schools typically have a single multigrade classroom with 15 students on average (for brevity, throughout the text we will refer to CONAFE primary schools as *schools*).

The local instructors consist predominantly of community residents aged between 15 and 29 years old, who typically have minimal formal training (if any) as teachers. As a result of the very low compensation and extremely challenging conditions, about one quarter of the instructors drop out before completing their first school year. Schools frequently face closure because of these challenges. In fact, the average yearly rate of school closures in Chiapas stands at 11%. Parents organize local associations aimed at promoting community education, to which they contribute by maintaining the school's facilities and contributing to the school financially. 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 program (Asesores Pedagógicos Itinerantes; API) as an attempt to improve the quality of education provision in basic education. The program was implemented initially 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, or social services who have previous experience as community instructors and who speak an indigenous language. Prior to beginning work 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 2 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 allows for accommodating only 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% 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 locality, 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 antipoverty 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) 2-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 primarily held after regular instructional hours. The tutoring sessions are offered to the six weakest students in the class, identified through a diagnostic evaluation conducted at the beginning of the school year and an additional exam administered by the mentor. During regular school hours, mentors are tasked with observing and possibly improving 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 2 months in 2-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 *API Original*.

Starting from 2016, the government adopted a new modality of the program, which we will refer to as *API Plus*. API Plus encompasses all the features of API Original but with substantial changes in the training module. First and foremost, it extends the initial training period from 1 week to 2. 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 API Original, each API Plus 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, in collaboration, to devise strategies for more effective interaction and engagement with

[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 teacher support) are directed toward all children in the school.

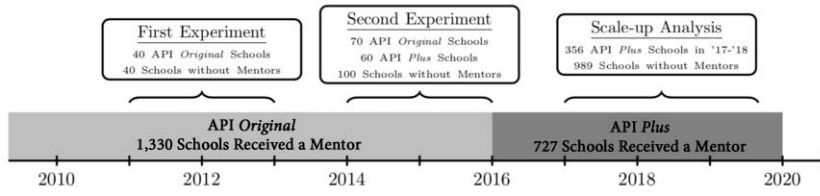

FIG. 1.—The mentoring program in Chiapas and the different study samples.

parents.[3] The cost of API Plus is US$332 per child, compared to US$285 per child for 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, API Original served over 1,300 schools and approximately 18,000 students. Following the transition to API Plus between 2016 and 2020, around 700 schools and 10,000 students received mentoring support. Because of the program's 2-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 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 every school in Chiapas had received a mentor because of the reassignment of mentors across schools. When, in the fall of 2016, the government converted the program to the API 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 2009–2016 (7 school years) is approximately double the cumulative number of schools that received the API Plus modality at least once during 2016–2020 (4 school years).[4]

[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, Cattan, and Meghir 2022; García and Heckman 2023). Moreover, this decision was influenced by the feedback received from the mentors during the initial implementation of the API Original modality. They reported that the interactions with parents were the most challenging aspect of their tasks in the local communities.

[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 fig. 1). At that point, the vast majority of the schools that did not receive mentoring were part of the experimental sample

B. *Experimental Design and Scale-Up Analysis*

Two consecutive and independent randomized evaluations took place following the nationwide implementation of the API 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 sec. III), 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 were selected among those that had never received the mentoring program before. Of those, 62 schools were located in the state of Chiapas and the remaining 18 schools were in the three states of Hidalgo, Queretaro, and Veracruz. Assignment to the mentoring program was randomized at the school level using a block design, with the strata represented by the Mexican states where schools were located. Forty schools were assigned to receive API Original starting from the 2011–2012 school year while the remaining half of the schools were 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 2 school years after the assignment of the API program through the performance in the national standardized test for students in grades three through six. A midline survey records parental behaviors and investments for 208 parents in 73 schools (the enumerators were not able to reach the parents in seven schools). Because of 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 subsample of 73 schools with parental outcomes. Out of the 10 schools that were part of the experimental sample but that we are unable to match in our final sample, five schools are in the treatment group and five are in the control group. Table B1 (tables A1, A2, and B1 to B15 are available online) 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.

(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 sec. II.A, which assessed whether communities had previously received a mentor, lost its significance in determining assignment priorities for the expansion of the API Plus program.

Second experiment.—Two hundred thirty program-eligible primary schools were selected in rural Chiapas from those that had never received the mentoring program before. Assignment of the mentors was carried out using a randomized block design at the school level, with the strata represented by the deciles of the 2012 school average in the national standardized achievement score in the Spanish test. As a result, 60 schools were assigned to receive API Plus mentors starting from the 2014–2015 school year, another group of 70 schools were assigned to receive API Original mentors over the same time period, and the remaining 100 schools were 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 could not implement the program due to high political instability. Out of the six schools that dropped out of the sample, two schools were in the control group, two were in the API Original group, and two were in the API Plus group. The number of schools that took part in the second experiment was 224. Table B2 shows that a large array of predetermined 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 5-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 across 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 sec. II.C). This sample selection allows for 2 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 API Plus modality after

[5] The distribution of missing schools in the analysis of transition to secondary school is 18 schools in the control group, 14 in API Original and 16 in API Plus. Because of the different individual identifiers, we are not able to match this dataset to the survey data. The estimates reported in table B3 document no program effects on grade repetition and attrition, which suggest that conditioning on grade attainment is not problematic in our context.

