When Effective teacher training falls short in the classroom: Evidence from an experiment in primary schools

Suzanne Bellue, Adrien Bouguen, Marc Gurgand, Valerie Munier, André Tricot

When Effective Teacher Training Falls Short in the Classroom: Evidence from an Experiment in Primary Schools*

Suzanne Bellue† Adrien Bouguen‡ Marc Gurgand§
Valerie Munier¶ André Tricot‖

October 14, 2024

Abstract

Although in-service teacher training programs are designed to enhance the performance of several cohorts of students, there is little evidence on the persistence of their effects. We present the two-year results of a randomized study of an intensive in-service teacher training program conducted in France during and after the training program's implementation. Our results highlight the short-run effectiveness of the training program: it successfully improves students' performance but only during the implementation year. A detailed analysis of teachers' outcomes indicates that teachers changed their pedagogical vision and practices but afterwards struggled to apply skills to contents not directly covered during training.

JEL classification: I20

Keywords: in-service teacher training, professional development, teacher effect

*We are very thankful to the Foundation *La Main à la Pâte*, in particular David Jasmin and Elena Pasquinelli, as well as to the *Maisons pour la science* and their teams. We also thank the teachers who entered the research project. We thank the numerous field enumerators who made this project possible and the valued contribution of our Research Assistant, Todor Tochev. The project received financial support from ANR (ANR-13-APPR-0004-01). The experiment received the Paris School of Economics IRB approval IN/2005009, and was registered on the AEA registry (AEARCTR-0001864).

†CREST-ENSAE, France. Suzanne Bellue gratefully acknowledges financial support from the German Academic Exchange Service (DAAD) and the DFG, German Research Foundation through CRC-TR-224 (project A03).

‡Santa Clara University, CA, USA. Adrien Bouguen also acknowledges financial support from the German Research Foundation (DFG) project SFB 884 during his stay at the University of Mannheim.

§Paris School of Economics, CNRS, ENS-PSL, France

¶LIRDEF, Université de Montpellier et Université Paul Valéry Montpellier 3, France

‖Université Paul Valery Montpellier 3, France

1 Introduction

In-service teacher training programs are designed to improve teachers' pedagogical skills with the ultimate goal of enhancing the performance of several cohorts of students. Since training a single teacher can impact the academic outcomes of many students over several years, effective training programs can be regarded as one of the most cost-effective educational policy tools. However, the long-term success of these programs relies on teachers' ability to routinely apply the acquired teaching skills in their regular classes. While short-term evaluations provide valuable insights into the immediate effectiveness of the skills taught during the training, it is crucial to also assess whether the training program's effect persists over time.

In this article, we assess the effectiveness of an intensive in-service teacher training program in inquiry-based learning during and after its implementation. The training program targets French school teachers of Grades 3 to 5, who generally teach all subjects to a single class of students. French primary school teachers follow a national curriculum that specifies, among other things, the eight science topics to be covered and recommends the use of the inquiry-based learning method, which the United States National Research Council has long endorsed as one of the best approaches for teaching science (Council et al., 2000). Inquiry-based learning involves encouraging students to actively conduct and design experiments, define research questions, identify scientific problems, formulate hypotheses, and find solutions. It aims to enhance student engagement, motivation, and scientific skills and knowledge. These principles inspire the training program that involves designing teaching sequences based on two or three of the eight science topics. These sequences are easily reusable in the classroom, sometimes with the help of a trainer. The training program comprises 80 hours over two years, making it intensive compared to the amount of training routinely provided to French teachers (9.5 hours per year in the regions where the study was conducted according to our survey).The training program is also policy-relevant as it is implemented at scale. It is independently administered by nine local training centers, known as *Maisons pour la Science*, located in three regional school districts.[1]

We conduct a randomized evaluation in which we allocate 134 teachers who registered for the training program into a treatment and a control group. Only teachers in the treatment group benefit from the two-year training program. We collect information for all registered teachers and their respective students over two academic

[1]Metropolitan France is composed of twenty-five regional school districts.

years: during the second (and final) year of the in-service teacher training program and one year afterward, once teachers no longer receive support from trainers. We assess the scientific skills, knowledge, and motivation of two successive cohorts of students at the beginning and end of both academic years. Each cohort consists of about 2,500 students. In addition, we gather detailed information on teachers' pedagogical skills and practices in science, including the specific topics they covered in the classroom—a well-defined and easily measurable aspect of their practices. We can cross-reference the topics covered in the classroom with those covered in each training center within each regional school district. Importantly, since the control teachers are all within the catchment area of a training center, we can identify which topics they would have been exposed to if they had been assigned to the treatment group. The variation in topics covered in training sessions enables us to assess the impact of those sessions on teaching practices. Specifically, we can evaluate whether covering a specific topic during the training increases both the likelihood a teacher teaches that topic in the classroom and how they approach it. Furthermore, we can observe whether teaching practices evolve for topics not covered during the training sessions. This design, which includes information at the training center, teacher, and student levels over two years, helps interpret the mechanisms underlying the potential effectiveness of our program. Specifically, the two-year evaluation period allows us to capture both the immediate and medium-term effects of the training program, including when trained teachers independently implement the sequence in their classroom—a feature often missing in the literature.

While we encountered inherent difficulties associated with conducting and evaluating a teacher training program at scale, we observed high program adherence and low levels of differential attrition. Treatment teachers also expressed high satisfaction with the training program. They almost perfectly adhered to it, receiving an average of 66 hours of in-service training over the two years, which is reasonably close to the objective of delivering 80 hours of training. The experimental setting generates a 61-hour net increase in in-service teacher training compared to the control group. Because Grade 3 to 5 teachers can switch levels and teach Grade 1 or 2, teacher attrition increases by seven percentage points (or ten teachers) between the two evaluation years. However, our experiment does not suffer from statistically significant differential attrition, and our samples of students in both treatment arms are statistically identical in both evaluation years. We also noticed a distinct pattern in teacher enrollment. By comparing

registered teachers to their non-registered peers working in the same schools, we find that registered teachers are more likely to hold a scientific degree, are older and more experienced, and teach more hours of science. Since our experimental context closely mimics a policy implemented at scale, these findings highlight the potential challenges policymakers face when aiming to target intensive training programs at teachers with low levels of science knowledge.

Our first main results highlight the short-run effectiveness of the training program in improving students' scientific knowledge. At the end of the first evaluation year, while students' motivation and scientific skills are similar across the two treatment arms, students in the treatment group demonstrate scientific knowledge about 10% of a standard deviation (SD) higher than those in the control group. However, this positive effect vanishes by the end of the second evaluation year, when teachers no longer receive support from trainers. At this point, while the average scientific knowledge of the second cohort of students is statistically identical across the control and the treatment groups, students' scientific motivation in the treatment group is lower by about 10% of a SD than in the control group.

We then investigate teachers' practices to understand the short-run effectiveness of the program. Estimates on teacher outcomes suggest that our training program shifts teaching practices toward inquiry-based learning guidelines. In both evaluation years, treatment teachers tend to teach more science per week (+0.17 and +0.22 hours, barely significant), and their inquiry-based declared practices and knowledge indices tend to outperform those of control teachers. While the effects on pedagogical skills and knowledge indices are insignificantly different from zero, they are reasonably large, particularly one year after the end of the training program (+0.24 and +0.27 SD). The fact that all teacher practice metrics are consistently positive and become more pronounced after the program ends suggests that the training program encourages teachers to follow inquiry-based learning guidelines even after its conclusion.

Importantly, however, we observe two main differences in teachers' practices between the two evaluation years. While treatment teachers concentrate their teaching on topics covered during the training sessions during the training year, one year later, they revert to teaching a broader range of topics, including those not covered in the training. In addition, the program's positive effect on the frequency of hands-on experiments—a core activity of the program in which students actively participate—disappears in the second evaluation year. We can show that the initial aggregate increase in the

hands-on experiment is entirely driven by a change in teaching practices for the specific topics covered during the training sessions. Since hands-on experiments need to be tailored to each science topic, this suggests that, without ongoing support from the trainers, treatment teachers struggle to design new experiments that involve student participation. Furthermore, the negative effect of our program on students' motivation in the second evaluation year could indicate the challenges teachers face when independently implementing inquiry-based pedagogy to a different set of science topics. This interpretation aligns with the literature findings that the effectiveness of inquiry-based learning is highly sensitive to the quality of its implementation and the guidance provided to students (Kirschner et al., 2006; Crawford, 2007; Lazonder and Harmsen, 2016). Overall, these results call for teacher training programs that offer longer-term support from trainers and comprehensive curriculum coverage, equipping teachers with a broad range of tools they can use in their regular classes.

Our results relate to three strands of the literature and provide new insights into the effectiveness of an intensive in-service training program in both the short and longer term. Our study first contributes to the growing literature that rigorously evaluates the effectiveness of in-service teacher training programs. This literature covers a wide variety of training programs that differ in contexts, subject matters, grade levels, frequency, intensity, and impacts. Historically, this body of research has only rarely demonstrated large treatment effects on student achievement. Recent large experiments (Garet et al., 2016; Pianta et al., 2017; Loyalka et al., 2019; Drummond et al., 2020) and a recent meta-analysis on *general* professional development (Fryer, 2017) confirm this trend. However, when focusing on *managed* professional development program only—training programs that include tutoring, new curriculum materials, and extended implementation period— Fryer (2017) report significant positive impacts in his meta-analysis (+0.05 SD for math and 0.4 SD in reading). Similarly, a meta-analysis conducted by Kraft et al. (2018) reports a substantial treatment coefficient (+0.18 SD) for professional developments that include *teacher coaching*, broadly defined as all in-service programs where coaches or peers observe teachers' instruction and provide feedback. Additionally, when restricted to STEM, Lynch et al. (2019) find a large meta-coefficient of +0.21 SD, with larger impacts when the program improved teacher's content knowledge, has a pedagogical focus and includes teacher coaching. Our paper's results are consistent with the conclusions of these recent reviews. Our intervention program includes 80 hours of training, largely focuses on designing teaching sequences that could

be easily re-used in class, and includes workshops, teacher meetings, and in-class support from a teacher trainer. We find positive impacts on student performance during the program implementation period and suggestive evidence of positive impacts on teachers' knowledge and practices. However, our short-term results at the student level are smaller than those reported by Lynch et al. (2019) and Kraft et al. (2018). Similarly, the impacts on teacher knowledge and practices (about +0.25 SD in our case) are at the lower end of the range reported by the most recent meta-analysis on the effect of professional development on teacher practices (Gonzalez et al., 2022).

Secondly, our study evaluates the effects of the teacher training program one school year after its end, assessing whether it achieves its intended goals of enhancing the academic outcomes of several cohorts of students—a key feature when considering the program's overall cost-effectiveness. Very few studies have investigated the effectiveness of teacher training programs after the intervention period ends. In their meta-analysis of 60 teacher coaching studies, Kraft et al. (2018) notice that only five reported teachers' and/or students' outcomes from a follow-up year after coaching had ended, and the long-term evidence is described as "very mixed". Some exceptions include Borman et al. (2007), Allen et al. (2011), and Sirinides et al. (2018).[2] While we find that the training program's effect on students disappears after its end, Borman et al. (2007) and Sirinides et al. (2018) report a positive effect on students' achievement during and after the training program. One key distinction between these studies and ours lies in the ongoing support provided to teachers during the evaluation period. For instance, the Success for All program offers approximately 15 days of additional training each year after the initial training year (Borman et al., 2007), and Reading Recovery teachers continue to receive coaching and participate in feedback sessions on teaching practices periodically (Sirinides et al., 2018). In contrast, Allen et al. (2011) only find significant effects after the program ends. The authors evaluate a coaching intervention that includes workshop-based training and a year of personalized coaching with regular feedback about class sessions. In our case, teachers are left without ongoing support after completing the training program. Although the program incorporates meetings for teachers to discuss challenges faced during class sessions and in-class help, there

[2]Newman et al. (2012) could only capture the results after one year out of the two years of intervention because the randomization was compromised in the second year. Garet et al. (2008); Meyers et al. (2016) also evaluate their program over two years. However, Garet et al. (2008)'s evaluated program is ineffective even during implementation. In the case of Meyers et al. (2016), the randomization is done at the school level, and the authors note that it is possible that trained teachers taught surveyed students in lower grades, potentially exposing students for several years.

is no personalized feedback on their self-designed teaching sessions. Despite the program's initial success in improving students' performance, teachers likely encountered difficulties implementing inquiry-based learning for different topics without continued support from trainers. Our findings, therefore, underscore the need for professional development programs that offer ongoing support and relevant feedback, akin to those in the coaching literature analyzed by Kraft et al. (2018).

Finally, our experiment contributes to the literature that investigates the impact of training teachers on the inquiry-based learning method. Since Bruner (1961), the inquiry-based approach has gained popularity and has been extensively discussed in the theoretical literature. However, empirical evidence, particularly studies reporting mid- or long-term impacts, is relatively scarce. Newman et al. (2012); Harris et al. (2015); Meyers et al. (2016); Nugent et al. (2018) are a few examples. Except Newman et al. (2012), these studies, which typically offer intensive training programs and sometimes a teacher coaching component, report positive effects on students' achievement. However, none of them include measurements after the end of the training. Our study provides additional evidence that an intensive inquiry-based program can effectively influence teacher practices and bolster students' scientific knowledge, even among lower grade levels (3-5). In addition, it raises questions about the effectiveness of independently implementing inquiry-based learning guidelines by teachers.

In the rest of the article, we describe the teacher training program conducted by the *Maisons pour la Science* in Section 2. Section 3 then presents the experimental setting, data, and compliance. Section 4 gives the evaluation results for students, and section 5 investigates teachers' outcomes and interprets our findings.

2 Training Program: Background and Content

This paper studies a teacher training program delivered by local training centers, the *Maisons pour la Science*, referred to as *Maisons* in the rest of the paper. The objective of the program is to train teachers in the inquiry-based learning method with the ultimate goal of improving their students' scientific skills, knowledge, and motivation. We begin by providing an overview of the program, its origin, and the French primary school context.

