Discussion Paper Series – CRC TR 224

Discussion Paper No. 455
Project A 03

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

Suzanne Bellue [1]
Adrien Bouguen [2]
Marc Gurgand [3]
Valerie Munier [4]
André Tricot [5]

August 2023

[1] University of Mannheim, Email: sbellue@uni-mannheim.de
[2] Santa Clara University, California, USA
[3] Paris School of Economics, CRNS & ENS-PSL
[4] Université de Montpellier et Université Paul Valéry Montpellier 3, France
[5] Université Paul Valéry Montpellier 3, France

Support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through CRC TR 224 is gratefully acknowledged.

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‖

August 22, 2023

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 large-scale randomized study of an intensive in-service teacher training program conducted in France during and after the training program's implementation year. 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 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 that entered the research project. 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).

†University of Mannheim, Germany. 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 multiple cohorts, effective training programs can be regarded as one of the most cost-effective educational policy tools. However, the long-term effectiveness of these programs hinges on teachers' ability to apply the acquired teaching skills in their regular classes routinely. While short-term evaluations can provide valuable insights into the effectiveness of skills taught during the training, it is crucial to 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 of inquiry-based learning during and after its implementation. Our training program targets French school teachers of Grades 3 to 5, who generally teach all the subjects to one class of students. French primary school teachers follow the national curriculum that specifies, among other things, the eight science topics to be covered and recommends the use of the inquiry-based learning method—a method considered one of the best pedagogical approaches to teaching science that the United States National Research Council has long endorsed (Council et al., 2000).[1] Our training program is independently implemented by nine local training centers, *Maisons pour la Science*, in three regional school districts. A large part of the training content consists in designing a teaching sequence based on one or two of the eight science topics. The content of such a sequence can be easily re-used in class, sometimes with the help of a trainer. The training program consists of a total of 80 hours over two years, making it intensive with respect to routine training programs—only 9.5 hours per year in the three regional school districts[2]—and likely to be effective with regard to the education literature.[3]

We conduct a large randomized evaluation in which we allocate the 134 teachers who registered for our training program into a treatment and a control group. Only

[1]Inquiry-based learning in science 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.

[2]We obtain this number by surveying the peers of registered teachers.

[3] Compared with the international literature, the intensity of the program we investigate is also in line with the programs that have proven effective in the past: equivalent to Newman et al. (2012) but lower than Meyers et al. (2016) (see Table A1).

teachers in the treatment group benefit from the two-year training program. We collect information for all registered teachers and their respective students during two academic years: during the second (and last) year of our in-service teacher training program and one year after, once teachers no longer benefit from trainers' support. We assess the scientific skills, knowledge, and motivation of the two successive cohorts of students at the beginning and the end of both academic years. Each of the student cohorts is of about 2,500 students. In addition, we collect information on teachers' pedagogical skills and practices in science, including very detailed information on topics covered in the classroom, a well-defined and easily measurable dimension of practices. We can cross this data with information on the topics covered in each training center within each regional school district, which vary. Importantly, control teachers are all in the catchment area of a training center, so we know which topics they would have been exposed to if they had been in the treatment group. This design, with information at the training center, teacher and student levels, and over two years, helps interpret the mechanisms underlying the potential effectiveness of our program. Two years of evaluation allow us to capture the immediate and medium-term effects of the training program. With teacher's and student-level data, we can see whether our program affects teachers' practices and ultimately affects students' achievement and motivation. Finally, the variation in topics covered in training sessions enables us to assess the impact of those sessions on teaching practices. Specifically, we can determine whether covering a specific topic in the training affects both the likelihood a teacher teaches that topic in the classroom and the manner she does so. Furthermore, we can observe whether teaching practices evolve for topics that were not covered during the training sessions.

While we face inherent difficulties associated with conducting and evaluating a teacher training program at a large scale, we see high program adherence and low levels of differential attrition. The first difficulty regards selection in teacher enrollment. By comparing registered teachers to their non-registered peers who work at the same schools, we observe 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 scale-up context, those results highlight potential challenges that would face policymakers who wish to target intensive science training programs to teachers with low science knowledge levels. Nevertheless, treatment teachers are highly satisfied with the training program and almost perfectly adhere

to it: they receive an average of 66 hours of in-service training over the two years, which is fairly close to the objective of 80 hours.[4] 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 to teaching in 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.