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 API 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 12 and 14 years old and (ii) the rate of child literacy for children between eight and 14 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 sec. IV.A).

C. 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, List, and Suskind (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 column 5 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

[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 United States 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).

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, 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 nonexperimental 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%) with the 2020 population census containing village-level educational outcomes for the quasi-universe of the schools and 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 based on the assignment's priority criteria, while the remaining 989 localities that did not receive mentors are used as a comparison group (see fig. 1). This variation enables us to estimate the program's impact at scale by comparing educational outcomes from the 2019 school year, as recorded in the census, coinciding with the termination of the 2-year program cycle for those schools. The census-matched sample comprises 1,161 nonexperimental schools and 184 schools previously engaged in the second experiment. These two subsamples maintain their representativeness in terms of observable characteristics in relation to the overall targeted populations in Chiapas (see table B4).[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 was randomized independently of the criteria that would later dictate program assignment priorities at a larger scale. The estimates

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

[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 one primary school. For further details on the census sampling design, please refer to INEGI (2020).

[9] In our scale-up sample, 44 schools that were part of the control group in the second experiment also did not receive a mentor during the national scale-up of the API Plus modality.

TABLE 1
DIFFERENCES ACROSS POPULATIONS

	All Chiapas (1)	First Experiment (2)	Second Experiment (3)	Chiapas vs. First Experiment (4)	Chiapas vs. Second Experiment (5)
	A. School Characteristics				
Average test score (Spanish)	424.503	399.116	431.340	−25.387	6.837
	(56.466)	(32.631)	(60.810)	[.000]	[.139]
Average test score (Math)	414.921	379.165	421.333	−35.756	6.412
	(75.300)	(45.339)	(80.895)	[.000]	[.297]
Students (no.)	14.049	15.507	15.009	1.458	.960
	(8.468)	(8.781)	(6.053)	[.175]	[.037]
Teachers (no.)	1.231	1.333	1.217	.102	−.014
	(.467)	(.505)	(.413)	[.099]	[.638]
Overaged students (share)	.349	.230	.324	−.119	−.025
	(.797)	(.552)	(.659)	[.088]	[.610]
Schools (no.)	1,523	80	230	1,603	1,753

	B. Locality Characteristics				
Total population	118.758	247.280	121.389	128.522	2.630
	(221.648)	(549.923)	(240.562)	[.043]	[.879]
Rate of extreme poverty	.490	.486	.473	−.004	−.017
	(.500)	(.503)	(.500)	[.949]	[.644]
Incidence of social conflicts	.190	.150	.187	−.040	−.003
	(.392)	(.359)	(.391)	[.335]	[.919]
Rate of illiteracy	.313	.321	.295	.008	−.018
	(.160)	(.157)	(.153)	[.662]	[.127]
Labor force participation	.297	.289	.303	−.008	.006
	(.076)	(.071)	(.070)	[.352]	[.259]
Locality access without road	.216	.203	.179	−.013	−.037
	(.411)	(.404)	(.384)	[.777]	[.181]
Water network (Y/N)	.028	.050	.022	.022	−.006
	(.164)	(.219)	(.146)	[.365]	[.578]
Sewage system (Y/N)	.011	.038	.009	.026	−.002
	(.105)	(.191)	(.093)	[.219]	[.712]
Garbage collection (Y/N)	.022	.038	.022	.016	.000
	(.146)	(.191)	(.146)	[.463]	[.994]
Localities (no.)	1,523	80	230	1,603	1,753

NOTE.—Panel A shows school-level variables from the school census (2010), whereas panel B displays community-level characteristics from the population census (2010). Columns 1–3 show means and SDs in parentheses for various characteristics collected before the introduction of the API program. Columns 4 and 5 show asymptotic *p*-values in brackets 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 app. A1 for more details on the data sources.

of program impacts on student outcomes reported in table B5 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.

III. 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 API Original by analyzing the results of two independent field experiments. Subsequently, we quantify the positive effects of API Plus in both the randomized evaluation and during the government's implementation of the mentoring intervention.