2.1 Background of the Training Program

The teacher training program was initiated by the *Grand emprunt*—a 57 billion euro loan contracted by the French Government to stimulate the economy in the aftermath of the 2008 financial crisis. The objective of the loan was to finance innovative projects in strategic domains such as scientific knowledge, innovation, and education. The foundation *La Main à la Pâte*, an influential and experienced actor in the field of scientific awareness at school, received a grant to support a vast project aiming at improving the scientific knowledge and motivation of French students. It established local training centers, the *Maisons*, within local universities of several regional school districts. The *Maisons* then designed and implemented the training program for primary school teachers teaching in Grades 3 to 5.

2.2 French Primary Schools

In France, primary school covers Grades 1 to 5—children from ages six to eleven. Primary school teachers' initial education typically includes one undergrad degree and one teacher's initial certification obtained in national certification/training centers.[3] To join a certification center, most teachers have to pass a competitive national exam. The initial certification typically lasts one year, during which the new teachers take theoretical lectures and conduct in-class teaching sequences. Teachers' initial and in-service training in science varies depending on the age of the teacher, her initial education (scientific or not), and her pedagogical training.

In the vast majority of schools, primary school teachers are responsible for teaching all the subjects to one class of students in the same Grade. They can teach any Grade between 1 and 5. French teachers follow the national curriculum that specifies the teaching time, the topics to be covered, and the competencies that the pupils should master at the end of each school year. For instance, Grade 3 to 5 teachers should devote two hours per week to science and technology and cover eight science topics.[4] While the French curriculum recommends the inquiry-based learning method and scientific experiments, teachers are free to use the most suitable pedagogical method. The *Maisons*' training program aims at providing primary school teachers with inquiry-based learning teaching skills for science and technology.

[3]These centers are also responsible for in-service training. The training we analyzed in this paper was mostly conducted by trainers from these national training centers.

[4]Note that we are here referring to the curriculum in 2014 as it has been modified since then.

2.3 The *Maisons*' Training Program

Inquiry-based learning in science is considered one of the best pedagogical approaches to teaching science that the United States National Research Council has long endorsed (Council et al., 2000). It involves encouraging students' active role in conducting and designing experiments, defining research questions, scientific problems, and hypotheses, and finding solutions to enhance students' engagement in class, motivation in the discipline, and, eventually, scientific skills and knowledge. The training program aims to help teachers implement in-class hands-on experiments by supplying designs of experiments and materials for science topics included in the national curriculum. The program lasts two consecutive school years and comprises 80 hours of training. It includes training sessions at the *Maisons*, engaged discussions between trained teachers, attendance at scientific conferences, and in-class educational support.[5]

The training sessions at the training centers occurred during regular teaching hours. The school district managers appointed a substitute teacher, ensuring the program did not reduce normal teaching hours. Substitute teachers have the same qualifications as regular teachers, and 40 hours of absence yearly represents less than 4% of the overall yearly teaching time. We, therefore, do not believe that the substitution had any adverse effect on students' performance.

3 Experimental Setting, Data and Compliance

3.1 Survey Protocol and Teacher Selection

We evaluate the teacher training program in three regional school districts: Auvergne, Lorraine, and Midi-Pyrénées.[6] Any primary school teachers in one of those school districts could register for the training program. A total of 134 teachers registered

[5]In a companion paper that specifically focuses on the qualitative analysis and relies in part on videos taken during the training sessions and in-class with some of the volunteer teachers, Munier et al. (2021) precisely describe the different stages of the training program in one of the *Maisons*. We summarize the information about the intervention in Appendix Table A1. The training was conducted by three contributors: professional trainers from the national certification and training centers, field trainers who observed and assisted the teachers during the science sequences in her classrooms, and scientists who gave lectures on a specific curriculum topic. The training program covered four topics in this *Maison*; most hours were conducted in person at the training center, and additional hours were conducted in class. During these in-class sessions, the teachers implemented the science sequence designed at the training centers with the support of a field trainer.

[6]A fourth region was initially included but dropped out due to a lack of teacher enrollment and difficulty in finding substitute teachers.

in 2014. They work in 111 schools across three school districts and nine local school districts. In this paper, we follow those teachers for three years. Each *registered* teacher first fills out a registration form in 2014 called "Q0". When registering, we provided teachers with all the information about the intervention settings and about the research design, i.e., the limited number of available slots, the randomization, and the fact that the performance of their students will be assessed during two years. The registered teachers and their respective students answer follow-up surveys at the beginning and at the end of two academic years: the second and last year of training (year 2) and one year after the training program ends (year 3). Figure 1 depicts the survey protocol.[7] Note that because teachers usually remain in the same Grade level between two consecutive school years, they teach different students in years 2 and 3.[8] Consequently, our data set comprises a panel of teachers and a repeated cross-section of students.

[7]Note that due to implementation difficulties, the teachers in one local school district started their training one year later in 2015/2016 and were surveyed in 2016/2017 and 2017/2018.

[8]In a few cases, the teacher "followed" her students, i.e., she moved to the upper-level grade and therefore had the same students in year 2 and in year 3.

Figure 1: Survey Protocol

Experimental Year	School years	School month	Training	Teacher Survey	Student Survey
Year 0	**2013-2014**	June		Q0 survey	Q0 survey
			Random Assignment		
Year 1	**2014-2015**	Sept	**1st year of training**		
		June			
Year 2	**2015-2016**	Sept	**2nde year of training**	Q1 Survey	Q1 Survey
		June		Q2 Survey	Q2 Survey
Year 3	**2016-2017**	Sept	**No training**		Q3 Survey
		June		Q4 Survey	Q4 Survey

The figure presents the survey protocol for the main sample for the school year 2013-2014 until the school year 2016-2017. Note that an additional school district, not shown here, implemented the same protocol but one year later, i.e., the training program happened in the school year 2015-2016 and 2016-2017, and the surveys happened in 2016-2017 and 2017-2018. However, those schools filled out the Q0 survey and were randomized at the same time as the rest of the sample. Besides, note that an additional teacher survey was implemented in September of the school year 2016-2017 (Q3): since we do not rely on this survey in this paper, we do not report it here.

The second panel of Table 1 shows that registered teachers are more experienced and interested in science than their non-registered peers. This panel describes the characteristics of the registered teachers and of a group of non-registered teachers called "peer teachers". To create the group of peer teachers, we asked each registered teacher to name one of their school colleagues in Grades 3 to 5 who would be willing to participate in our survey.[9] Column *Registered v. Peer* shows that registered teachers are significantly older (+2.2 years), have more teaching experience (+3.8 years), and are more likely to hold a degree in science (+ 16 pp) than peer teachers.[10] Prior to applying for the *Maisons*' training, registered teachers are also more likely to have

[9] Peer teachers first fill out our survey in year 2. We use this information to characterize our sample of registered teachers.

[10] Compared to the national average (DEPP, 2018), registered teachers are less likely to be female (74% against 81.6%) and are older (45 years old against 42).

benefited from any in-service training in science (+3.2 hours per year) and to have benefited from training from the *Maisons* or from other *La Main à la pâte* organizations than their peers.[11]

Overall, these results suggest that this very intensive teacher training program attracted teachers who are already fairly interested in and accustomed to the topic being taught. While the *Maisons* and the school district managers were aware of this potential selection issue, their efforts to attract teachers with a lower level of scientific awareness mostly failed. More generally, we believe that this selection effect reflects an important conundrum for teacher training: targeting the program to teachers who need it the most is challenging.

Anecdotally, even though registered teachers are more exposed to in-service science training than their peer counterparts, the average number of science training hours they declared receiving per year remains relatively small (two hours per school year), especially when compared with the number of hours of training provided by our intervention (forty hours per school year). The intensity gap between our training intervention and the usual number of hours of science training among peer teachers is even larger: peer teachers declare receiving an average of one hour of training per year (conditioning on those who benefit from a science training program, the average training program is of nine and a half hours per year). The intervention, therefore, constitutes a significant increase in in-service training exposure, even for registered teachers.

Appendix Table A3 shows that while registered teachers are very specific teachers, their students are similar to the students of their peers, indicating no selection at the student level. Our cohorts of students are, therefore, likely comparable to the average students in similar schools and school districts.

3.2 Randomization

To measure the program's effectiveness, we randomly assigned the 134 registered teachers into a control and a treatment group. To do so, we used the information registered teachers provided when filling out the registration questionnaire Q0. However, because each local school district administered Q0 at different points in time (between June and September 2014), we conducted the randomization in each district separately. In

[11] The *Maisons* were already open a few years before the beginning of the experiment. The *Maisons* provided training hours in science, but the intensity was not comparable with the training program we study in this experiment. The *Main à la pâte* foundation has other interventions about science in primary schools.

most districts (five school districts), we stratified the randomization using baseline teaching experience only.[12] In three other school districts, we additionally used both experience and an index of teaching practices. Whenever several teachers from the same school registered to the program, we offered school district managers to randomize at the school level.[13] Finally, in one district, teachers did not fill out Q0 before randomization; in this case, we used the municipality as a stratification variable. Since each training center had a fixed number of available teacher training slots, each local school district had a different probability of assignment to the treatment group. In the remainder of the paper, we account for those specificities by including sampling weights proportional to the inverse of the assignment probability in regressions; we also include strata fixed effects to account for the specificity of each district randomization strategy.

Columns "Treatment v. Control" in Table 1 displays the balance checks at the teacher level. Fourteen out of fifteen relevant teachers' characteristics are statistically identical between the treatment and control groups. Using the multi-hypothesis testing step-up method developed by Benjamini and Hochberg (1995), we find that none of the coefficients is significant at the 10% level.[14] Nevertheless, we observe that the treatment group had declared teaching science slightly more than the control group one year before the intervention started. Since this variable is one of the intermediary outcomes, this unfortunate imbalance is a potential source of concern. In the remainder, as a robustness test, we will add baseline hours of science to our regressions.

Note that by design, students are not directly randomly allocated to control and treatment groups. Only teachers are. There is, therefore, a risk that post-randomization treatment teachers are assigned to better or worse students than control teachers. In

[12] We took the median of baseline teaching experience to assign teachers to an above-median experience and below-median experience strata.

[13] By randomizing at the school level, all teachers in the same school are either in the treatment or the control group, avoiding risks of contamination or spillover. However, a few school district managers refused this strategy. In seven schools, we stratify at the school level and randomize at the teacher level, leading to a treatment and a control teacher within those schools. While those cases present a risk of spillovers, several mitigating factors alleviate the concerns of downward-biased estimates. First, only seven schools out of 111 are concerned (seven control teachers), suggesting, if any, relatively minimal impacts of spillovers. Second, the intensive nature of the program (80 hours) makes it unlikely for teachers to transfer full content. Moreover, crucial aspects of the program, such as teacher assistance, were exclusively available to treatment teachers. Finally, teachers who registered for the training program from the same school typically taught different grade levels. If one of them constructed grade-specific teaching sequences, the other teacher could not directly borrow these.

[14] The minimum value to reject at least one of the 15 outcomes tested is 59%. In other words, the minimum Q-value is 59% for these fifteen coefficients, far from standard significant levels.

Appendix Table A3, we compare average students' test scores at the beginning of each of the two evaluation years (baselines) between control and treated teachers.[15] All indices are balanced at both baselines —i.e., at the beginning of each school year— suggesting an absence of selection in students' assignments between treatment arms.

[15] As mentioned previously, because teachers generally teach the same Grade year after year, the student questionnaires are generally administered to two different cohorts of students.

Table 1: Pre-Randomization Teacher Characteristics

	Treatment v. Control			Registered v. Peer		
	Obs.	Control	(1)	Obs.	Peer	(2)
Socio-economic characteristics						
Gender, 1= female	134	0.740	-0.013	223	0.679	0.060
			(0.079)			(0.066)
Birth year	132	1969.97	-0.631	214	1971.27	-2.159*
			(1.147)			(1.148)
Higher education in years	132	2.836	0.353	215	3.157	-0.168
			(0.221)			(0.194)
Holds a scientific degree	132	0.637	-0.124	212	0.396	0.157*
			(0.086)			(0.082)
Had a career in science	132	0.146	-0.019	213	0.107	0.022
			(0.057)			(0.051)
Teaching experience	132	17.456	0.379	214	14.343	3.838***
			(1.086)			(1.192)
In-service training last year						
Received some training	132	0.287	-0.007	214	0.145	0.170**
			(0.061)			(0.067)
Total training hours	116	4.660	0.848	197	3.460	3.836
			(2.220)			(2.852)
Total training hours in science	118	2.077	1.470	199	0.996	3.229**
			(1.067)			(1.482)
Received Maisons training	132	0.203	-0.042	214	0.015	0.175***
			(0.062)			(0.046)
Received La Main à la Pâte	132	0.174	-0.071	214	0.011	0.137***
			(0.048)			(0.040)
Teaching practices last year						
# of hours of sciences	132	1.920	0.278**	189	1.234	0.820***
			(0.111)			(0.114)
# of topics covered (max 8)	132	5.098	0.267	205	4.746	0.388
			(0.262)			(0.254)
% of sessions with expe.	132	0.569	0.030	205	0.588	-0.001
			(0.035)			(0.037)
Practices inquiry-based	132	0.814	0.083	.	.	.
			(0.059)			.
Observations	134	62		402	134	

The table shows the differences between treatment and control teachers before randomization at Q0 in the first three columns. In the last three columns, the table shows the difference between the volunteer teachers (treatment and control) and the peer teachers (collected in Q1). Column *Obs.* gives the number of observations, column *Control* the average in the control group, *Peer* the average for the peer teachers, and columns *(1) and (2)* the results of the regression of the dependent variables against the treatment variable or the *registered* teacher variable. All regressions are weighed and include strata fixed effects. Standard errors are given below the regression coefficients in parentheses.

p<0.01 ***, p<0.05 ** p<0.1 *

3.3 Differential Attrition across Survey Waves

Our study suffers from a relatively high level of attrition due to an increasing number of teachers switching to teaching in Grades 1 and 2 over time. Still, there is no statistically significant differential attrition. Table 2 investigates the class, student, and teacher attrition rates. In Panel A of Table 2, class attrition rises from 8% (ten teachers) in year 1 to 15% (twenty teachers) in year 2 and is almost fully driven by teachers who switch to teaching in Grades 1 and 2. Indeed, only one teacher refused to remain in our study. All other teachers accepted that we surveyed their students. The only reason we cannot survey their students is if they are not teaching in Grades 3 to 5. The differential attrition decreases from -5.3% to -7.5% over the two years. Still, it remains insignificant, suggesting that the treatment did not significantly reduce teachers' incentives to teach in the lower Grades of primary schools. On the contrary, differential student attrition rates are close to zero and remain constant across the two years. This attrition is due to students refusing to answer or being absent on the evaluation day. Finally, because our teacher questionnaire concerns teaching practices in science for Grades 3 to 5, there are three reasons for teacher attrition displayed in Panel B of Table 2. If a teacher (i) does not teach science, (ii) does not teach in Grades 3 to 5, (iii) refuses to answer our survey. Comparing class and teacher attrition indicates that very few teachers refused to answer our survey (2% in year 2 (Q2) and 4% in year 3 (Q4)) or do not teach science (1% in year 2 (Q2) and 5% in year 3 (Q4)).[16] The bulk of the class and teacher attrition rates are driven by teachers who switch to teach in Grades 1 to 2. Treatment teachers tend to continue teaching in Grades 3 to 5 longer than control teachers, explaining the small and non-significant differential attrition rates.