Our first main results highlight the short-run effectiveness of the training program in improving students' achievement. At the end of the first evaluation year, while students' motivation and scientific skills are similar across the two treatment arms, students' scientific knowledge is about ten percentage points of a standard deviation (SD) higher in the treatment group than in the control group. However, the positive effects of our training program vanish a year later when teachers no longer receive support from trainers. At the end of the second evaluation year, 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 ten percentage points of an SD than in the control group.

We investigate teachers' practices to understand the short-run effectiveness of the program on students' achievement. Estimates on teacher outcomes suggest that our training program leads to a change in teaching practices toward inquiry-based learning guidelines. In both evaluation years, treatment teachers teach more science per week (+0.17 and +0.22 hours), and their inquiry-based declared practices and knowledge indices tend to outperform those of control teachers. The effects on pedagogical skills and knowledge indices are insignificantly different from zero but rather large, especially one year after the end of the training program (+0.24 and +0.27 SD). The fact that all the point estimates are positive and become more prominent after the end of the program is consistent with an effective teacher training program that leads teachers to follow inquiry-based learning guidelines even after the training program ends.

Importantly, however, we observe two main differences in teachers' practices between the two evaluation years. While during the training year, treatment teachers concentrate their teaching on topics covered during the training sessions, one year af-

[4]See Table 3. The difference between the 80 scheduled hours and the 66 effective hours, as reported by treatment teachers, is likely due to imperfect take-up (95% take-up in the treatment group), absenteeism, and recall bias.

ter, they revert to also teaching uncovered topics. In addition, the program's positive effect on the frequency of hands-on experiments—experiments at the heart 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 approach is entirely driven by a change in teaching practices for the very topics covered during the training sessions. Because hands-on experiments must be tailored to each science topic, this suggests that, without the trainers' support, treatment teachers fail to design new experiments that involve students' participation. Furthermore, the negative effect of our program on students' motivation in the second evaluation year could indicate difficulties teachers face when independently implementing inquiry-based pedagogy to a different set of science topics.[5] Overall, those results call for teacher training programs that offer longer-term support from trainers or comprehensive curriculum coverage, providing teachers with a broad range of tools they can use in their regular classes.

Contributions and Related Literature

Our results relate to three strands of the literature and provide new insights into the effectiveness of an intensive in-service training program in the short and longer term.

Our results first relate to the literature that rigorously evaluates the effectiveness of in-service teacher training programs in a policy-relevant context. Despite the relevance of evaluating those programs, rigorous evidence is relatively scarce (Gersten et al. (2014); Yoon et al. (2007)).[6] Our paper probably most closely relates to the recent study of Loyalka et al. (2019), who conduct a thorough evaluation of a 15-day government-sponsored teacher training program in China, with measures at the teacher and student levels. Our intervention settings mimic a policy scale-up context, are based on a validated pedagogical method, and we also find high satisfaction rates among treatment teachers. However, in contrast to Loyalka et al. (2019)'s study, our intervention program consists of 80 hours of training spread out over two years—against a one-shot 15 days of training. In addition, we find positive impacts on students' performance during the program implementation period—while they find no impacts

[5]This interpretation is consistent with findings in the literature that the effectiveness of inquiry-based learning is very sensitive to the quality of its implementation and the guidance provided to students (Kirschner et al., 2006; Crawford, 2007; Lazonder and Harmsen, 2016).

[6]For instance, of the 643 United States studies of math teacher training interventions in Grades K–12, Gersten et al. (2014) find that only five of them meet What Works Clearinghouse evidence standard. Over 1,300 identified PD studies in the United States, Yoon et al. (2007) find that only nine have pre-and post-test data and a control group.

at all. Our result is consistent with recent reviews that suggest that only intensive programs can be effective (Yoon et al. (2007); Fryer (2017)).[7] In Appendix Table A1, which displays the estimated effectiveness of 18 recent evaluation studies, only studies that tend to be very intensive show positive impacts.[8] For instance, Borman et al. (2007) find positive effects of the Success for All program in the United States, and Sirinides et al. (2018) find that the Reading Recovery program improves first graders' reading skills. Other results based on non-experimental variation, such as Bouguen (2016); Machin and McNally (2008); Angrist and Lavy (2001), confirm that intensive training programs, based on validated pedagogical approaches, can efficiently modify teaching practices and students' performance.