A. Empirical Model

We analyze the two experiments through linear regression models on the treatment-assignment indicators for the API Original and the API Plus modalities after 2 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 table B2, panel B). For this reason, we include the indicator for indigenous language in the regression analysis of the second experiment. All models further include the strata control variables that account for the block randomization designs, as well as student 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 week and survey route 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 sec. II.B). 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

sec. II.A, to assess the impact of API Plus on a large scale. To do this, we estimate the following linear regression model:

$$Y_j = \alpha_0 + \alpha_1 \text{Plus}_j + \boldsymbol{\delta}' \boldsymbol{X}_j + \epsilon_j, \quad (1)$$

where Y_j is a locality-level outcome on children's education attainment for locality j, while Plus_j takes a value of 1 if the school in locality j receives a mentor in the school year 2017–2018 and 0 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 sec. II.A). We also control for the number of hostile events related to property in land, religion, elections, crime, or drug addiction, as reported at the locality level in the 2010 population census, 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 ordinary least squares (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 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 table B6).[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 API Plus modality. Table B7 displays the results. The assignment of the mentoring program outside of the experiments is not unconditionally random (cols. 1, 3, and 5 of table B7), as priority is given to more disadvantaged communities. Instead, when we control for the vector $\boldsymbol{X}_j$, the estimated coefficients displayed in columns 2, 4, and 6 of table B7 are very small and statistically insignificant.

[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.

To curb the possibility of detecting false positives, we go beyond the conventional asymptotic inference by employing three additional procedures. First, we present *p*-values based on randomization inference, which are accurate irrespective of 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 than in the second experiment and the scale-up sample. Second, 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, Shaikh, and Xu 2019).[11] Third, building upon the insights in Maniadis, Tufano, and List (2014) and Al-Ubaydli, List, and Suskind (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 rollout). For each program modality, we calculate *p*-values for joint null hypotheses across the different study samples used throughout the analysis.[12]

B. Evidence on API Original

Tables 2 and 3 display the impacts of the API Original modality on children's schooling outcomes, collected 2 years after the introduction of the mentoring program in each experiment. 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 (cols. 1–4 of table 3), as well as a measure of educational attainment (col. 5 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

[11] The Romano-Wolf correction (Romano and Wolf 2005a, 2005b, 2016) asymptotically controls the familywise 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 resampling from the original data.

[12] To test hypotheses across study samples, we employ Fisher's combined probability test: $-2\sum_{i=1}^{k} \log(p_i) \sim \chi^2_{2k}$, where $p_i \sim U[0, 1]$ is the *p*-value for the *i*th 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.

[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 app. A1). The test was terminated in 2014, 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 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.

TABLE 2
CHILDREN'S ACHIEVEMENT—FIRST EXPERIMENT

	Reading Score	Math Score	Science Score	Overall Index
API Original	−.053	.083	−.082	−.022
	[.737]	[.655]	[.585]	[.902]
	{.750}	{.669}	{.591}	{.910}
	(.779)	(.739)	(.717)	(.878)
Schools (no.)	70	70	70	70
Observations	599	599	599	599

NOTE.—This table shows OLS estimates and the associated *p*-values on student outcomes, measured after 2 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 app. A1. The dependent variables are standardized with respect to their means and the SD 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 stepwise procedure described in Romano and Wolf (2005a, 2005b, 2016). All *p*-values account for clustering at the school level.

TABLE 3
CHILDREN'S ACHIEVEMENT AND ATTAINMENT—SECOND EXPERIMENT

	Survey-Based Test Scores				Administrative Records	
	Reading (1)	Math (2)	Socio-Emotional (3)	Overall Index (4)	Enroll Secondary (5)	(6)
API Original	.126	.056	.071	.126	.073	.081
	[.104]	[.455]	[.418]	[.182]	[.255]	[.519]
	{.134}	{.486}	{.446}	{.222}	{.281}	{.567}
	(.150)	(.558)	(.558)	(.240)	(.311)	(.478)
API Plus	.315	.237	.199	.368	.124	.298
	[.001]	[.008]	[.022]	[.001]	[.074]	[.030]
	{.001}	{.014}	{.032}	{.001}	{.089}	{.052}
	(.001)	(.005)	(.011)	(.001)	(.030)	(.030)
API Original = API Plus	[.043]	[.043]	[.178]	[.021]	[.469]	[.134]
	{.077}	{.112}	{.221}	{.024}	{.568}	{.230}
	(.045)	(.045)	(.100)	(.024)	(.372)	(.156)
Schools (no.)	224	224	224	224	182	76
Observations	1,044	1,044	1,045	1,045	468	106

NOTE.—This table shows OLS estimates and the associated *p*-values on student outcomes measured after 2 school 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 app. A2. The dependent variables in cols. 1–4 are standardized with respect to their means and the SD in the control group. The dependent variables in cols. 5 and 6 are computed from administrative school records (see app. A1). 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 administrative records) through the stepwise procedure described in Romano and Wolf (2005a, 2005b, 2016). All *p*-values account for clustering at the school level.

consistently inconclusive evidence regarding the effectiveness of the API 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 (col. 4 of table 2)—an average weighted by generalized least squares (GLS) 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 row 1 of table 3), although none of the estimated coefficients gets close to 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 SD—a nonnegligible effect size that is nonetheless not statistically different from zero (p-value $= 0.24$ after adjusting for multiple hypothesis testing). The effect of the API Original modality of the mentoring program on the transition rates to lower secondary school are shown in columns 5 and 6 of table 3. The estimated effect sizes are noisy, with an increase of 7–8 percentage points out of a basis of 62% enrollment rate in seventh grade in the control group.