Nevertheless, attrition rates at the end of the two evaluation years (Q2 and Q4) are relatively high and may distort our sample. Appendix Table A2 displays the balance checks for the relevant baseline teacher characteristics using the sample of respondents of our teacher questionnaire in Q1 (year 1), Q2 (year 2), and Q4 (year 3).[17] An unfortunate imbalance rate appears significant for the baseline variable "practices inquiry-based" in year 2. In the remainder, as a robustness test, we will also add baseline "practices inquiry-based" to our regressions.

[16] As mentioned in section 2.2, primary school teachers usually teach all subjects.

[17] Using the terminology of Ghanem et al. (2022), this corresponds to a selective attrition test that examines whether, conditional on non-attrition status, baseline observable characteristics differ between the treatment and control groups.

Table 2: Differential Attrition

	Obs.	Control	(1)	(2)
Panel A: Teacher surveys				
Teacher attrition				
...at Q0	134	0.018	-0.007	
			(0.022)	
...at Q1	134	0.013	0.038	
			(0.030)	
...at Q2	134	0.156	-0.080	
			(0.056)	
...at Q4	134	0.246	-0.024	
			(0.075)	
Panel B: Student surveys				
Class attrition				
...at Q2	134	0.106	-0.053	
			(0.046)	
...at Q4	134	0.185	-0.075	
			(0.064)	
Student attrition				
...at Q2	2,935	0.083	-0.004	-0.005
			(0.011)	(0.011)
...at Q4	2,724	0.083	0.002	0.001
			(0.012)	(0.012)
Number of clusters			134	134
Controlling for grade level			N	Y

The table provides the attrition rates at the teacher level for each survey wave and at the student level in years 2 and 3. "Column Obs." gives the number of observations, "Control" the average in the control group, "(1)" the differential attrition without controlling for Grade level, and "(2)" with grade level control. p<0.01 ***, p<0.05 ** p<0.1 *

3.4 Exposure to Training

Treatment teachers almost perfectly adhere to the training program. Table 3 presents the exposure to the training program using data from the Q1 and Q2 teacher surveys.[18] In particular, we explicitly ask teachers whether or not they had received training from the *Maisons*, and if so, for how long. Being assigned to the treatment group significantly increases the teacher's probability of being enrolled in the *Maisons* training program (+ 72 pp in the first year; +87 pp in the second year). The difference between the experimental groups in terms of hours of training is large and significant. Over the two years, our results indicate that 95% of the treatment teachers received some form of training provided by the *Maisons*. The program also significantly increases the average hours of training received: compared to the control group, treatment teachers reported 32 additional hours of *Maisons* training the first year and another 30 additional hours the second year. Overall, treatment teachers reported approximately 78 hours of training over the two years and about 66 hours offered by the *Maisons*, close to the objective of offering 80 hours.

Because one of the *Maisons* authorized a few control teachers to attend some hours of training sessions, 15.5% of control teachers report having benefited from training conducted by a *Maison*. However, control teachers only received an average of four and a half hours of the *Maisons*' training over the two years, meaning that treated control teachers only received 29 hours of training over the two years.[19] These treated control teachers, therefore, received a relatively weak treatment, incomparable with the intensity received by treatment teachers.

In addition, we look at whether enrolling into the *Maisons* training program had spillover effects on other training programs' enrollment during or after the intervention. Specifically, we could be worried that the intensive training program offered by the *Maisons* is a substitute for other training programs provided by other institutions in science or other topics. We do not find evidence of such substitution. In both year 1 and year 2, the impact on "# of hours any training" is comparable to the impact on "# of hours of training from Maisons," indicating no systematic substitution patterns.

[18] This data is based on teacher reports; they are consistent with the monitoring data collected by the *Maisons* (results not shown here). For instance, according to the *Maisons*, treatment teachers benefited from an extra 35 hours the first year (against 32 hours as declared by teachers) and 22 hours the second year (against 28 as declared by teachers) for a total differential take-up of 57 hours compared to 60 hours when using self-declared teacher measures.

[19] $4.54/0.155 = 29.3$ hours. This number is consistent with our monitoring data from the *Maisons*.

Table 3: Exposure to Training

	Obs.	Control	Impact
Year 1 & Year 2			
Received any training	132	0.444	0.542***
			(0.064)
... from *Maisons*	132	0.155	0.791***
			(0.053)
# of hours of any training	132	10.751	67.19***
			(4.894)
... from *Maisons*	132	4.539	61.01***
			(4.886)
Year 1			
Received any training	129	0.275	0.662***
			(0.066)
... from *Maisons*	129	0.142	0.721***
			(0.060)
# of hours of any training	129	6.187	36.36***
			(3.423)
... from *Maisons*	129	3.747	31.72***
			(3.460)
Year 2			
Received any training	127	0.266	0.649***
			(0.069)
... from *Maisons*	127	0.031	0.868***
			(0.042)
# of hours of any training	127	4.974	29.76***
			(3.187)
... from *Maisons*	127	0.942	28.00***
			(2.439)
Year 3			
Received any training	115	0.243	0.024
			(0.085)
... from *Maisons*	115	0.111	-0.058
			(0.057)
# of hours of any training	115	5.228	0.124
			(2.329)
... from *Maisons*	115	3.579	-1.629
			(2.254)

The table shows differences between the treatment and control groups (column *Impact*) in terms of the exposure to training programs, both overall exposure and exposure to the training program provided by the *Maisons*. Column *Obs.* gives the number of *volunteer* teachers surveyed, *Control* the average in the control group, and *Impact* the treatment coefficient. All regressions are weighted and include strata fixed effects. Standard errors are below the regression coefficients in parentheses.
$p<0.01$ ***, $p<0.05$ ** $p<0.1$ *

Likewise, we do not find any significant spillover on enrolling in other training programs one year after the end of the *Maisons* program. By year 3, teachers in both groups find themselves back in the same pre-intervention situation with about 3 hours of training per year provided by the *Maisons*, not significantly different in the treatment and control groups.

Finally, Appendix Figure A3 and A4 show that treatment teachers expressed high satisfaction with the program: 87% were somewhat or very satisfied after the first training year and 85% after the second training year. They expressed satisfaction with all aspects of the training: in-class visits (95% are satisfied), in-person training sessions (92% are satisfied), and group work (87% are satisfied).

We now turn to our main results, the estimated impact of the program on students' performance.

4 Training Program's Effects on Students' Performance

4.1 Measures of Student Outcomes

To estimate the intervention's effect on students' science achievement, we construct three students' test scores: scientific knowledge, scientific skills, and scientific motivation. We developed our tests using the expertise of developmental psychologists. Most of the questions are taken from statistically validated standardized tests documented in an extensive literature review on student science assessments (Djeriouat, 2015). We describe below each of our three indices:[20]

- *Scientific knowledge*: This index assesses students' scientific knowledge. All the questions are based on the French science curriculum for Grades 3 to 5. Following it, while a few questions are specific to a given grade, a few others are common to multiple Grade levels (i.e., when a given topic must be covered at multiple Grade levels).[21] This feature makes it possible to observe the progression of pupils over time.

- *Scientific skills*: This index assesses skills developed with inquiry-based learning, such as scientific or analogical reasoning. The questions refer to situations con-

[20] Our companion paper, Munier et al. (2021), describes our instruments in greater detail.
[21] We provide an example of a question in Appendix Figure A1.

sistent with the French curriculum and, whenever possible, are taken from the existing literature.[22]

- *Scientific motivation*: This index captures students' attitudes toward science. This instrument is primarily taken from Kind et al. (2007), which develops measures of students' attitudes toward science. This questionnaire is common to the three grade levels.

The observed correlations between test scores, over time and with student characteristics, support the validity of our student instrument. First, as shown in Appendix Table A4, our student knowledge and skills measures are highly correlated with each other ($\rho = 0.501$) but not perfectly correlated, suggesting that they measure different dimensions of scientific performance. Second, the baseline correlation between knowledge and motivation is much lower and insignificant ($\rho = 0.059$), while the one between skills and motivation is null. Yet, the motivation test significantly correlates with endline knowledge, a slight influence of initial motivation on academic performance throughout the school year. Third, the three test scores correlate over baseline and endline (ρ well above 0.5). Because our indices are standardized using the control group's *baseline* scores, the averages in *Mean* column of Table 4 directly measure control group students' progress expressed in control group standard deviation (SD) over each academic year. Our tests properly capture the natural progression of students over time. For instance, over year 2, students' average performances in the control group increased by 0.74 SD and 0.54 SD in knowledge and skills, respectively. Students' motivation tends to decline during the school year, a dynamic already described by Gillet et al. (2012); Opdenakker et al. (2012). Last, Table A4 provides a set of correlations between the different test scores and some student characteristics. Our tests are properly correlated with the Grade level, and *late students*—students who were held back at least once— perform about 0.15 SD below the rest of the class in scientific skills and knowledge.[23]

4.2 Impacts on Students' Performance

Columns (1) and (2) of Table 4 present the estimated effects of the training program on the three students' indices at the end of survey years 2 and 3. Specifically, this

[22] We provide an example of a question in Appendix Figure A2.

[23] Because at least some items of our tests are the same across grades, we expect a progression across grades.

table shows the difference between the test scores of students in treatment and control groups that were measured by γ_1 when running the following OLS regression:

$$y_{i,t} = \gamma_0 + \gamma_1 T_{i,t} + \boldsymbol{X}_{i,t}\boldsymbol{\gamma_2} + \nu_{i,t},$$

with y the outcome of interest (knowledge, skills, or motivation) of student i in the evaluation year t. T is the treatment status of the teacher of student i in year t, and $\boldsymbol{X}$ is a set of control variables at the student level.[24] In both columns of Table 4, we control for the Grade level and the strata fixed effects. In column (2), we add baseline hours of science taught, baseline practices of inquiry-based learning, and the baseline score, which increases the precision of the estimates. Column (2) contains our preferred specification. Unless otherwise indicated, we refer to column (2) estimates in the rest of the text. For about 10% of the sample, the baseline score is missing: as we have no interest in the coefficient of that variable and use it only to increase precision, we set the value to zero when it is missing and add a dummy variable that indicates missing values. This is sufficient information to increase precision, and it avoids dropping observations.[25] All OLS regressions have robust standard errors clustered at the teacher level as the randomization occurred at the teacher level.[26]. All observations are weighted by randomization probabilities, and regressions include strata fixed effects. Moreover, because we are testing several treatment parameters, we provide the q-value for the false discovery rate in brackets (Anderson, 2008), which can be interpreted as a p-value, robust to multiple hypothesis testing. Because our three test scores are standardized using the control group's *pre-test* scores, our estimates can be directly interpreted as effect sizes.

At the end of year 2, the year the teacher receives her final year of training, students in the treatment group outperform control group students in scientific knowledge. After controlling for baseline variables (column (2) of Table 4), the impact is positive (+ 0.1

[24]For any given year, we identify "treatment students" as students currently taught by a treatment teacher and "control students" as students currently taught by a control teacher.

[25]Missing values on baseline hours of science taught and baseline practices of inquiry-based learning are more problematic because these variables support the ignorability assumption of the treatment variable. Yet, only nine students in Year 2 and twelve students in Year 3 have missing values (because one teacher did not fill out the baseline questionnaire), so we set the values to zero and added a dummy to indicate they are missing. In results not shown here, we re-estimated Table 4, excluding the observations with missing values; the results were unaffected.

[26]Since some teachers worked in the same school, we could have clustered the standard errors at the school level instead. In results not shown here, clustering at the school level does not make any significant difference.

SD) and significant at 5%.[27] The result is also significant when accounting for multi-hypothesis testing but less so, with a q-value of 5.8%. Scientific skills and motivation are, however, unaffected in year 2, although both dimensions were the program's prime objective.[28]

One year after the end of the training in year 3 (same teachers but different students), the impact on knowledge vanishes, and scientific skills remain unaffected. Furthermore, students' motivation is negatively affected (-0.10 SD). This result remains significant at 5% even when controlling for multi-hypothesis testing. This somewhat unexpected result is very robust. We decompose the motivation index into three sub-indices in the Appendix Table C3.[29] All of them are negatively impacted in year 3. Furthermore, this phenomenon is observed in each of the three regional school districts (F-test 0.14, p-value 0.87 for the test of equal effect in each region).

Our experiment is relatively well-powered because the minimal detectable effect is about 0.08 SD. Still, comparing coefficients between years is more demanding in terms of statistical precision. Column (3) of Table 4 shows that the differences in estimated impact between years 2 and 3 are not significantly different from each other, especially when accounting for multi-hypothesis testing (FDR p-values are provided in square brackets). However, the differences in point estimates between the two evaluation years are quantitatively substantial. The point estimate of the program's effect on students' knowledge is close to zero in year 3 and five times lower than in year 2 (0.018 SD versus 0.097 SD). Similarly, while the program's impact on students' motivation is close to zero in year 2, it is negative and five times larger in year 3 (-0.018 SD versus -0.094 SD).

The change in class attrition rates across the two evaluation years does not drive the difference in the program's impact between years 2 and 3. Indeed, column 2 of Table 4 includes control variables for all imbalance rates, including those that occur with attrition, and the point estimate of the effect on scientific knowledge drops to a fairly precise zero in the second year (0.02 SD with a standard error of 0.048). In addition, Appendix Table B1 shows the program's impact on students' outcomes, restricting the

[27] Comparing the estimations between columns (1) and (2), we see that controlling for baseline variables at the teacher and student level only marginally affects point estimates, with lower standard errors in column (2).