Importantly, our study evaluates the effects of the teacher training program one school year after its end, assessing whether our program achieves its intended goals of enhancing the academic outcomes of several cohorts of students. Very few studies have investigated the teacher-training program's effectiveness after the intervention's end. Some exceptions are Borman et al. (2007); Sirinides et al. (2018).[9] While we find that the training program effect disappears after its end, Borman et al. (2007) and Sirinides et al. (2018) find a positive effect on students' achievement during and after the training program. However, one key difference between those two studies and ours is the ongoing support teachers receive during the evaluation. Indeed, after the initial training year, the Success for All program includes approximately 15 days of additional training each 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 our case, teachers are entirely left on their own once the training program concludes. Our findings suggest that, even though our

[7] In his recent review of the experimental literature, Fryer (2017) finds that among 21 rigorous and experimental professional development studies in high-income countries, only five show positive impacts. Several encouraging results were found in low-income countries (Piper et al., 2018; Cilliers et al., 2020; Kerwin and Thornton, 2021; Cilliers et al., 2022), but the low level of initial teacher training makes these results hard to compare to our context.

[8] We consider post-2000 studies that are rigorous and sufficiently powered. We did not conduct a systematic literature review. Eleven studies have been identified in Fryer (2017) or in Yoon et al. (2007). The rest of the studies are more recent, conducted outside the United States, or use quasi-experimental methods. A study is well-powered if it has an ex-ante MDE below 0.30 SD.

[9] The study of 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.

training program effectively improves students' scores during its implementation, it falls short in the classroom as teachers struggle to implement inquiry-based learning guidelines for a different topic set without the trainers' support.

If our results do not directly speak to the effectiveness of follow-up interventions, they call for professional development programs that provide relevant follow-up and contribute to the implementation research literature on training practitioners.[10] This literature identifies key components for training programs' success, such as the need to motivate practitioners (Kealey et al., 2000), the need for specific training for a given curriculum (Ross et al., 1991), and the need for guidance on the flexible use of specific guidelines (Dansereau and Dees, 2002). In our context, we see that the program successfully motivates the teachers who express high satisfaction levels and successfully change their practices, mainly when teaching topics covered during the training sessions. However, it failed to provide flexible enough tools so that teachers could apply them to all regular classes.

Finally, our experiment relates to the literature that investigates the impact of training teachers on the inquiry-based learning method. Since Bruner (1961), the inquiry-based approach has become popular and has been the subject of abundant theoretical literature. However, the benefit of training teachers using this approach has rarely been rigorously evaluated at scale. To the best of our knowledge, among the 18 recent rigorous studies we identified in the literature (see Table A1), only two specifically analyze the impact of an inquiry-based teacher training program on either science, technology, or mathematics (STEM). Both studies find positive effects on the achievement of Grade 4 to 8's students. In the first study, conducted in Alabama, Newman et al. (2012) find that a reasonably intensive inquiry-based training program has small but significant effects on mathematical skills (+0.05 SD) but no significant impacts on scientific skills.[11] In the second study, in Missouri Meyers et al. (2016) find that a very intensive program (240 hours of training over two years) impacts students' performance in mathematics.[12] Our study provides additional evidence that

[10]For a review, see Fixsen et al. (2005). The importance of follow-up interventions is frequently mentioned, such as in Parsad et al. (2001); Guskey (2002); Popova et al. (2022).

[11]The intervention is relatively similar to ours and includes ten days of teacher training during the summer and several follow-up training sessions over two years. The evaluation analyzes the impact on Grade 4 to 8's students on math, science, and technology.

[12]The program is comprehensive (at the school level), intensive, and relies heavily on inquiry-based principles. The study targets seventh- and eighth-grade students and finds impacts on mathematics skills between 0.13 SD and 0.18 SD three years after the beginning of the program. However, since the randomization is done at the school level, the authors cannot exclude that surveyed students were

an intensive inquiry-based program can lead to a change in teacher practices and to an increase in students' scientific knowledge, even in low Grades (3-5).

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. Finally, section 6 concludes.