The evidence is consistent 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.

C. *Evidence on API Plus*

The second row of table 3 displays the estimated coefficients for the average impact of the API Plus modality of the API program when compared to the control group. Children who are enrolled in schools that receive the API Plus mentors increase their reading scores by 0.32 SD (p-values ≤ 0.001). Quantitatively, the effect of API Plus is approximately 2.5 times larger than the effect of API Original. The difference between the two program

[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.

effects is statistically different from zero after adjusting for multiple hypothesis testing (p-value = 0.045). We find similar patterns when we look at math scores (col. 2), which show a sizable and highly significant effect of the API Plus modality, with an estimated treatment effect of 0.24 SD.

The API Plus program also generates a sizable improvement in the socio-emotional score of 0.2 SD (col. 3). While the difference with respect to the API Original modality is at the margin of statistical significance (p-value = 0.100), the larger effect of the API Plus modality is consistent with qualitative evidence documenting that mentors with enhanced training acquired more effective skills to deal with children's emotions during the bimonthly sessions.[15] The effect size of the API Plus modality on the GLS-weighted index of achievement displayed in column 4 of table 3 is very large (0.37 SD), precisely estimated (p-values ≤ 0.001), and statistically different at the 5% level from the effect of the API Original modality.[16]

Columns 5 and 6 in table 3 report the estimated effects on the average transition rate to secondary school. Less than two-thirds of the sixth graders in the control group enroll in seventh grade, while the corresponding national average is 95%. 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% increase in the share of students who transition to secondary school, relative to the mean in the control group. The size of the effect more than doubles when we focus on the subsample of overaged sixth graders (13 years old or older; col. 6). 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

[15] We conducted a series of in-depth interviews in the spring of 2022 for a small and representative subsample of 16 mentors and 12 community instructors who were part of our study. Appendix A3 reports more details about these interviews. Tables A1 and A2 show that the characteristics of these survey respondents are broadly comparable to those of the mentors and the local instructors in the second experiment.

[16] In table B8 we report the results by subdomains of the reading scores (panel A) and math scores (panel B). While the estimates are erratic and not statistically significant for the API Original modality, the API Plus modality is shown to increase students' proficiency in reading across various domains (familiar-word reading, reading comprehension, and dictation). For math scores, the API Plus modality seems particularly effective on numbers' identification and discrimination, as well as addition. Similarly, in table B9 we report the effects of the two program modalities for each individual component of the socio-emotional score.

TABLE 4
CHILDREN'S ACHIEVEMENT AND ATTAINMENT—API PLUS SCALE-UP

	NONEXPERIMENTAL SCHOOLS		EXPERIMENTAL SCHOOLS	
	Enroll Secondary (1)	Child Literacy (2)	Enroll Secondary (3)	Child Literacy (4)
API Plus	.056	.028	.091	.035
	[.010]	[.012]	[.022]	[.078]
	{.011}	{.014}	{.021}	{.075}
	(.020)	(.020)	(.041)	(.075)
Schools (no.)	1,161	1,161	184	184

NOTE.—This table shows OLS estimates and the associated robust p-values on locality-level outcomes measured after 2 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 app. A1. 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 subsamples of schools (nonexperimental and experimental) through the stepwise procedure described in Romano and Wolf (2005a, 2005b, 2016). All p-values allow for heteroskedasticity of unknown form.

then a quarter of the children aged 12–14 in Chiapas are out of school. Column 1 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% 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 (+9.1 percentage points; p-value = 0.041; see col. 3 in table 4). These effects of the mentoring intervention during the government's program implementation are statistically similar across the two school subsamples, and they are in line with the experimental findings of the API Plus modality on the enrollment in seventh grade documented in column 5 of table 3.[17]

The estimates of the impact of the government-run API Plus mentoring modality on child literacy are displayed in columns 2 and 4 of table 4. This is another relevant education outcome for the disadvantaged communities that are targeted by the intervention, akin to the achievement test scores reported in tables 2 and 3. In our sample, 13% of school-aged children are still illiterate. After 2 years of exposure, we find that villages that received mentors display a 2.8 percentage point (p-value = 0.020) increase in child literacy rates when compared to villages without mentors.

[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 school.

The magnitude of this effect implies a reduction of illiteracy rates by 21% with respect to the sample average. The estimated program effect for the subsample of experimental schools is quantitatively similar, although a bit noisier (+3.5 percentage points; p-value = 0.075).

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 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 2-year intervention, enhancing its potential for scalability. Leveraging the initial randomization from the second experiment, figure B1 (figs. B1 to B3 are available online) shows that the effect of the API Plus program on secondary-school enrollment continues beyond the 2-year time frame of the program's cycle. Figure B2 shows that extended exposure to the program beyond the initial 2-year cycle further enhances the program's impact.