[28] Given the high level of precision at the student level, our estimates are unlikely to suffer from type I error: the confidence intervals for skills and motivation closely lie around zero, and we are able to detect impacts as low as 0.1 SD (effect on knowledge in year 2).

[29] The sub-indices are described in Table C1 and balancing over those sub-indices is provided in Table C2.

sample to the 112 teachers' classes that participated in both evaluation years. Even though precision is lower due to a smaller sample of classes, results in both years are qualitatively the same, with an increase of +0.07 SD in students' scientific knowledge in the first evaluation year.

Finally, we find no heterogeneous effects of the training program by students' and teachers' baseline characteristics. Appendix Table A5 and A6 show the interaction coefficients between the treatment and different sets of baseline characteristics (student scores, student gender, initial teacher training in science, or teacher gender). They are all insignificant and close to zero. This suggests that the intervention would not have fared much better should the training program had attracted teachers with different characteristics (e.g. younger or less experienced teachers).

In the next section, we investigate how the training program affects teacher outcomes to help rationalize its decreasing effects on students' test scores.

Table 4: Impacts on Students' Scores

			Treatment effect		*Y2 vs Y3*
	Obs.	Mean	(1)	(2)	(3)
Year 2					
Endline knowledge	2,694	0.737	0.116**	0.097**	
			(0.057)	(0.041)	
			[0.15]	[0.06]	
Endline skills	2,694	0.542	0.013	0.015	
			(0.048)	(0.035)	
			[1.00]	[0.82]	
Endline motivation	2,686	-0.071	-0.036	-0.018	
			(0.040)	(0.037)	
			[0.59]	[0.82]	
Number of clusters			124	124	
Year 3					
Endline knowledge	2,489	0.514	0.029	0.018	*0.18*
			(0.061)	(0.048)	*[0.37]*
			[0.73]	[1.00]	
Endline skills	2,489	0.374	-0.030	-0.010	*0.59*
			(0.054)	(0.044)	*[0.37]*
			[0.73]	[1.00]	
Endline motivation	2,488	-0.051	-0.131***	-0.094**	*0.12*
			(0.045)	(0.038)	*[0.37]*
			[0.01]	[0.05]	
Number of clusters			114	114	
Additional controls			N	Y	

This table gives the impact of the program on students' endline test scores in year 2 (upper panel) and in year 3 (lower panel). Columns *Obs.* gives the number of students surveyed, *Mean* the average in the control group, which can be read as the progression during the year in terms of baseline standard deviations. In column *(1)*, we only control for Grade fixed effects. In column *(2)*, we add baseline scores, baseline hours of science taught, and baseline inquiry-based learning practices. All regressions include strata fixed effects and are weighted by sampling probabilities. Standard errors are clustered at the teacher level and are given below the regression coefficients in parentheses. The coefficients in square brackets are the p-values robust to multiple hypothesis testing. In column *(3)*, we provide the p-value of the statistical comparison of the coefficients across years 2 and 3 using column (2) estimates.
p<0.01 ***, p<0.05 ** p<0.1 *

5 Training Program's Effects on Teachers' Practices

5.1 Measures of Teacher Outcomes

To construct our teacher outcomes, we leverage the rich data from teacher questionnaires covering two years: the second training year (year 2, Q2) and the post-training year (year 3, Q4). We create two inquiry-based learning indices:

- *Declared Practices*: The teacher survey contains questions about the five main practices related to inquiry-based learning: Introducing a scientific problem, formulating a hypothesis, linking models and observations, framing students' vision, and evaluating students. For each dimension, through sub-items, we ask teachers if they implemented them in class. For each dimension, we test the consistency of the sub-items using Cronbach alphas. We keep the internally consistent dimensions, with a Cronbach alpha above 0.7, and aggregate them in one index.[30]

- *Normative Statement*: The teacher survey also contains questions about the perceived importance of the five main practices for teaching science.[31] We aggregate four of the five normative dimensions into a *Normative statements* index.[32]

In addition, we consider more quantitative measures of *science intensity*, namely the declared number of hours of science taught per week, the number of topics covered in class, the share of topics covered that include scientific experiments, and whether these experiments were hands-on or not.[33] We also monitor which science topics teachers choose to teach in years 1, 2, and 3.

Because we want to compare the evolution of teaching practices between the two evaluation years, in the following, we will restrict the sample to teachers who answered

[30] As a result, for the *Declared practices* index, we are left with "introducing scientific problems", "framing students' vision", and "evaluating students" (see Table B6). We also submitted a questionnaire eliciting the teacher's vision of science at the end of the third year. The treatment did not affect this questionnaire, suffered from differential attrition, and answers do not correlate with student performance, so we do not present this data here.

[31] For instance, we ask *Should inquiry-based learning include introducing a problem that should be solved: always, often, etc.*; or *Do you think helping students to separate models from reality is: very important, important, etc.*

[32] For the *Normative statements* index, we only keep the teachers' normative statements on "the importance of introducing a scientific problem", "formulating a hypothesis", "linking model and observation", and "evaluating students" (see Table B6).

[33] We consider an experiment as *hands-on* if the students were directly involved in the design and implementation of the experiment, as opposed to an experiment conducted by the teacher only.

in both survey waves, in years 2 and 3.[34] We start by analyzing the relationship between the topics covered during the training program and those covered in class.

5.2 Impacts on Topics Covered in Class

The primary school science curriculum in France can be divided into eight topics (e.g., "Earth and the Universe", "Energy" or "Technical objects"). A large part of the *Maisons*' training content consisted of designing a teaching sequence based on one (sometimes two) of these topics. For instance, one training center used medieval machinery to illustrate the operation of levers and pulleys, a sequence that belongs to the topic "Technical objects". The content of such a sequence covered during the training can be easily re-used in class, sometimes with the help and presence of a trainer. Each year, we collected information on the topics covered in each training center within each regional school district (the three regions are then divided into nine local school districts). The sample is therefore composed of 15 local district-year observations that generate variation in the topics covered during training.[35]

In our teacher surveys, we list all possible topics and ask each teacher to list the ones they covered with their students during the year so that we have information for each of the three years. Using this data, we measure how much the training sessions influenced teaching by estimating whether a topic covered during training was more likely to be taught in class subsequently (in the same or the following years).

We define W_{jpt} as a dummy variable that takes the value one if a teacher j taught topic p in class in year t. Accordingly, define $Z_{c(j)pt}$ if topic p has been covered during the training program $c(j)$ where teacher j belongs during year t. Importantly, $Z_{c(j)pt}$ is also defined for the control teachers: because we know where they teach, we also know which topics they would have been exposed to if they had been trained. We estimate the following regression separately for every year t:

$$W_{jpt} = \beta_0 + \beta_1 Z_{c(j)pt} + \beta_2 Z_{c(j)pt} \times T_j + \beta_3 T_j + \varepsilon_{jpt}$$

where T_j is a treatment group dummy. In this model, β_1 should be zero because

[34] In Table B5, we present the results for the unrestricted sample: they are qualitatively the same. In addition, Table B2 presents the program's impact on students' outcomes considering the restricted sample of teachers who answered both survey waves. Results are qualitatively unchanged.

[35] In one of the regional districts that contain three local districts, all of the sessions occurred during the first year of training, while in the others, the topic sessions were spread over the two years.

the control teachers' science sequences should be unaffected by the training topics. The parameter of interest is β_2; it is positive if trained teachers are more likely to teach the training topics, suggesting they use the training materials in their class. Finally, β_3 would be negative if treatment teachers are less likely to teach the topics not covered during the training program and positive otherwise. Finally, β_0 indicates the average likelihood that a control teacher teaches a topic not covered during the training program. We also estimate a variant of this equation using $Z_{c(j)pt-1}$ to learn whether training from earlier years remains influential.

Table 5: Effects of Training Topics on Class Topics

	Topics covered in class					
	Year 1		Year 2		Year 3	
	(1)	(2)	(3)	(4)	(5)	(6)
(Y1 training topic)	0.325***		0.080		0.164*	
×(Treatment)	(0.081)		(0.082)		(0.083)	
	[0.00]		[0.20]		[0.12]	
			0.01			
			[0.03]			
(Y2 training topic)		-0.041		0.205***		0.117
×(Treatment)		(0.095)		(0.076)		(0.102)
		[0.50]		[0.02]		[0.15]
				0.03		
				[0.03]		
Y1 training topic	-0.010		-0.020		-0.084	
	(0.066)		(0.063)		(0.058)	
Y2 training topic		0.091		-0.045		-0.078
		(0.079)		(0.051)		(0.079)
Treatment	-0.083*	0.004	-0.100**	-0.114***	-0.025	-0.004
	(0.042)	(0.040)	(0.042)	(0.040)	(0.045)	(0.042)
N teachers	93	93	96	96	96	96
N observations	744	744	768	768	768	768

This table shows the regression of a dummy for covering each of the eight possible topics in each class, each year, on a treatment dummy and dummies for that very topic being covered by the local training center. The estimated coefficients are conditional on baseline hours of science taught and baseline practices of inquiry-based learning. Each column is a different regression. The sample is restricted to teachers who answered in both survey rounds. All regressions include strata fixed effects and are weighted by sampling probabilities. Standard errors are clustered at the teacher level and are given below the regression coefficients in parentheses. In italics in columns (3) and (4), we provide the p-values of the statistical comparison of the coefficients between years 1 and 2. Below in square brackets, we provide the same p-values corrected for multi-hypothesis testing. $p<0.01$ ***, $p<0.05$ ** $p<0.1$ *

Table 5 gives the above regression results.[36] Columns (1) and (4) of Table 5 show same-year relationships, whereas the other columns verify whether training topics from another year influenced the class topics in the current year. In columns (1) and (4), the interaction terms (β_2) indicate that the probability of teaching a topic that has been covered during the training program in that same year (relative to another topic) increases by 33 and 21 percentage points in the treatment group in years 1 and 2, respectively. As expected, the training topics covered during the training do not affect the topics covered in class in the control group (the training topic coefficient is close to 0). In addition, the interaction term coefficient in column (2) indicates that training topics that will be covered in year 2 have no effect on the topics covered in class in year 1.

The relationship between topics in training and in class weakens in subsequent years. Column (3) illustrates that the training topics in year 1 have little influence on the topics covered in class in year 2 in the treatment group (+0.08). Comparing year 1 and 2 effects, we first see, in italics column (4), that the year 2 topics are significantly more covered in year 2 than in year 1, with a p-value of 0.03, this result remaining significant even after controlling for multi-hypothesis testing. Besides, the decrease in the influence of training topics in year 1 is significant, with a p-value of 0.01 (column (3) in italics), still significant after accounting for multi-hypothesis testing (in square brackets). These results confirm that the training program affected the choice of topics during the training year, but that influence faded out one year later. Columns (5) and (6) indicate that in year 3, when the training program is over, training topics covered in years 1 or 2 are covered in class more often in the treatment group, but by only 16 and 12 percentage points, respectively.

Interestingly, in years 1 and 2, the negative coefficient on the treatment dummy in columns (1) and (4) indicate that treatment teachers teach fewer topics that were not covered during the training program. In year 3 (columns (5) and (6)), this effect disappears, suggesting that treatment teachers revert to teaching topics not covered during the training program as often as control teachers. In Table 6 below, we directly regress the number of different topics covered in the year on the treatment dummy: it

[36] In this analysis, we have a maximum of 134 teachers * 8 topics = 1072 data points. However, in Table 5, we restrict the sample to teachers who filled out both survey rounds (less than 100 teachers thus less than 800 observations). Appendix Tables B3 show the results if we omit baseline covariates, and Table B4 presents the results on the entire sample. The coefficients are of similar magnitude. The constant coefficients of Appendix Table B3 indicate that slightly more than half of the topics not covered during the training program were taught in class, on average, in the control group.

is lower in the treatment group in year 2 but not in year 3 anymore. This finding is compatible with the fact that French teachers must cover the full science curriculum, which includes topics not covered in training, and make an exception only during training. Overall, our results indicate that trainers influenced teachers' topic choices during the training program, but much less so when it is over.

We now analyze the program's effects on inquiry-based learning and science teaching practices.

5.3 Impacts on Reported Teaching Practices

We estimate the effects of the program on teaching practices through the following OLS regression:

$$y_{j,t} = \alpha_0 + \alpha_1 T_j + \boldsymbol{X}_{j,t}\boldsymbol{\alpha_2} + \nu_{j,t},$$

where $y_{j,t}$ represents the outcome of interest for teacher j in year t. The assignment status is denoted by T, and $\boldsymbol{X}$ represents a set of control variables. Table 6 presents the estimated effects measured by α_1. While columns (1) and (3) only include strata fixed effects as control, we account for unfortunate imbalance rates by including the baseline number of hours in science and inquiry-based learning variables as control variables in columns (2) and (4). Given initial imbalances, columns (2) and (4) contain our preferred specification. All observations are weighted by randomization probabilities, and we present robust standard errors.

5.3.1 Inquiry-based learning

Table 6 reports positive, but often insignificant, effects of the program on the *Declared practices* and *Normative statements* indices, which capture whether teachers implement and understand inquiry-based learning guidelines.[37] Even though the effects in columns (2) and (4) are insignificant, the point estimates of the effects are reasonably large and close to acceptance standards, providing suggestive evidence of changes in teaching practices.[38] This is particularly noticeable one year after teachers have completed the entire training program, in year 3 (0.24 SD and 0.27 SD controlling for

[37] Appendix Table B6 displays the effects of the program on all the sub-indices of each index, described in Section 5.1.

[38] Note that our analysis at the teacher level is based on a maximum of 134 data points, and 96 with attrition, therefore suffering from limited statistical power.

baseline variables, column (4)).[39] These results suggest that the program effectively improved teachers' knowledge about inquiry-based learning guidelines and induced them to implement those guidelines more frequently, even one year after the program ends.

Note that these measures are self-declared and may be subject to desirability bias *or reference point bias*. A desirability bias would be at play if teachers only answered our questions based on what they heard in the training sessions and not on what they actually implemented in class, leading to an overestimation of the impacts. Nevertheless, this reasoning implies that teachers have learned inquiry-based learning guidelines, suggesting the program successfully conveyed those guidelines. On the contrary, with a reference point bias, treatment teachers' perception of the optimal inquiry-based learning practices would have changed over time, leading to an underestimation of the impacts on practices.