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 school districts. The *Maisons* then designed and implemented our teacher 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

exposed for several years to trained teachers.

teacher's initial certification obtained in national certification/training centers.[13] 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.[14] 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.

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.[15]

[13] 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.

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

[15] 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 A2. The training was conducted

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.[16] Any primary school teachers in one of those school districts could register for the training program. A total of 134 teachers registered 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". 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.[17] Note that because teachers usually change student groups between two consecutive school years, students in years 2 and 3 are, in most cases, different students.[18] Consequently, our data set comprises a panel of teachers and a repeated cross-section of students.

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

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.

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

[17] 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.

[18] 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.

"peer teachers". To create the group of peer teachers, we asked each of the registered teachers to name one of their school colleagues in Grades 3 to 5 who would be willing to participate in our survey.[19] 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.[20] Prior to applying for the *Maisons*' training, they are also more likely to have 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.[21]

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 effort 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 A4 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.

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

[20] 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).

[21] 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.

3.2 Teacher Randomization

To measure the effectiveness of the program, 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 most districts (five school districts), we stratified the randomization using baseline teaching experience only. In three other school districts, we additionally used both experience and an index of teaching practices. In the school districts where several teachers from the same school applied to the program, we stratified at the school level. Finally, in one district, teachers did not fill out Q0 before randomization; in this case, we used the municipality as a stratification variable. Finally, 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 this specificity by including sampling weights proportional to the inverse of the assignment probability in regressions; we also include strata fixed effects.

The first panel of 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. [22] 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 Appendix Table A4, we compare average students' test scores at the beginning of each of the two evaluation years (baselines) between control and treated teachers.[23] All indices are balanced at both baselines —i.e., at the beginning of each school year—

[22] 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.

[23] 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.

suggesting an absence of selection in students' assignments between treatment arms.

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. 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 concern teaching practices in science for Grade 3 to 5, there are three reasons for teacher attrition displayed in the last panel. 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)).[24] The bulk of the class and teacher attrition rates is 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 A3 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).[25] 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

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

[25] 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 1: Pre-Randomization Teacher Characteristics

	Treatment v. Control			Volunteer 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	1970	-0.631	214	1971	-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 panel. In the second panel, 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 *

baseline "practices inquiry-based" to our regressions.

Table 2: Differential Attrition

	Obs.	Control	(1)	(2)
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)	
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 *

Figure 1: Survey Protocol

School years	School month	Training	Teacher Survey	Student Survey
2013-2014	June		Q0 survey	Q0 survey
		Random Assignment		
2014-2015	Sept	**1st year of training**		
	June			
2015-2016	Sept	**2nde year of training**	Q1 Survey	Q1 Survey
	June		Q2 Survey	Q2 Survey
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.

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.[26] Being assigned to the treatment group significantly increases the teacher's probability of being enrolled in the 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.[27] 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 spill-over 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. Likewise, we do not find any significant spill-over on enrolling in other training programs one year after the end of the *Maisons* program. By year 3, teachers in both

[26] 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.

[27] $4.54/0.155 = 29.3$ 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 *

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, in results not presented here, treatment teachers expressed high satisfaction with the program: 86% were somewhat or very satisfied after the first training year and 87% after the second training year. They expressed satisfaction with all aspects of the training: in-class visits (95% are satisfied), on-site 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 effect of the intervention on students' achievement in science, 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:[28]

- *Scientific knowledge*: This index assesses the scientific knowledge of students. All the questions are based on the French curriculum in science for Grades 3 to 5. In accordance with 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). This feature makes it possible to observe the progression of pupils over time.

- *Scientific skills*: This index aims at assessing skills developed with inquiry-based learning, such as scientific reasoning or analogical reasoning. The questions refer to situations consistent with the French curriculum and, whenever possible, are taken from the existing literature.

[28] Our companion paper, Munier et al. (2021), describes our instruments in greater detail.