IV. 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 API Plus modality. The literature discusses various mechanisms that can cause a voltage drop in the new situation (Al-Ubaydli, List, and Suskind 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 API 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.

A. 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 between how mentors were recruited and assigned to communities within the randomized trial and how recruitment and assignment was conducted in the ongoing government intervention

(API Original). First, in the randomized trial, 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 could potentially influence the effectiveness of the mentoring service.

We begin by examining the extent to which the population of mentors was 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] Despite a limited set of common variables across these two datasets, table B10 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 affect 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 API Plus program (see app. A2 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. Figure 2*A* 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

[18] The survey data comprises responses from 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 assigned to schools and those who were awaiting a role. In the 2016 survey, for instance, the 139 mentors were either assigned to the 107 unique schools included in the survey or they were currently awaiting a role within the program.

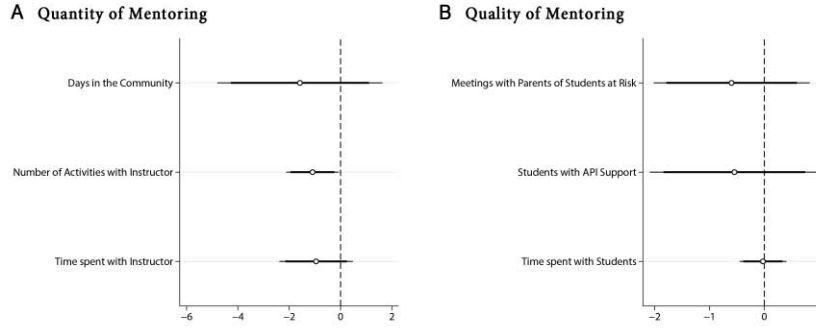

FIG. 2.—Differences in the mentoring practices between experiment and scale-up. The figure shows the comparison in the quantity and quality of API Plus mentors between the second experiment and the government implementation. This information is collected during the surveys of the local instructors, in 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 table B11. All the regressions include the same set of controls as in table 4.

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 0 to 5) 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 1 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 1 minute less during their last visit to the community. In two out of three cases we cannot reject the null hypothesis of zero effect at conventional levels of significance.

In terms of the quality of the mentoring programs, our results also show a small and statistically insignificant reduction in our observed measures between the field experiment and the government setting. The estimates of the mean differences across situations are shown in figure 2*B*. Both the number of meetings with parents of underperforming 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.

[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.

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 to close a school 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 2 school years of the study period. As a result, only two schools closed in the experimental sample of 230 schools, compared with an average 11% school closing rate in the rest of Chiapas for the 3 years before the experiment and with a 19% 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 API 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 is recorded as permanently closed in the administrative school census during the fall of 2019.

Figure 3 shows that the government implementation of the API Plus modality induces a significant and large effect, in percent terms, on school closures.[20] Both experimental schools and nonexperimental 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

[20] For the overall sample of schools in the scale-up analysis (both experimental and nonexperimental schools), only 1.6% of the schools with API Plus mentors are permanently closed after 2 years of exposure to the government-run program, against 9.1% of closures among schools without mentors.

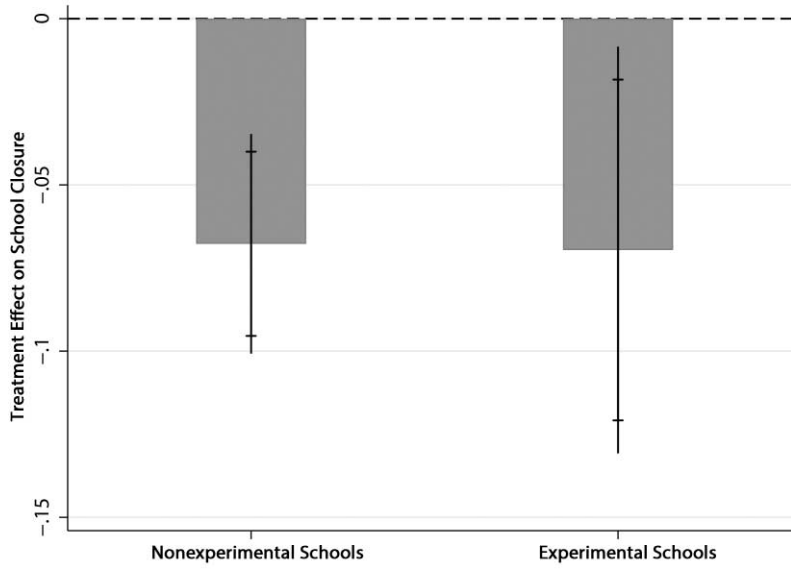

FIG. 3.—The impact of the API Plus program on school closures. The bars in the figure represents the OLS estimates of the assignment to the API program during the government implementation of the Plus modality (same as in equation (1)) on the rate of school closures, as measured over the subsequent 2 years. Vertical lines overlaid on each bar display the 95% and 90% confidence intervals, respectively. Confidence intervals are based on asymptotic inference. The OLS estimates, *p*-values, and number of observations for the two subsamples of schools are also reported in table B12.

decrease in school closures during the government implementation of 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 API 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 the program's underlying effectiveness endures during the government implementation.