5.3.2 Science intensity

The training program also impacts our more quantitative measures of teachers' practices in science: the weekly hours of science instruction, the number of covered topics, and the share of topics that include hands-on experiments. In both evaluation years, treatment teachers report teaching slightly more hours of science than control teachers. The point estimates, while positive, are insignificant and not statistically different between years 2 and 3: +0.17 and +0.22 weekly hours of science, or +10 and +13 minutes per week (columns (2) and (4) Table 6).[40]

Interestingly, as mentioned above, the program's effect on the number of topics and hands-on experiments differs between years 2 and 3. While in year 2, treatment teachers cover fewer science topics in class (-0.639, significant, column (2)), they revert to teaching a similar number of topics as control teachers in year 3 (+0.125, insignificant, column (4)), this difference across survey years is significant (p-value=0.01), even after accounting for multi-hypothesis testing (p-value=0.03). In addition, we see a +10 pp increase in the fraction of topics in which teachers include hands-on experiments, but

[39] Teacher attrition is not driving the positive estimates on our teachers' practice indices. Appendix Table B5 reports the estimated effects of the program on those indices for all observed teachers. Our sample increases to a bit more than a hundred teachers, and the effect size remains substantial, especially in year 3, above 0.2 SD.

[40] Comparing columns (1) and (3) to columns (2) and (4), we see that controlling for the imbalanced baseline number of hours and inquiry-based learning variables lowers the point estimates that fall below the significance level. As for our inquiry-based metrics, we favor the specification that controls for initial imbalance.

Table 6: Impacts on Teacher Practice Indices

	Year 2 (Y2)				**Year 3 (Y3)**				
			Treatment effect				Treatment effect		Y2 vs Y3
	Obs.	Mean	(1)	(2)	Obs.	Mean	(3)	(4)	(5)
Inquiry-based learning									
Declared practices	96	-0.05	0.262 (0.208) [0.20]	0.106 (0.217) [0.34]	96	0.04	0.365* (0.215) [0.23]	0.239 (0.221) [0.62]	*0.39* *[0.65]*
Normative statements	96	-0.01	0.292 (0.187) [0.14]	0.231 (0.208) [0.22]	95	-0.01	0.318 (0.227) [0.28]	0.266 (0.249) [0.62]	*0.81* *[0.94]*
Science intensity									
Weekly hours	96	1.42	0.250** (0.100) [0.07]	0.166 (0.103) [0.21]	96	1.34	0.257** (0.113) [0.15]	0.218* (0.115) [0.44]	*0.58* *[0.78]*
Number of topics	96	4.36	-0.307 (0.347) [0.26]	-0.639* (0.339) [0.21]	96	4.19	0.183 (0.319) [0.38]	0.125 (0.344) [0.92]	*0.01* *[0.03]*
% hands-on experiments	96	0.65	0.116** (0.051) [0.07]	0.096* (0.052) [0.21]	96	0.65	0.032 (0.050) [0.38]	0.010 (0.055) [0.92]	*0.12* *[0.30]*
Additional controls			N	Y			N	Y	Y

The table gives the program's impacts on the teachers' practice indices. We restrict the sample to teachers who answered both surveys. Columns *Obs.* give the number of teachers, *Mean* the average in the control group, *(1), (3)* the treatment coefficients *(2), (4)* the treatment coefficients conditional on baseline hours of science taught and baseline practices of inquiry-based learning. All regressions are weighted by sampling probabilities and include strata fixed effects. Robust standard errors are given below the regression coefficients in parentheses. In column *(5)*, we provide the p-value of the statistical comparison of the coefficients between years 2 and 3 using columns (2) and (4).
p<0.01 ***, p<0.05 ** p<0.1 *

only in year 2.[41] Indeed, in year 3, the point estimate drops to 0.01 (columns (2) and (4) of Table 6), the difference across years being almost significant (p-value=0.12), although this result does not withhold the multi-hypothesis correction (p-value=0.3). Taken at face value, these last results suggest that, in addition to influencing the topic choice of teachers, the program successfully induced teachers to use the material developed during the training program and translate it into their own science sequences in class. However, this effect is only salient during the training program implementation and is only suggestive given the lack of statistical precision.[42]

The decreasing effects of the program on the share of topics with hands-on experiments can be driven by treatment teachers stopping implementing experiments or by treatment teachers facing difficulties in extrapolating their skills to a different set of topics than those covered during training. Table 7 provides evidence for both explanations. The positive effect on the share of topics with hands-on experiments in year 2 is primarily driven by topics covered by the training program in year 2 (+0.29, column (2)). In year 3, the effect for covered topics remains positive but smaller and insignificant (+0.10, column (4)). Treatment teachers continue to implement the experiments they have learned during the training program one year after its end, but to a lesser extent. However, the effect on topics not covered by the training program in year 2 is close to zero in both years (+0.04 and -0.03, respectively, columns (2) and (4)). Since hands-on experiments must be tailored to each science topic, this indicates that treatment teachers were unable to design hands-on experiments for topics different from those covered by the program. The change in the topic sets between the two years then contributes to explaining the overall drop in the intensity of hands-on experiments. Once again, given the small sample size, the differences across years are hardly significant, especially when accounting for multi-hypothesis testing. We, therefore, consider this interpretation as suggestive.

[41] Notice that the percentage of science topics taught with experiments conducted solely by the teacher remains unaffected (see Appendix Table B6). This result can be explained by the fact that preparing and implementing science sequences with hands-on experiments in class is an aspect of the training program.

[42] Remember that the sample size at the teacher level is small: any comparison across years is therefore suggestive.

Table 7: Impacts on Hands-On Approach

	Year 2 (Y2)				Year 3 (Y3)				
			Treatment effect				Treatment effect		Y2 vs Y3
	Obs.	Mean	(1)	(2)	Obs.	Mean	(3)	(4)	(5)
% topics with hands-on experiments									
if covered	61	0.61	0.269**	0.294**	56	0.59	0.136	0.103	*0.24*
topics			(0.122)	(0.138)			(0.123)	(0.139)	*[0.47]*
			[0.07]	[0.08]			[1.00]	[1.00]	
if not	96	0.65	0.065	0.037	96	0.66	-0.008	-0.029	*0.32*
			(0.069)	(0.070)			(0.058)	(0.062)	*[0.47]*
			[0.21]	[0.43]			[1.00]	[1.00]	
Baseline covariates			N	Y			N	Y	Y

The table gives the impacts of the program on the share of topics with hands-on experiments depending on whether the topics have been covered by the training program in year 2 or not. We restrict the sample to teachers who answered both surveys. Columns *Obs.* give the number of teachers, *Mean* the average in the control group, *(1), (3)* the treatment coefficients *(2), (4)* the treatment coefficients conditional on baseline hours of science taught and baseline practices of inquiry-based learning. All regressions are weighted by sampling probabilities and include strata fixed effects. Robust standard errors are given below the regression coefficients in parentheses. In column *(5)*, we provide the p-value of the statistical comparison of the coefficients between years 2 and 3 using columns (2) and (4).
p<0.01 ***, p<0.05 ** p<0.1 *

5.4 Discussion

Overall, our findings on teachers' outcomes indicate that the training program influenced teachers' practices. During the program implementation, treatment teachers spend slightly more time teaching science. They are also more likely to understand and declare implementing inquiry-based learning guidelines (though not significantly so), teach topics covered by the program, and use hands-on experiments developed during the training for their science sequences. We also see that in year 2, teachers concentrate their efforts on a smaller number of topics, likely with more preparation and a stronger focus on hands-on science experiments. One year after the program ends, treatment teachers still declare implementing inquiry-based learning guidelines more frequently (although not significantly so) and spend slightly more time teaching science than control teachers.

Even though the point estimates of the program's impact on teachers' practices may seem large, they are small in comparison to those reported in other successful inquiry-based learning programs. Gonzalez et al. (2022)'s meta-analysis on STEM teacher training programs finds an average effect on teachers' knowledge and instruction outcomes of 0.49 SD. The authors associate a 1 SD increase in teacher knowledge and instruction outcomes with a 0.21 SD improvement in student achievement. Our relatively small effects on teachers could help rationalize the absence of impact on students' performance. Nevertheless, we find a 0.10 SD increase in students' knowledge in the first evaluation year, suggesting that changes in teachers' practice positively influenced students' learning, at least in the very short term.

The measurable differences in teaching practices between the two evaluation years reside in teachers' science topic choices and the number of hands-on experiments. Between the two evaluation years, the number of covered topics reverted to pre-treatment levels, and the share of topics taught with hands-on experiments dropped to reach baseline levels. Given the high level of reported satisfaction with the training program, the decline in the utilization of the training materials in the classroom is unlikely caused by teacher dissatisfaction with them.[43] This decline could, however, be attributed to the fact that teachers must cover the full science curriculum and may have lacked time to prepare additional teaching materials in year 3. Indeed, in year 3, treatment teachers cover more topics but do not increase the number of hours dedicated to science compared to year 2. They may have chosen to include fewer time-consuming hands-on

[43] Appendix Figures A3 and A4 display teachers' satisfaction levels.

experiments in their science lessons.

A second complementary explanation stems from the fact that not all topics were covered by the training program, and hands-on experiments must be tailored to each science topic. Our results indicate that, without the guidance of the trainers, treatment teachers faced difficulties in designing new experiments that involved students' participation and likely did not uphold the quality of the previous year's activities. This interpretation aligns with the main conclusions of the qualitative analysis of training sessions and class sequences of the same program by Munier et al. (2021): "It is as if training courses were designed solely to enhance or supplement teachers' scientific knowledge, and provide them with "turnkey" pedagogical situations, with no attempt to develop structured knowledge of classroom implementation, either didactic or general pedagogical".

The decay in the program's effect on students' scientific knowledge and its negative impact on students' motivation in the second evaluation year are consistent with the challenges teachers may have encountered when independently implementing inquiry-based learning pedagogy to a different set of science topics. Indeed, there is evidence in the literature that the effectiveness of inquiry-based learning is sensitive to the quality of its implementation and the guidance provided to students (Kirschner et al., 2006; Crawford, 2007; Lazonder and Harmsen, 2016).[44] Students may not enjoy inquiry-based teaching if it lacks sufficient guidance, which can leave them feeling "lost and frustrated" (Kirschner et al., 2006). Blazar and Pollard (2023) also observe that students' engagement is stronger when the sequences are organized around clearer routines, as opposed to when they have to take more responsibility for their learning. Our negative effect on motivation once teachers are no longer supported by the trainers may reflect these mechanisms.

While fade-out is relatively common in the education literature and does not necessarily indicate an absence of long-term impacts (Bailey et al., 2017), in our case, the fact that teachers seem unable to implement the inquiry-based activities independently jeopardizes the effectiveness and sustained impacts of the training program. Overall, our results suggest that trainers' support was key in improving children's outcomes. They call for teacher training programs that offer sustained support from trainers, akin

[44] An alternative explanation could be that in the first evaluation year, trainers' direct class input positively influenced students' knowledge. However, direct trainers' interventions were limited to a few hours, and we believe the program's effects in the first evaluation year most likely came from a change in teachers' practices.

to the assistance provided in coaching programs (Kraft et al., 2018), and comprehensive curriculum coverage, equipping teachers with a broad array of tools that can be used in their regular classes.

6 Conclusion

The primary objective of in-service teacher training programs is to improve teachers' practices and enhance the academic outcomes of several cohorts of students. Since the cost-effectiveness of training programs crucially hinges on teachers' capacity to independently implement the practices they learned during training, assessing the effectiveness of such programs requires long-term evaluations that span beyond the duration of the training itself. In this paper, we use a randomized control trial to evaluate the impact of an intensive inquiry-based learning teacher training program over two years: during and after the program implementation. Our study suffers from attrition inherent to teacher evaluation programs when implemented at scale but benefits from high adherence rates and high satisfaction rates among treatment teachers.

Our first findings reveal the effectiveness of the training program. During its implementation, students' scientific knowledge improves. The change in teachers' practices is consistent with this improvement: teachers spend more time teaching science, conduct more hands-on experiments in topics covered by the program, and implement the inquiry-based learning guidelines more often.

However, these effects are short-lived. We find that the positive effects on students' performance vanish one year after the completion of the program. Furthermore, the program negatively impacts students' motivation in this second evaluation year. We analyze teaching practices and identify two main differences across the two evaluation years. First, teachers cover science topics that weren't necessarily covered in training. Second, they include fewer hands-on experiments in their sequences. Those differences, coupled with the negative effect on students' motivation, are indicative of difficulties that teachers may have faced when implementing inquiry-based learning guidelines for a different set of science topics in the absence of ongoing support from the trainers.

Overall, our results call for sustained support from trainers or comprehensive curriculum coverage that would equip teachers with a broad range of tools applicable to their regular classroom practices. Additionally, our results underscore the importance of conducting large-scale, long-term evaluations of teacher training programs to assess

their effectiveness in enhancing the academic outcomes of not one but multiple student cohorts.

References

Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y., and Lun, J. (2011). An interaction-based approach to enhancing secondary school instruction and student achievement. *Science*, 333(6045):1034–1037.

Anderson, M. L. (2008). Multiple inference and gender differences in the effects of early intervention: A reevaluation of the abecedarian, perry preschool, and early training projects. *Journal of the American Statistical Association*, 103(484):1481–1495.

Bailey, D., Duncan, G. J., Odgers, C. L., and Yu, W. (2017). Persistence and fadeout in the impacts of child and adolescent interventions. *Journal of research on educational effectiveness*, 10(1):7–39.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Blazar, D. and Pollard, C. (2023). Challenges and tradeoffs of "good" teaching: The pursuit of multiple educational outcomes. *Journal of Teacher Education*, 74(3):229–244.

Borman, G. D., Slavin, R. E., Cheung, A. C., Chamberlain, A. M., Madden, N. A., and Chambers, B. (2007). Final reading outcomes of the national randomized field trial of success for all. *American Educational Research Journal*, 44(3):701–731.

Bruner, J. S. (1961). The act of discovery. *Harvard educational review*.

Council, N. R. et al. (2000). *Inquiry and the national science education standards: A guide for teaching and learning*. National Academies Press.