- *Scientific motivation*: This index captures students' attitudes toward science. This instrument is largely 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 A5, our measures of student knowledge and skills are highly correlated with each other ($\rho = 0.501$) but not perfectly correlated, which suggests that they measure different dimensions of scientific performance. The correlation between knowledge and motivation is much lower ($\rho = 0.05$), while the one between skills and motivation is null. Second, the three test scores are correlated over time between baseline and endline (ρ well above 0.5). Because our indices are standardized using the control group's *baseline* scores, the averages in *control 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 A5 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.[29]

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 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:

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

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

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.[30] In the first column of Table 4, we control for the Grade level and the strata fixed effects, whereas in the second column, in addition to baseline hours of science taught and baseline practices of inquiry-based learning, we also control for baseline scores, which increases the precision of the estimates.[31] All OLS regressions are with robust standard errors clustered at the teacher level. 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 therefore be directly interpreted as effect sizes.

At the end of year 2 – the year during which the teacher receives her final year of training – students in the treatment group outperform students in the control group. After controlling for baseline results (Column 2 of Table 4), the impact is positive (+ 0.1 SD) and significant at 5%. 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 prime objective of the program.[32]

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 1% even when controlling for multi-hypothesis testing. This somewhat unexpected result is very robust. In the Appendix Table C3, we decompose the motivation index into three sub-indices. All of them are negatively impacted in year 3. Furthermore, this phenomenon is observed in each of the three regional school districts (F-test 0.06, p-value 0.94 for the test of equal effect in each region).

The change in class attrition rates across the two evaluation years does not drive the

[30] 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.

[31] We impute missing observations at baseline and add a dummy variable that indicates imputation as a control variable. This imputation strategy avoids losing observations.

[32] 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).

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 which 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 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 +0.07SD 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 A6 and A7 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.

In the next section, we investigate our teachers' outcomes to help rationalize the decreasing effects of the training program on students' achievement.

5 Training Program's Effects on Teachers' Practices

5.1 Measures of Teacher Outcomes

To construct our teacher outcomes, we leverage the rich data obtained from teacher questionnaires covering the two years: the second training year (year 2, Q2), and the post-training year (year 3, Q4).[33] 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 dimensions that are internally consistent, with a Cronbach alpha above 0.7, and aggregate them in one index.[34]

[33] In the Appendix, we also present the results for the first training year (year 1, Q1).

[34] As a result, for the *Declared practices* index, we are left with "introducing scientific problems", "framing students' vision", and "evaluating students". 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 at all, and

Table 4: Impacts on Students' Scores

		Treatment v. Control		
	Obs.	Control	(1)	(2)
Year 2				
Endline knowledge	2,694	0.737	0.116**	0.097**
			(0.057)	(0.041)
			[0.152]	[0.058]
Endline skills	2,694	0.542	0.013	0.015
			(0.048)	(0.035)
			[1.000]	[0.821]
Endline motivation	2,686	-0.071	-0.036	-0.018
			(0.040)	(0.037)
			[0.587]	[0.821]
Year 3				
Endline knowledge	2,489	0.514	0.029	0.018
			(0.061)	(0.048)
			[0.734]	[1.000]
Endline skills	2,489	0.374	-0.030	-0.010
			(0.054)	(0.044)
			[0.734]	[1.000]
Endline motivation	2,488	-0.051	-0.131***	-0.094**
			(0.045)	(0.038)
			[0.012]	[0.048]
Number of clusters			124	114
Controlling for baseline variables			N	Y

The table gives the impact of the program on student performance. Column *Obs.* gives the number of students surveyed, *Control* 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 brackets p-values robust to multiple testing.
$p<0.01$ ***, $p<0.05$ ** $p<0.1$ *

- *Normative Statement*: The teacher survey also contains questions about the perceived importance of the five main practices for teaching science.[35] We aggregate four of the five normative dimensions into a *Normative statements* index.[36]

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.[37] 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 in both survey waves, in years 2 and 3.[38] 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 in 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

answers do not correlate with student performance, so we do not present this data here.

[35]For instance, we ask *Should inquiry-based teaching 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.*

[36]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).

[37]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.