B. Plus versus Original: Channels

In this subsection, we study the possible mechanisms behind the success of the API 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 app. A2).[22] Panel A displays the estimates of the API Original modality in the first experiment, while panel B shows the corresponding figures for both of the API modalities 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 SD.

Parents appear to be systematically more invested in their children's education activities under the API 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 hypothesis testing, three out of four coefficients are statistically significant at the 5% level, with a very large effect size for the overall index of parenting practices of 0.36 SD. 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 API 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 2 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 despite quite noisy estimates (due to missing observations and reduced sample

[21] As discussed in sec. II.C, 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 B4), as well as in terms of program impacts (see table 4 and fig. 3).

[22] We also estimate the impacts of both the API Original and API Plus modalities for each of the individual measures of the parental behavior collected in the survey that have been aggregated in the summary measures displayed in table 5. Table B13 reports the results, which are broadly comparable to the estimates discussed in the text. They show large and significant effects for the API Plus modality on food donations to the instructors, the management of the school resources, helping with homework, enrolling their children in extracurricular activities, expecting their children to complete secondary education or more, and meeting periodically with the instructor.

[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

TABLE 5
PARENTAL INVESTMENT

	Engage at School (1)	Manage School Resources (2)	Engage With Child (3)	Overall Index (4)
	A. First Experiment			
API Original	.198	−.135	.149	.101
	[.259]	[.415]	[.399]	[.580]
	{.261}	{.422}	{.399}	{.578}
	(.338)	(.511)	(.511)	(.511)
Schools (no.)	73	73	73	73
Observations	208	208	208	208
	B. Second Experiment			
API Original	−.188	−.124	.167	−.034
	[.049]	[.176]	[.015]	[.684]
	{.070}	{.216}	{.030}	{.709}
	(.058)	(.197)	(.015)	(.630)
API Plus	.217	.087	.353	.359
	[.034]	[.344]	[.001]	[.001]
	{.047}	{.393}	{.001}	{.001}
	(.037)	(.247)	(.000)	(.001)
API Original = API Plus	[.001]	[.056]	[.029]	[.001]
	{.000}	{.058}	{.171}	{.001}
	(.002)	(.036)	(.036)	(.001)
Schools (no.)	224	224	224	224
Observations	1,045	1,045	1,045	1,045

NOTE.—This table shows OLS estimates and the associated *p*-values on survey-based measures of parental behavior measured after 2 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 app. A2. 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, 2005b, 2016). All *p*-values account for clustering at the school level.

size). Over a 2-month period, mentors in the API Plus modality met 1 additional time with parents at school and 0.7 additional times at home compared to those in the API Original modality (sample means in the API Original group are 5 and 3, respectively). The GLS-weighted index shown in column 3 documents that the quantity of parent-mentor interactions increased by 0.36 SD under the API Plus modality, which is significant at the

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

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	A. PARENTS AND MENTORS INTERACTIONS (as Reported by Parents)						
	QUANTITY (Last 60 Days)			QUALITY			
	Meetings	Visits	Index	Inform About Child	Advise About Child	Index	
API Plus	1.039	.726	.362	.102	.100	.251	
	[.147]	[.125]	[.062]	[.057]	[.034]	[.040]	
	{.194}	{.171}	{.094}	{.097}	{.056}	{.070}	
	(.194)	(.194)	(.100)	(.078)	(.078)	(.078)	
Observations	482	491	504	354	353	357	
	B. PARENTING STYLES PROMOTED BY MENTORS (as Reported by Mentors)						
	EDUCATIVE STYLE			EMOTIONAL STYLE			
	Communication	Learning	Index	Share Feelings	Self-Knowledge	Manage Transitions	Index
API Plus	.178	.168	.494	.049	.030	.142	.194
	[.038]	[.077]	[.018]	[.627]	[.756]	[.123]	[.312]
	{.049}	{.092}	{.024}	{.637}	{.749}	{.118}	{.315}
	(.070)	(.070)	(.040)	(.846)	(.846)	(.295)	(.542)
Observations	107	107	107	107	107	107	107

NOTE.—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 (panel B). For a detailed description of the outcome variables used in this table, see app. A2. 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 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, 2005b, 2016).

10% level. Columns 3 and 4 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 and (ii) 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% increase in the probability of informing parents relative to the respective sample means in the API Original group (70%). The estimated coefficient for the GLS-weighted quality index is 0.25 SD, which is significant at the 5%–10% level depending on the inference procedure (col. 5).

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 (see app. A2).[24] Mentors with enhanced training are more inclined to foster attitudes that are centered on educative parenting styles, such as communicating with the child (col. 1), as well as learning activities (col. 2). The overall educative style GLS-weighted index displayed in column 3 of panel B shows a sizable and significant effect (p-value = 0.040) of the API Plus modality, with an increase of 0.49 SD in the promotion of educative parenting styles to parents during the home visits. The estimates in columns 4–7 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 and 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 this hypothesis. We report here a few quotes from mentors who have participated to the training sessions of the API Plus modality (see app. A3 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."