Crawford, B. A. (2007). Learning to teach science in the rough and tumble of practice. *Journal of Research in Science Teaching*, 44:613 – 642.

DEPP (2018). Bilan social du ministère de l'éducation nationale et de la jeunesse.

Djeriouat, H. (2015). Comment évaluer des connaissances scientifiques des élèves ? analyse de l'existant. validation du questionnaire d'évaluation des connaissances et attitudes vis à vis de la science. Rapports non publiés.

Drummond, K. V., Tucker-Bradway, N., Smith, D.-M., Hubbard, D., Meakin, J., and Salinger, T. (2020). Children's literacy initiative: Final report of the i3 scale-up study. *American Institutes for Research*.

Fryer, R. (2017). The production of human capital in developed countries: Evidence from 196 randomized field experiments. In *Handbook of economic field experiments*, volume 2, pages 95–322. Elsevier.

Garet, M. S., Cronen, S., Eaton, M., Kurki, A., Ludwig, M., Jones, W., Uekawa, K.,

Falk, A., Bloom, H. S., Doolittle, F., et al. (2008). The impact of two professional development interventions on early reading instruction and achievement. ncee 2008-4030. *National Center for Education Evaluation and Regional Assistance*.

Garet, M. S., Heppen, J. B., Walters, K., Parkinson, J., Smith, T. M., Song, M., Garrett, R., Yang, R., and Borman, G. D. (2016). Focusing on mathematical knowledge: The impact of content-intensive teacher professional development. ncee 2016-4010. *National Center for Education Evaluation and Regional Assistance*.

Ghanem, D., Hirshleifer, S., and Ortiz-Becerra, K. (2022). Testing attrition bias in field experiments. Working Paper 202218, University of California at Riverside, Department of Economics.

Gillet, N., Vallerand, R. J., and Lafrenière, M.-A. K. (2012). Intrinsic and extrinsic school motivation as a function of age: The mediating role of autonomy support. *Social Psychology of Education*, 15(1):77–95.

Gonzalez, K., Lynch, K., and Hill, H. C. (2022). A meta-analysis of the experimental evidence linking stem classroom interventions to teacher knowledge, classroom instruction, and student achievement. *EdWorkingPaper*, (22-515).

Harris, C. J., Penuel, W. R., D'Angelo, C. M., DeBarger, A. H., Gallagher, L. P., Kennedy, C. A., Cheng, B. H., and Krajcik, J. S. (2015). Impact of project-based curriculum materials on student learning in science: Results of a randomized controlled trial. *Journal of Research in Science Teaching*, 52(10):1362–1385.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and psychological measurement*, 20(1):141–151.

Kind, P., Jones, K., and Barmby, P. (2007). Developing attitudes towards science measures. *International journal of science education*, 29(7):871–893.

Kirschner, P., Sweller, J., and Clark, R. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, 41(2):75–86.

Kraft, M. A., Blazar, D., and Hogan, D. (2018). The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of educational research*, 88(4):547–588.

Lazonder, A. and Harmsen, R. (2016). Meta-analysis of inquiry-based learning: Effects of guidance. *Review of Educational Research*, 86(3):681–718.

Loyalka, P., Popova, A., Li, G., and Shi, Z. (2019). Does teacher training actually work? evidence from a large-scale randomized evaluation of a national teacher training program. *American Economic Journal: Applied Economics*, 11(3):128–54.

Lynch, K., Hill, H. C., Gonzalez, K. E., and Pollard, C. (2019). Strengthening the research base that informs stem instructional improvement efforts: A meta-analysis. *Educational Evaluation and Policy Analysis*, 41(3):260–293.

Meyers, C. V., Molefe, A., Brandt, W. C., Zhu, B., and Dhillon, S. (2016). Impact

results of the emints professional development validation study. *Educational Evaluation and Policy Analysis*, 38(3):455–476.

Munier, V., Bächtold, M., Cross, D., Chesnais, A., Lepareur, C., K., M., Gurgand, M., and Tricot, A. (2021). Etude didactique de l'impact d'un dispositif de formation continue à un enseignement des sciences fondé sur l'investigation. impact sur les élèves / impact sur les enseignants. *Recherches en Didactique des Sciences et des Technologies*.

Newman, D., Finney, P. B., Bell, S., Turner, H., Jaciw, A. P., Zacamy, J. L., and Gould, L. F. (2012). Evaluation of the effectiveness of the alabama math, science, and technology initiative (amsti). final report. ncee 2012-4008. *National Center for Education Evaluation and Regional Assistance*.

Nugent, G., Kunz, G., Houston, J., Wu, C., Patwardhan, I., Lee, S., DeChenne-Peters, S. E., and Luo, L. (2018). The effectiveness of a summer institute and remotely delivered science instructional coaching in middle and high school. *Journal of Science Teacher Education*, 29(8):760–784.

Opdenakker, M.-C., Maulana, R., and den Brok, P. (2012). Teacher–student interpersonal relationships and academic motivation within one school year: Developmental changes and linkage. *School Effectiveness and School Improvement*, 23(1):95–119.

Pianta, R., Hamre, B., Downer, J., Burchinal, M., Williford, A., Locasale-Crouch, J., Howes, C., La Paro, K., and Scott-Little, C. (2017). Early childhood professional development: Coaching and coursework effects on indicators of children's school readiness. *Early Education and Development*, 28(8):956–975.

Sirinides, P., Gray, A., and May, H. (2018). The impacts of reading recovery at scale: Results from the 4-year i3 external evaluation. *Educational Evaluation and Policy Analysis*, 40(3):316–335.

Appendix A: Additional Tables

Table A1: Description of a training program in a *Maison* during one year

Session topics	Lectures	In-class with field trainer	Preparation time	Trainer
Machinery	6 hrs	2 hrs	1 hr	1 ESPE trainer & 1 field trainer
Light and astronomy	12 hrs			1 scientist & 1 ESPE trainer
Technical objects	6 hrs			ESPE trainer
Wood	12 hrs	2 hrs		1 scientist & 1 field trainer
Inquiry-based method	3 hrs			1 ESPE trainer & 1 field trainer
Total	39 hrs	4 hrs	1 hr	

The table describes the activities and the corresponding number of hours of one training program conducted in one of the Maison during one year. ESPE trainers are trainers from the certification centers, and field trainers are trainers (usually students) coming to the trainee's classroom to help implement a teaching sequence about science. The information contained in this table is taken from Munier et al. (2021).

Table A2: Pre-Randomization Teacher Characteristics on Respondent Teachers in Years 1, 2 and 3

	Year 1			Year 2			Year 3		
	Obs.	Control	(1)	Obs.	Control	(2)	Obs.	Control	(3)
Socio-economic characteristics									
Gender, 1= female	129	0.737	-0.023	119	0.713	-0.003	102	0.672	0.030
			(0.081)			(0.089)			(0.107)
Birth year	128	1970.00	-0.590	119	1969.92	0.093	101	1969.77	-0.795
			(1.175)			(1.231)			(1.364)
Higher education in years	128	2.847	0.335	119	2.866	0.397	101	2.893	0.290
			(0.231)			(0.246)			(0.280)
Holds a scientific degree	128	0.645	-0.128	119	0.613	-0.077	101	0.677	-0.092
			(0.088)			(0.096)			(0.098)
Had a career in science	128	0.148	-0.026	119	0.114	-0.005	101	0.172	-0.037
			(0.058)			(0.063)			(0.066)
Teaching experience	128	17.422	0.589	119	17.474	-0.400	101	17.358	0.845
			(1.099)			(1.219)			(1.315)
In-service training in year 0									
Received some training	128	0.291	-0.017	119	0.281	-0.036	101	0.285	-0.012
			(0.061)			(0.065)			(0.067)
Total training hours	115	11.348	-1.234	108	11.239	-1.310	89	12.223	-0.316
			(6.879)			(7.871)			(10.54)
Total training hours in science	115	2.108	1.581	108	1.088	1.469	89	1.064	0.880
			(1.092)			(1.152)			(1.242)
Received Maisons training	128	0.206	-0.056	119	0.179	0.007	101	0.160	-0.007
			(0.063)			(0.062)			(0.067)
Received La Main à la Pâte	128	0.176	-0.082*	119	0.180	-0.071	101	0.140	-0.059
			(0.049)			(0.053)			(0.054)
Teaching practices in year 0									
# of hours of sciences	128	1.925	0.297***	119	1.933	0.231*	101	1.934	0.219
			(0.112)			(0.123)			(0.136)
# of topics covered (max 8)	128	5.113	0.217	119	5.145	0.187	101	5.066	0.154
			(0.263)			(0.284)			(0.342)
% of sessions with expe.	128	0.570	0.031	119	0.564	0.047	101	0.574	0.035
			(0.036)			(0.038)			(0.041)
Practices inquiry-based	128	0.825	0.069	119	0.784	0.151**	101	0.793	0.114
			(0.059)			(0.064)			(0.069)
Observations	129	61		119	53		102	46	

The table shows the differences between treatment and control teachers before randomization at Q0 on the sample of teachers who responded to the teacher questionnaire in years 1, 2 or 3. Column *Obs.* gives the number of observations, and column *Control* the average in the control group. All regressions are weighed and include strata fixed effects. Standard errors are given below the regression coefficients in parentheses.

Table A3: Balance Checks - Students' Outcomes and Characteristics

	Treatment v. Control			Volunteer v. Peer		
	Obs.	Control	(1)	Obs.	Peer	(2)
Year 2						
Baseline knowledge	2,689	-0.035	0.017	4,495	-0.064	0.049
			(0.058)			(0.044)
Baseline skills	2,689	-0.020	-0.056	4,495	-0.132	0.084
			(0.074)			(0.063)
Baseline motivation	2,672	-0.009	0.015	4,461	0.079	-0.075**
			(0.046)			(0.036)
Grade 3	3,053	0.223	0.076	5,207	0.248	0.017
			(0.064)			(0.062)
Grade 4	3,053	0.330	-0.054	5,207	0.394	-0.097*
			(0.065)			(0.056)
Grade 5	3,053	0.447	-0.022	5,207	0.358	0.080
			(0.072)			(0.068)
Late student	3,053	0.087	-0.022*	5,207	0.073	0.003
			(0.012)			(0.008)
Female student	3,053	0.474	0.023	5,207	0.498	-0.010
			(0.016)			(0.014)
Year 3						
Baseline knowledge	2,529	-0.027	-0.003	3,971	-0.039	0.005
			(0.062)			(0.062)
Baseline skills	2,529	-0.039	-0.097	3,971	-0.143	0.007
			(0.085)			(0.086)
Baseline motivation	2,516	0.012	-0.053	3,951	-0.001	-0.006
			(0.053)			(0.044)
Grade 3	2,883	0.235	0.115	4,408	0.260	0.051
			(0.072)			(0.069)
Grade 4	2,883	0.316	-0.037	4,408	0.423	-0.111
			(0.066)			(0.072)
Grade 5	2,883	0.448	-0.079	4,408	0.317	0.060
			(0.069)			(0.081)
Late student	2,826	0.083	-0.025*	4,351	0.110	-0.033*
			(0.013)			(0.019)
Female student	2,826	0.476	0.011	4,351	0.511	-0.022
			(0.016)			(0.017)

The table provides the baseline difference, in years 2 and 3, between the students in the treatment and control group and the difference between peer students and students of volunteer teachers. Column *Obs.* gives the number of students, *Control* the average in the control group, *Peer* the average of the peer students, and column (1) and (2) the difference between treatment and control and peer student and students of volunteer teachers respectively. All observations are weighted by sampling probabilities. We control for strata-fixed effects. Standard errors are clustered at the teacher level and given below the regression coefficients in parentheses.
$p<0.01$ ***, $p<0.05$ ** $p<0.1$ *

Table A4: Correlations between Test Scores and Student Characteristics

	Baseline knowledge	Baseline skills	Baseline motivation
Baseline scores			
Knowledge	1	0.501	0.059
	[0.000]	[0.000]	[0.452]
Skills	0.501	1	-0.006
	[0.000]	[0.000]	[1]
Motivation	0.059	-0.006	1
	[0.452]	[1]	[0.000]
Endline scores			
Knowledge	0.551	0.463	0.076
	[0.000]	[0.000]	[0.034]
Skills	0.439	0.603	0.022
	[0.000]	[0.000]	[1]
Motivation	0.058	0.005	0.529
	[0.499]	[1]	[0.000]
Student characteristics			
Grade 3	-0.13	-0.308	0.081
	[0.000]	[0.000]	[0.016]
Grade 5	0.153	0.325	-0.078
	[0.000]	[0.000]	[0.028]
Late student	-0.165	-0.143	0.023
	[0.000]	[0.000]	[1]
Female student	-0.051	0.056	-0.046
	[1]	[0.652]	[1]
Observations	2016	2016	2005

The table provides the correlation between our student test score measures and student characteristics across time (baseline versus endline) in the control group, years 2 and 3 pooled. In square brackets, we provide the Bonferroni-adjusted p-values.