[38]In the Appendix, we present the Tables for the unrestricted sample. Results are qualitatively the same. In addition, Table B2 presents the program's impact on students' outcomes considering the restricting sample of teachers who answered both survey waves. Results are qualitatively unchanged.

generate variation in the topics covered during training.[39]

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 covered 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 covered topic p in year t. Accordingly, define $Z_{c(j)pt}$ if topic p has been used in the training center $c(j)$ 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 training topics should not affect control teachers. The parameter of interest is β_2; it is positive if trained teachers use the training material in their class. Finally, β_3 would be positive if treatment teachers covered more topics and negative otherwise. We also estimate a variant of this equation using $Z_{c(j)pt-1}$ to learn whether training from earlier years remains influential.[40]

Table 5 gives the above regression results.[41] The constant coefficients of Table 5 indicate that slightly more than half of the existing topics were covered in class in year one, on average, in the control group. Columns (1) and (4) show same-year relationships, whereas the other columns verify whether training topics from an other year influenced the class topics in the current year. In Columns (1) and (4), the interaction terms (β_2) indicate that this probability of teaching a topic that has been covered by the training program in that same year increases by 33 and 24 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

[39] In one of the regional districts that contains 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.

[40] Of note, in this analysis, we have a maximum of 134 teachers * 8 topics = 1072 data points.

[41] Appendix Tables B3 show the results if we control for baseline covariates, and Table B4 presents the results on the entire sample. The coefficients are of similar magnitude.

Table 5: Effects of Training Topics on Class Topics

	Year 1 Class topic		Year 2 Class topic		Year 3 Class topic	
	(1)	(2)	(3)	(4)	(5)	(6)
Training topic Year 1 x Treatment	0.325***		0.079		0.160*	
	(0.082)		(0.082)		(0.083)	
Training topic Year 2 x Treatment		-0.015		0.236***		0.129
		(0.096)		(0.077)		(0.101)
Training topic Year 1	-0.010		-0.020		-0.082	
	(0.066)		(0.062)		(0.058)	
Training topic Year 2		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)
	744	744	768	768	768	768

The 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 in the local training center. Each column is a different regression. 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. $p<0.01$ ***, $p<0.05$ ** $p<0.1$ *

coefficient in Column (2) indicates that training topics from 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 (+0.08). 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 13 percentage points, respectively.

Interestingly, in year 2, the negative coefficient on the treatment dummy indicates that fewer different topics were covered in the treatment group, suggesting that the training may have induced teachers to cover fewer topics. In year 3, this effect disappears, suggesting that treatment teachers revert to teaching topics uncovered by the training program. In Table 6, we more directly regress the number of different topics covered in the year on the treatment dummy: it is indeed 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 take a different approach only during training. Overall, our results indicate that trainers influenced teachers' topic choices, especially during the training program.

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). All observations are weighted by randomization probabilities, and we present robust standard errors.

Table 6: Impacts on Teacher Practice Indices

	Year 2 (Y2)				Year 3 (Y3)				
			Treatment effect				Treatment effect		
	Obs.	Mean	(1)	(2)	Obs.	Mean	(3)	(4)	Y2 vs Y3
Inquiry-based learning									
Declared practices	96	-0.05	0.262 (0.208)	0.106 (0.217)	96	0.04	0.365* (0.215)	0.239 (0.221)	0.39
Normative statements	96	-0.01	0.292 (0.187)	0.231 (0.208)	95	-0.01	0.318 (0.227)	0.266 (0.249)	0.81
Science intensity									
Weekly hours	96	1.42	0.250** (0.100)	0.166 (0.103)	96	1.34	0.257** (0.113)	0.218* (0.115)	0.58
Number of topics	96	4.36	-0.307 (0.347)	-0.639* (0.339)	96	4.19	0.183 (0.319)	0.125 (0.344)	0.01
% hands-on experiments	96	0.65	0.116** (0.051)	0.096* (0.052)	96	0.65	0.032 (0.050)	0.010 (0.055)	0.12
Baseline Covariates			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 *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 *

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.[42] Even though the effects are insignificant, the point estimates of the effects are large and close to the acceptance standards.[43] This is particularly true one year after teachers have completed the entire training program, in year 3 (0.24SD and 0.27SD controlling for baseline variables).[44] These results suggest that the program effectively improved teachers' knowledge about inquiry-based learning guidelines and induced them to implement those guidelines more often, even one year after the program ends.[45]

5.3.2 Science intensity

The training program also affects our more quantitative measures of teachers' practices in science: the weekly hours of science, the number of covered topics, and the share of topics that includes hands-on experiments. In both evaluation years, treatment teachers report teaching +0.17 and +0.22 weekly hours of science more than control teachers (Table 6).