[24] Of a total of 126 schools that received mentors between the API Original and API 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-values = 0.514).

- "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 API Plus program compared to the API 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 sec. II.A for details on the eligibility). The estimates displayed in table B14 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 fig. B3).

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 API Plus modality on children's outcomes. table B15 reports estimates of the effect of 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 app. A2).[25] The results show erratic patterns of positive and negative signs with no statistically significant effects of either API modality.

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 API Plus program. The available evidence suggests that a greater parental involvement, which was likely triggered by enhanced parent-mentor interactions, played a central role.

C. Parents as Potential Means of Scalability

Given the results documented in sections IV.A and IV.B, we next explore the link between school closures and the engagement of parents with the school community under the randomized program assignment. The

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

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 sec. II.A). 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. Columns 1 and 2 of table 7 show the reduced-form effects of the two randomized program modalities—in both the first experiment (col. 1) and the second experiment (col. 2)—on the probability that schools close 2 years after the program intervention. The API 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.

Column 2 of table 7 shows that the API Plus modality, which substantially boosts parental engagement during the experiment, has a large and

TABLE 7
SCHOOL CLOSURES AND PARENTAL ENGAGEMENT

	OUTCOME: SCHOOL CLOSURES		
	First Experiment (1)	Second Experiment (2)	Second Experiment, IV (3)
API Original	.063	−.031	−.031
	[.225]	[.396]	[.410]
API Plus		−.083	
		[.030]	
Overall parental engagement			−.217
			[.021]
Observations	73	224	1,045
Clusters			224
F-statistic (excl. instrument)			13.833

NOTE.—Columns 1 and 2 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 2 years after the program intervention. Column 3 reports the IV 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 IV while the randomized API Original modality is included as a control variable. The variable "Overall Parental Engagement" is the same variable used in col. 4 of table 5. The asymptotic *p*-values reported in brackets allow for heteroskedasticity of unknown form in cols. 1 and 2, while they are clustered at the school level in col. 3.

significant impact on school closures 2 years after the API 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 API Plus modality, which is shown in figure 3.

The IV estimates shown in column 3 of table 7 go a step further and quantify the causal effect of parental engagement on the probability of school closures. An increase of 0.1 SD 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 API Plus modality 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 sec. IV.B 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 API Original mentoring program in both independent field experiments on parental investments and school closures (see tables 5 and 7) serves as further corroboration 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 result common to past successful interventions (Heckman and Mosso 2014; Zhou et al. 2021; García and Heckman 2023)—but also 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.

[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."

V. 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 vs. government implementation), as well as across study samples (experimental schools vs. 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.

Data Availability

Code replicating the tables and figures in this article can be found in Agostinelli et al. (2024) in the Harvard Dataverse, https://doi.org/10.7910/DVN/TOTKSS.

References

Agostinelli, F. 2018. "Investing in Children's Skills: An Equilibrium Analysis of Social Interactions and Parental Investments." Manuscript, Dept. Econ., Univ. Pennsylvania.

Agostinelli, F., C. Avitabile, and M. Bobba. 2024. "Replication Data for: 'Enhancing Human Capital in Children: A Case Study on Scaling.'" Harvard Dataverse, https://doi.org/10.7910/DVN/TOTKSS.

Agostinelli, F., M. Doepke, G. Sorrenti, and F. Zilibotti. 2020. "It Takes a Village: The Economics of Parenting with Neighborhood and Peer Effects." Working Paper no. 27050 (April), NBER, Cambridge, MA.

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." *Internat. Econ. Rev.* 61 (4): 1387–409.

Allcott, H. 2015. "Site Selection Bias in Program Evaluation." *Q.J.E.* 130 (3): 1117–65.

Araujo, M. C., and K. Macours. 2021. "Education, Income and Mobility: Experimental Impacts of Childhood Exposure to Progresa after 20 Years." PSE Working Paper no. 2021-57 (October), Paris School of Economics, Paris.

Attanasio, O., H. Baker-Henningham, R. Bernal, C. Meghir, D. Pineda, and M. Rubio-Codina. 2022. "Early Stimulation and Nutrition: The Impacts of a Scalable Intervention." *J. European Econ. Assoc.* 20 (4): 1395–432.

Attanasio, O., S. Cattan, and C. Meghir. 2022. "Early Childhood Development, Human Capital, and Poverty." *Ann. Rev. Econ.* 14:853–92.

August, G., M. Bloomquist, S. Lee, G. Realmuto, and J. Hektner. 2006. "Can Evidence-Based Prevention Programs Be Sustained in Community Practice Settings? The Early Risers' Advanced-Stage Effectiveness Trial." *Prevention Sci.* 7: 151–65.