Table A5: Treatment Heterogeneity - Year 2

			Student Heterogeneity				Teacher Heterogeneity			
			Top achiever		Girl		Science diploma		Woman	
		Obs.	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
Knowledge	T*H	2,415	-0.121*	-0.063	0.114	0.009	0.056	0.086	0.032	0.033
			(0.072)	(0.063)	(0.082)	(0.064)	(0.124)	(0.088)	(0.133)	(0.101)
	T		0.220***	0.131**	0.064	0.095*	0.070	0.044	0.096	0.072
			(0.066)	(0.059)	(0.071)	(0.051)	(0.096)	(0.072)	(0.103)	(0.080)
	H		0.904***	0.024	-0.148**	-0.071	-0.181*	-0.126*	0.083	0.017
			(0.058)	(0.073)	(0.067)	(0.047)	(0.094)	(0.070)	(0.099)	(0.070)
Skills	T*H	2,415	-0.028	0.008	0.107	0.023	0.010	0.036	-0.026	-0.050
			(0.074)	(0.065)	(0.073)	(0.060)	(0.106)	(0.078)	(0.121)	(0.093)
	T		0.058	0.015	-0.044	-0.001	0.007	-0.005	0.033	0.051
			(0.063)	(0.052)	(0.056)	(0.044)	(0.072)	(0.056)	(0.097)	(0.079)
	H		0.606***	-0.005	0.122**	0.155***	-0.068	-0.020	0.070	0.029
			(0.061)	(0.065)	(0.059)	(0.046)	(0.078)	(0.061)	(0.087)	(0.065)
Motivation	T*H	2,409	-0.016	-0.063	-0.094	-0.087	-0.006	0.007	0.121	0.095
			(0.089)	(0.073)	(0.086)	(0.077)	(0.084)	(0.071)	(0.093)	(0.079)
	T		-0.015	0.029	0.012	0.026	-0.027	-0.020	-0.122*	-0.086
			(0.063)	(0.057)	(0.056)	(0.052)	(0.062)	(0.052)	(0.066)	(0.064)
	H		0.170**	0.103	-0.049	-0.042	0.029	0.026	-0.102	-0.116*
			(0.066)	(0.069)	(0.066)	(0.055)	(0.069)	(0.056)	(0.066)	(0.060)
Covariates			N	Y	N	Y	N	Y	N	Y

The table provides the result of the heterogeneous treatment analysis in year 2 on the three endline student test scores (knowledge, skills, and motivation). In rows, T*H gives the interaction between the treatment variable and the heterogeneity variables, T gives the coefficient of the treatment variable, and H is the coefficient of the heterogeneity variable. In the first set of columns (*student Heterogeneity*), the heterogeneity is based on baseline student variables (being a top achiever at baseline, i.e., top 50% of the knowledge score at baseline and being a girl). In the second set of columns, we analyze the heterogeneity by baseline teacher characteristics (having a diploma in science and being a woman). Column (1) gives the result of the regression without baseline covariate, while column (2) the result conditional on baseline covariates, baseline hours of science taught, and baseline practices of inquiry-based learning. All regressions are weighted by sampling probabilities and include strata fixed effects. Standard errors are clustered at the teacher level and given below the regression coefficients in parentheses.
p<0.01 ***, p<0.05 ** p<0.1 *

Table A6: Treatment Heterogeneity - Year 3

			Student Heterogeneity				Teacher Heterogeneity			
			Top achiever		Girl		Science diploma		Woman	
		Obs.	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
Knowledge	T*H	2,216	0.001	-0.002	-0.210**	-0.142*	-0.247*	-0.131	0.281	0.171
			(0.092)	(0.080)	(0.087)	(0.077)	(0.127)	(0.089)	(0.171)	(0.111)
	T		0.063	0.035	0.140*	0.101*	0.174*	0.096	-0.166	-0.105
			(0.077)	(0.063)	(0.080)	(0.059)	(0.099)	(0.070)	(0.142)	(0.092)
	H		0.801***	0.038	0.081	0.009	0.084	0.045	-0.133	-0.054
			(0.071)	(0.067)	(0.067)	(0.058)	(0.095)	(0.071)	(0.101)	(0.067)
Skills	T*H	2,216	-0.023	-0.029	-0.017	0.023	-0.170	-0.065	0.176	0.070
			(0.085)	(0.067)	(0.080)	(0.070)	(0.118)	(0.089)	(0.150)	(0.102)
	T		0.034	0.044	-0.007	-0.001	0.068	0.026	-0.153	-0.061
			(0.075)	(0.058)	(0.071)	(0.054)	(0.092)	(0.069)	(0.121)	(0.078)
	H		0.616***	0.097	0.176***	0.084	0.022	-0.016	-0.058	-0.006
			(0.063)	(0.060)	(0.059)	(0.052)	(0.088)	(0.068)	(0.095)	(0.068)
Motivation	T*H	2,216	0.010	-0.051	-0.068	-0.075	-0.075	-0.021	-0.103	-0.070
			(0.074)	(0.061)	(0.090)	(0.077)	(0.091)	(0.070)	(0.119)	(0.099)
	T		-0.141**	-0.076	-0.104	-0.062	-0.082	-0.078	-0.059	-0.044
			(0.067)	(0.055)	(0.068)	(0.056)	(0.071)	(0.059)	(0.096)	(0.079)
	H		0.205***	0.109*	-0.028	0.021	0.091	0.056	-0.022	0.028
			(0.054)	(0.063)	(0.067)	(0.057)	(0.067)	(0.052)	(0.095)	(0.077)
Covariates			N	Y	N	Y	N	Y	N	Y

idem than the previous note but for year 3.

Figure A1: Knowledge questionnaire Example

One year is the time it takes for the Earth to complete one revolution around...	A. itself.	Grades 4 and 5
	B. the solar system.	
	C. the sun.	
	D. the moon.	

Figure A2: Skill questionnaire Example

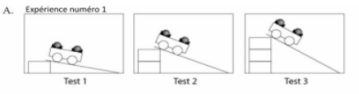
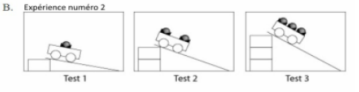
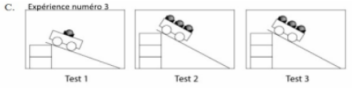
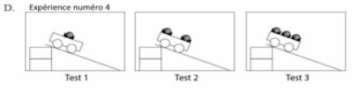
A classmate wants to experiment to test an idea: the heavier a car is, the faster it will go down a ramp. To do this, your classmate can change the car's weight by adding balls (one ball, two balls, or three balls), and he can change the height of the ramp by using different numbers of blocks (one block, two blocks, or three blocks). Your classmate hesitates between these four possibilities. What do you advise him to choose between A, B, C or D?	A. Expérience numéro 1  	Grades 4 and 5
	B. Expérience numéro 2  	
	C. Expérience numéro 3  	
	D. Expérience numéro 4  	

Figure A3: Satisfaction Survey - Year 1

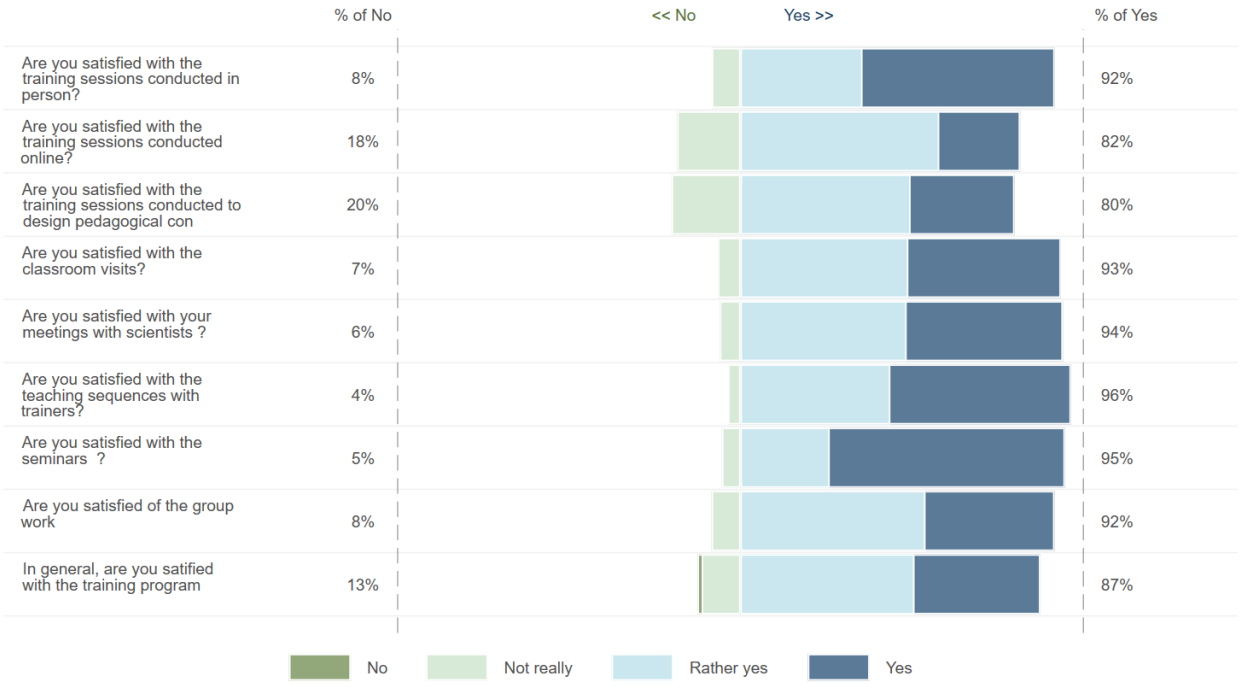


Figure A3 gives the result of the satisfaction survey conducted at the end of the first year of the training program (year 1). The survey only includes treatment teachers who participated in the training program.

Figure A4: Satisfaction Survey - Year 2

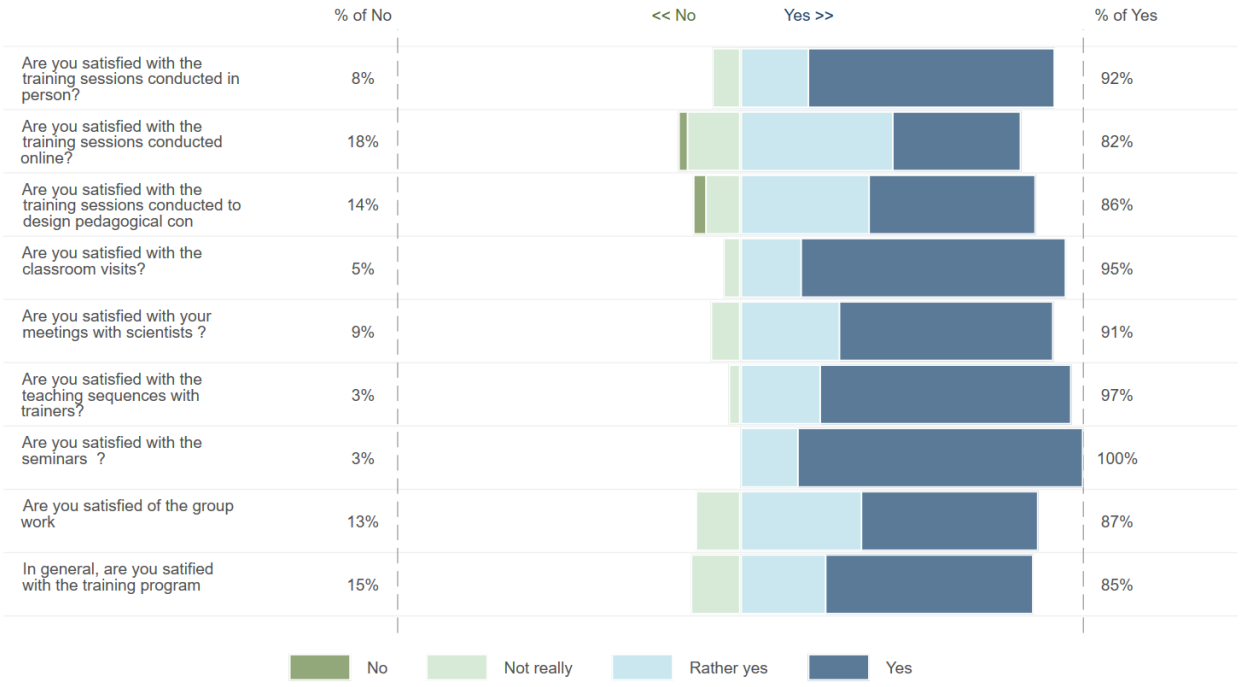


Figure A4 gives the result of the satisfaction survey conducted at the end of the second year of the training program (year 2). The survey only includes treatment teachers who participated in the training program.

Appendix B: Robustness Checks on Main Results

Table B1: Impacts on Students' Scores - Class Attrition Sample

			Treatment effect		Y2 vs Y3
	Obs.	Mean	(1)	(2)	(3)
Year 2					
Endline knowledge	2,427	0.742	0.090	0.071*	
			(0.060)	(0.043)	
			[0.68]	[0.42]	
Endline skills	2,427	0.531	0.000	0.002	
			(0.051)	(0.037)	
			[0.92]	[1.00]	
Endline motivation	2,420	-0.050	-0.043	-0.022	
			(0.043)	(0.040)	
			[0.68]	[1.00]	
Year 3					
Endline knowledge	2,473	0.512	0.031	0.021	*0.41*
			(0.063)	(0.049)	*[1.00]*
			[0.72]	[0.99]	
Endline skills	2,473	0.380	-0.034	-0.015	*0.72*
			(0.055)	(0.045)	*[1.00]*
			[0.72]	[0.99]	
Endline motivation	2,472	-0.053	-0.135***	-0.091**	*0.18*
			(0.046)	(0.039)	*[1.00]*
			[0.01]	[0.07]	
Number of clusters			112	112	
Additional controls			N	Y	

This table gives the impact of the program on students' endline test scores in year 2 (upper panel) and in year 3 (lower panel). The sample is restricted to classes that participated in both survey waves. Columns *Obs.* gives the number of students surveyed, *Mean* the average in the control group, which can be read as the progression during the year in terms of baseline standard deviations. In column *(1)*, we only control for Grade fixed effects. In column *(2)*, we add baseline scores, baseline hours of science taught, and baseline inquiry-based learning practices. All regressions include strata fixed effects and are weighted by sampling probabilities. Standard errors are clustered at the teacher level and are given below the regression coefficients in parentheses. The coefficients in square brackets are the p-values robust to multiple hypothesis testing. In column *(3)*, we provide the p-value of the statistical comparison of the coefficients across years 2 and 3 using column (2) estimates.
p<0.01 ***, p<0.05 ** p<0.1 *

Table B2: Impacts on Students' Scores - Teacher Attrition Sample

			Treatment effect		Y2 vs Y3
	Obs.	Mean	(1)	(2)	(3)
Year 2					
Endline knowledge	2,070	0.717	0.121*	0.086*	
			(0.062)	(0.044)	
			[0.19]	[0.19]	
Endline skills	2,070	0.488	0.022	0.001	
			(0.055)	(0.039)	
			[0.86]	[1.00]	
Endline motivation	2,067	-0.030	-0.035	-0.025	
			(0.046)	(0.044)	
			[0.80]	[1.00]	
Year 3					
Endline knowledge	2,107	0.481	0.016	0.022	*0.34*
			(0.070)	(0.052)	*[1.00]*
			[1.00]	[0.84]	
Endline skills	2,107	0.355	-0.017	0.021	*0.68*
			(0.065)	(0.051)	*[1.00]*
			[1.00]	[0.84]	
Endline motivation	2,106	-0.012	-0.153***	-0.098**	*0.21*
			(0.049)	(0.043)	*[1.00]*
			[0.01]	[0.08]	
Number of clusters			95	95	
Additional controls			N	Y	