Banerjee, A. V., R. Banerji, J. Berry, E. Duflo, H. Kannan, S. Mukerji, M. Shotland, and M. Walton. 2017. "From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application." *J. Econ. Perspectives* 31 (4): 73–102.

Becker, G. S. 1962. "Investment in Human Capital: A Theoretical Analysis." *J.P.E.* 70 (5): 9–49.

Bobba, M., V. Frisancho, and M. Pariguana. 2023. "Perceived Ability and School Choices: Experimental Evidence and Scale-up Effects." IZA Discussion Paper no. 16168 (May), Institute of Labor Economics, Bonn.

Bobba, M., and J. Gignoux. 2019. "Neighborhood Effects in Integrated Social Policies." *World Bank Econ. Rev.* 33 (1): 116–39.

Bold, T., M. Kimenyi, G. Mwabu, A. Ng'ang'a, and J. Sandefur. 2018. "Experimental Evidence on Scaling Up Education Reforms in Kenya." *J. Public Econ.* 168: 1–20.

Cameron, L., S. Olivia, and M. Shah. 2019. "Scaling up Sanitation: Evidence from an RCT in Indonesia." *J. Development Econ.* 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, 160–84. New York: Routledge.

CONEVAL (Consejo Nacional de Evaluación de la Política de Desarrollo Social). 2018. "Medición de la Pobreza serie 2008–2018." Technical report. https://www.coneval.org.mx/medicion/mp/documents/pobreza_18/pobreza_2018_coneval.pdf.

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, 143–60. New York: Routledge.

Doepke, M., and F. Zilibotti. 2017. "Parenting with Style: Altruism and Paternalism in Intergenerational Preference Transmission." *Econometrica* 85:1331–71.

Duflo, A., J. Kiessel, and A. M. Lucas. 2024. "Experimental Evidence on Four Policies to Increase Learning at Scale." *Econ. J.* 134:1985–2008.

García, J. L., and J. J. Heckman. 2023. "Parenting Promotes Social Mobility within and across Generations." *Ann. Rev. Econ.* 15 (1): 349–88.

Heckman, J. 1992. "Randomization and Social Policy Evaluation." In *Evaluating Welfare and Training Programs*. Edited by C. F. Manski and I. Garfinkel. Cambridge, MA: Harvard Univ. Press.

Heckman, J., and J. Zhou. 2021. "Interactions as Investments: The Microdynamics and Measurement of Early Childhood Learning." Working paper, Cent. Econ. Human Development, Univ. Chicago.

Heckman, J. J., and S. Mosso. 2014. "The Economics of Human Development and Social Mobility." *Ann. Rev. Econ.* 6:689–733.

INEGI (Instituto Nacional de Estadística y Geografía). 2020. "Censo de Población y Vivienda 2020: Diseño de la Muestra Censal." Technical report. https://www.inegi.org.mx/contenidos/productos/prod_serv/contenidos/espanol/bvinegi/productos/nueva_estruc/702825197629.pdf

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

List, J. A. 2024. "Optimally Generate Policy-Based Evidence before Scaling." *Nature* 626 (7999): 491–99.

List, J. A., F. Momeni, M. Vlassopoulos, and Y. Zenou. 2023. "Neighborhood Spillover Effects of Early Childhood Interventions." CEPR Discussion Paper no. 18134 (May), Centre for Economic Policy Research, Paris and London.

List, J. A., A. M. Shaikh, and Y. Xu. 2019. "Multiple Hypothesis Testing in Experimental Economics." *Experimental Econ.* 22 (4): 773–93.

Maniadis, Z., F. Tufano, and J. A. List. 2014. "One Swallow Doesn't Make a Summer: New Evidence on Anchoring Effects." *A.E.R.* 104 (1): 277–90.

Miguel, E., and M. Kremer. 2004. "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." *J. Econ. Perspectives* 31 (4): 103–24.

Muralidharan, K., and A. Singh. 2020. "Improving Public Sector Management at Scale? Experimental Evidence on School Governance India." Working Paper no. 28129 (November), NBER, Cambridge, MA.

O'Brien, P. C. 1984. "Procedures for Comparing Samples with Multiple Endpoints." *Biometrics* 40 (4): 1079–87.

Romano, J. P., and M. Wolf. 2005a. "Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing." *J. American Statis. Assoc.* 100:94–108.

Romano, J. P., and M. Wolf. 2005b. "Stepwise Multiple Testing as Formalized Data Snooping." *Econometrica* 73 (4): 1237–82.

Romano, J. P., and M. Wolf. 2016. "Efficient Computation of Adjusted *p*-Values for Resampling-Based Stepdown Multiple Testing." *Statis. and Probability Letters* 113:38–40.

Zhou, J., A. Baulos, J. J. Heckman, and B. Liu. 2021. "The Economics of Investing in Early Childhood." In *The Scale-up Effect in Early Childhood and Public Policy*. Edited by John A. List, Dana Suskind, and Lauren H. Supplee, 76–97. London: Routledge.