This table gives the impact of the program on students' endline test scores in year 2 (upper panel) and in year 3 (lower panel). The sample is restricted to teachers who participated in both teacher survey waves. Columns *Obs.* gives the number of students surveyed, *Mean* the average in the control group, which can be read as the progression during the year in terms of baseline standard deviations. In column *(1)*, we only control for Grade fixed effects. In column *(2)*, we add baseline scores, baseline hours of science taught, and baseline inquiry-based learning practices. All regressions include strata fixed effects and are weighted by sampling probabilities. Standard errors are clustered at the teacher level and are given below the regression coefficients in parentheses. The coefficients in square brackets are the p-values robust to multiple hypothesis testing. In column *(3)*, we provide the p-value of the statistical comparison of the coefficients across years 2 and 3 using column (2) estimates.
p<0.01 ***, p<0.05 ** p<0.1 *

Table B3: Effects of Training Topics on Class Topics - No Additional Controls

	Topics covered in class					
	Year 1		Year 2		Year 3	
	(1)	(2)	(3)	(4)	(5)	(6)
(Y1 training topic)	0.325***		0.079		0.160*	
.(Treatment)	(0.082)		(0.082)		(0.083)	
	[0.00]		[0.20]		[0.13]	
			0.01			
			[0.02]			
(Y2 training topic)		-0.015		0.236***		0.129
.(Treatment)		(0.096)		(0.077)		(0.101)
		[0.78]		[0.01]		[0.13]
				0.02		
				[0.02]		
Y1 training topic	-0.010		-0.020		-0.082	
	(0.066)		(0.062)		(0.058)	
Y2 training topic		0.076		-0.063		-0.085
		(0.079)		(0.052)		(0.079)
Treatment	-0.048	0.035	-0.058	-0.080*	-0.017	-0.000
	(0.041)	(0.039)	(0.045)	(0.041)	(0.043)	(0.041)
Constant	0.548***	0.533***	0.551***	0.558***	0.542***	0.537***
	(0.031)	(0.030)	(0.034)	(0.029)	(0.034)	(0.033)
N teachers	93	93	96	96	96	96
N observations	744	744	768	768	768	768

This table shows the regression of a dummy for covering each of the eight possible topics in each class, each year, on a treatment dummy and dummies for that very topic being covered by the local training center. Each column is a different regression. The sample is restricted to teachers who answered in both survey rounds. All regressions include strata fixed effects and are weighted by sampling probabilities. Standard errors are clustered at the teacher level and are given below the regression coefficients in parentheses. In italics is the p-value of the differences between the current and previous years' estimated coefficients.
p<0.01 ***, p<0.05 ** p<0.1 *

Table B4: Effects of Training Topics on Class Topics - Unrestricted Sample

	Topics covered in class					
	Year 1		Year 2		Year 3	
	(1)	(2)	(3)	(4)	(5)	(6)
(Y1 training topic)	0.221***		0.147**		0.128	
.(Treatment)	(0.067)		(0.073)		(0.081)	
	[0.00]		[0.03]		[0.31]	
			0.40			
			[0.40]			
(Y2 training topic)		0.019		0.241***		0.088
.(Treatment)		(0.079)		(0.068)		(0.102)
		[0.68]		[0.00]		[0.31]
				0.02		
				[0.02]		
Y1 training topic	0.044		-0.028		-0.058	
	(0.051)		(0.053)		(0.054)	
Y2 training topic		0.063		-0.071		-0.066
		(0.064)		(0.047)		(0.078)
Treatment	-0.051	-0.001	-0.093***	-0.101***	-0.027	-0.010
	(0.033)	(0.030)	(0.034)	(0.033)	(0.041)	(0.038)
N teachers	129	129	119	119	102	102
N observations	1032	1032	952	952	816	816

This table shows the regression of a dummy for covering each of the eight possible topics in each class, each year, on a treatment dummy and dummies for that very topic being covered by the local training center. The estimated coefficients are conditional on baseline hours of science taught and baseline practices of inquiry-based learning. Each column is a different regression. The sample is restricted to teachers who answered in both survey rounds. All regressions include strata fixed effects and are weighted by sampling probabilities. Standard errors are clustered at the teacher level and are given below the regression coefficients in parentheses. In italics is the p-value of the differences between the current and previous years' estimated coefficients.
p<0.01 ***, p<0.05 ** p<0.1 *

Table B5: Impacts on Teacher Practice Indices - Unrestricted Sample

			Year 2 (Y2) Treatment effect				**Year 3 (Y3)** Treatment effect		
	Obs.	Mean	(1)	(2)	Obs.	Mean	(3)	(4)	Y2 vs Y3
Inquiry-based learning									
Declared practices	119	-0.000	0.104 (0.176) [0.50]	-0.046 (0.176) [0.47]	101	-0.000	0.440** (0.202) [0.09]	0.297 (0.208) [0.64]	*0.03* *[0.15]*
Normative statements	119	0.000	0.146 (0.180) [0.46]	0.105 (0.195) [0.42]	100	0.000	0.295 (0.208) [0.19]	0.242 (0.231) [0.65]	*0.44* *[0.39]*
Science intensity									
Weekly hours	119	1.429	0.252*** (0.093) [0.04]	0.180* (0.095) [0.27]	102	1.349	0.252** (0.107) [0.09]	0.186* (0.108) [0.64]	*0.95* *[0.62]*
Number of topics	119	4.420	-0.272 (0.267) [0.45]	-0.467* (0.269) [0.27]	102	4.260	0.075 (0.283) [0.46]	0.038 (0.309) [0.99]	*0.06* *[0.15]*
% hands-on experiments	119	0.652	0.088** (0.044) [0.11]	0.068 (0.046) [0.27]	102	0.644	0.025 (0.049) [0.45]	-0.004 (0.050) [0.99]	*0.17* *[0.20]*
Baseline covariates			N	Y			N	Y	Y

The table gives the program's impacts on the teachers' practice indices. Column *Obs.* gives the number of teachers, *Control* the average in the control group, *(1)* the treatment coefficients *(2)* the treatment coefficients conditional on baseline hours of science taught and baseline practices of inquiry-based learning. All regressions are weighted by sampling probabilities and include strata fixed effects. Robust standard errors are given below the regression coefficients in parentheses. In column "Year comparison" we provide the p-value of the statistical comparison of the coefficients across years 2 and 3.
p<0.01 ***, p<0.05 ** p<0.1 *

Table B6: Impacts on Teacher Outcomes, Detailed - Restricted Sample

	Year 1			Year 2			Year 3		
	C	(1)	(2)	C	(1)	(2)	C	(1)	(2)
Science intensity									
Weekly hours	1.557	0.068	-0.018	1.420	0.250**	0.166	1.336	0.257**	0.218*
		(0.108)	(0.107)		(0.100)	(0.103)		(0.113)	(0.115)
Number of topics	4.339	0.257	-0.021	4.356	-0.307	-0.639*	4.193	0.183	0.125
		(0.304)	(0.318)		(0.347)	(0.339)		(0.319)	(0.344)
% sessions w/ hands-on expe.	0.611	0.129**	0.100*	0.648	0.116**	0.096*	0.645	0.032	0.010
		(0.051)	(0.051)		(0.051)	(0.052)		0.032	0.010
% sessions w/o hands on expe.	0.319	-0.039	-0.009	0.349	0.012	0.026	0.365	-0.052	-0.037
		(0.066)	(0.072)		(0.072)	(0.078)		(0.072)	(0.074)
Declared practices index									
Introduces sci. problem, sd	.	.	.	-0.051	0.524**	0.460*	0.100	0.406*	0.271
	.	.	.		(0.224)	(0.248)		(0.216)	(0.224)
Works on students vision, sd	.	.	.	-0.092	0.186	0.051	0.055	0.253	0.200
	.	.	.		(0.213)	(0.218)		(0.211)	(0.224)
Evaluates students, sd	.	.	.	-0.077	-0.060	-0.247	0.026	0.209	0.095
	.	.	.		(0.200)	(0.198)		(0.204)	(0.211)
Normative statements index									
Importance of ...									
...Introducing sci. pb., sd	.	.	.	-0.074	0.350**	0.354*	-0.004	0.204	0.180
	.	.	.		(0.175)	(0.187)		(0.209)	(0.232)
...formulating hyp., sd	.	.	.	-0.005	0.310	0.237	-0.037	0.194	0.157
	.	.	.		(0.211)	(0.234)		(0.207)	(0.220)
...linking model to obs., sd	.	.	.	0.129	-0.091	-0.136	0.039	0.156	0.108
	.	.	.		(0.184)	(0.206)		(0.192)	(0.192)
...evaluating student, sd	.	.	.	-0.085	0.249	0.159	-0.013	0.373	0.331
					(0.187)	(0.194)		(0.260)	(0.293)
Baseline covariates					N	Y		N	Y

The table provides the impact on quantitative teacher practices. Column C gives the average in the control group, (1) the treatment coefficient, (2) the treatment coefficient conditional on baseline hours of science taught and baseline practices of inquiry-based learning. All regressions are weighted by sampling probabilities and include strata fixed effects. Standard errors are given below the regression coefficients in parentheses. p<0.01 ***, p<0.05 ** p<0.1 *

Appendix C: Robustness of Results on Students Motivation

To understand how robust the negative effect on motivation is, we create sub-components of the motivation index using a Principal Component Analysis (PCA). Following the Kaiser criterion, we retained all the components with an eigenvalue greater than one (Kaiser (1960)). This gave us three main components, from which we create a simple averaged index of the (normalized) variables strongly loaded on each factor. Those three sub-dimensions are balanced at baseline (cf. Table C2) and have a relatively high Cronbach Alpha[45]. We label the three sub-dimensions "I like science", "Scientific mindset" and "Science is easy".[46]

Table C3 presents the causal effects of the training on those three dimensions of motivation. At the end of year 2 (upper panel), the three coefficients are slightly negative but not significant in both column (1) – controlling for grades only – and column (2) – controlling for baseline scores, baseline hours of science taught, and baseline practices of inquiry-based learning. In survey year 3 (bottom panel), the three coefficients are negative (between -0.4 SD and -0.09 SD) in both columns and very significant, even when controlling for multi-hypothesis testing. This indicates that the negative motivation effect is a robust feature of the data, not driven by a few items or mere chance.

[45]The first component has a Cronbach Alpha above 0.85, the second of about 0.6 and the third one of about 0.5.

[46]The details of those new indexes are in the Appendix Table C1.

Table C1: Sub-Components of the Motivation Index

"I like science"	**"Scientific mindset"**	**"Science is easy"**
Component 1	**Component 2**	**Component 3**
I love science	I am always curious about how new technologies work	I find science easy
Later, I plan to study science	To understand science, experiences are better than lessons	I do well in science
At home I like to play scientific games	I like to have scientific evidence before I think something is true	I like to observe plants and animals when I go for a walk.
I like to discuss science with my classmates	I prefer to learn science by doing experiments	I like to take my toys apart to try and figure out how they work.
I would like to participate in science competitions		
Science is my favorite subject		
I think I have a scientific mind		
I like to watch science shows on TV or on my computer.		
I like to read magazines and science books.		

The tables describe the item content of the three components of the motivation index.

Table C2: Balance Checks - Sub-Components of the Motivation Index

	Treatment v. Control			**Volunteer v. Peer**		
	Obs.	Control	(1)	Obs.	Peer	(2)
Year 2						
I like science	2,670	0.002	0.025	4,459	0.064	-0.044*
			(0.032)			(0.026)
Scientific mindset	2,658	-0.012	-0.054	4,439	-0.002	-0.038
			(0.033)			(0.027)
Science is easy	2,660	-0.008	0.035	4,437	0.043	-0.034
			(0.025)			(0.024)
Year 3						
I like science	2,516	0.026	-0.032	3,951	0.031	-0.017
			(0.038)			(0.031)
Scientific mindset	2,507	-0.030	-0.066**	3,938	-0.046	-0.006
			(0.031)			(0.028)
Science is easy	2,511	-0.008	0.015	3,943	-0.033	0.031
			(0.032)			(0.030)
Number of clusters			134			134

The table provides the baseline difference between the treatment and control students and the baseline difference between the students of the volunteer teachers and the peer students. Column *Obs.* gives the number of observations, *Control* the average in the control group, (1) the difference between treatment and control, *Peer* the average in the group of peer students, and (2) the difference between students of the volunteer teacher and the peer students. All regressions are weighted by sampling probabilities and include strata fixed effects. Standard errors are clustered at the teacher level and given below the regression coefficients in parentheses.
$p<0.01$ ***, $p<0.05$ ** $p<0.1$ *

Table C3: Impacts on Students' Scientific Motivation

			Treatment v. Control	
	Obs.	Control	(1)	(2)
Year 2				
I like science	2,686	-0.095	-0.015	-0.001
			(0.028)	(0.027)
			[1.000]	[1.000]
Scientific mindset	2,685	0.055	-0.028	-0.009
			(0.026)	(0.026)
			[1.000]	[1.000]
Science is easy	2,685	0.011	-0.020	-0.022
			(0.024)	(0.026)
			[1.000]	[1.000]
Year 3				
I like science	2,488	-0.059	-0.069**	-0.045*
			(0.032)	(0.025)
			[0.013]	[0.032]
Scientific mindset	2,488	0.043	-0.080***	-0.060**
			(0.025)	(0.026)
			[0.006]	[0.026]
Science is easy	2,487	-0.036	-0.076***	-0.071***
			(0.028)	(0.027)
			[0.009]	[0.026]
Number of clusters			124	114
Baseline covariates			N	Y

The table provides the impact of the program on the motivation index sub-components. Column *Obs.* gives the number of observations, *Control* the average in the control group, (1) the difference between treatment and control, (2) the treatment coefficient conditional on baseline hours of science taught and baseline practices of inquiry-based learning. All regressions are weighted by sampling probabilities and include strata fixed effects. Standard errors are clustered at the teacher level and given below the regression coefficients in parentheses. In square brackets, we provide the p-values robust to multiple testing. $p<0.01$ ***, $p<0.05$ ** $p<0.1$ *