Interestingly, 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 concentrate their class on fewer science topics (-0.64) that have been covered by the training program (see Table 5), they revert to teaching a similar number of topics as control teachers in year 3 (+0.13). In addition, we see a +10 pp increase in the fraction topics in which teachers include hands-on experiments, but only in year 2.[46] Indeed, in year 3, the point

[42] Appendix Table B6 displays the effects of the program on the subindices.

[43] Note that our analysis at the teacher level is based on a maximum of 134 data points and therefore suffers from limited statistical power.

[44] 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 0.2SD.

[45] These results are self-declared and may be subject to desirability bias. Teachers may only answer our questions based on what they heard in the training sessions. Nevertheless, this reasoning implies that teachers have learned inquiry-based learning guidelines, which is already a success of the program.

[46] Notice that the percentage of science topics taught with experiments conducted by the teacher only 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.

estimate drops to 0.01.[47] It suggests that, in addition to influencing the topic choice of teachers, the program successfully incentivized 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.

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. 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). In year 3, the effect for covered topics remains positive but smaller and insignificant (+0.10). 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). Since hands-on experiments must be tailored to each science topic, this indicates that treatment teachers were unable to design hands-on experiments for different topics than 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.

5.4 Discussion

Our findings on teachers' outcomes clearly indicate that the training program influenced teachers' practices. During the program implementation, treatment teachers spend more time teaching science. They understand and declare implementing inquiry-based learning guidelines, teach topics covered by the program, and use hands-on experiment designs developed during the training for their science sequences. We also see that in year 2, teachers concentrate their effort 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 often and spend more time teaching science than control teachers.

The measurable difference in teaching practices between the two evaluation years resides 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

[47] Even though the sample size is small, the difference in effects between year 2 and year 3 is almost significant with a p-value of comparison of coefficients of 0.12.

Table 7: Impacts on Hands-On Approach

	Year 2 (Y2)				Year 3 (Y3)				
			Treatment effect				Treatment effect		
	Obs.	Mean	(1)	(2)	Obs.	Mean	(3)	(4)	Y2 vs Y3
% topics with hands-on experiments									
if covered topics	61	0.61	0.269**	0.294**	56	0.59	0.136	0.103	0.24
			(0.122)	(0.138)			(0.123)	(0.139)	
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)	
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 *Y2 vs Y3*, 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 *

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 decay in the use of the training material in class is unlikely caused by teacher dissatisfaction with it. This could, however, be explained by the fact that teachers must cover the full science curriculum, and may have lacked time in the second evaluation year. Indeed, in year 3, treatment teachers teach more topics but do not increase the number of hours in science compared to year 2. They may have chosen to include fewer time-consuming hands-on experiments in their science sessions.

A second explanation relies on 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 help of the trainers, treatment teachers faced difficulties in designing new experiments that involve students' participation, and likely did not maintain the quality of previous year's activities. This interpretation is consistent 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 is consistent with difficulties teachers may have faced when implementing inquiry-based learning pedagogy to a different set of science topics independently. 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).[48] Overall, those results suggest that trainers' support was key in improving children's outcomes. They call for teacher training programs that offer longer-term support from trainers or comprehensive curriculum coverage, providing teachers with a broad range of tools they can use 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. Assessing the effectiveness of such programs hence 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 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

[48] 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.

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.

Angrist, J. D. and Lavy, V. (2001). Does teacher training affect pupil learning? evidence from matched comparisons in jerusalem public schools. *Journal of labor economics*, 19(2):343–369.

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.

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.

Bos, J. M., Sanchez, R. C., Tseng, F., Rayyes, N., Ortiz, L., and Sinicrope, C. (2012). Evaluation of quality teaching for english learners (qtel) professional development. final report. ncee 2012-4005. *National Center for Education Evaluation and Regional Assistance*.

Bouguen, A. (2016). Adjusting content to individual student needs: Further evidence from an in-service teacher training program. *Economics of Education Review*, 50:90–112.

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

Campbell, P. F. and Malkus, N. N. (2011). The impact of elementary mathematics coaches on student achievement. *The Elementary School Journal*, 111(3):430–454.

Cilliers, J., Fleisch, B., Kotze, J., Mohohlwane, N., Taylor, S., and Thulare, T. (2022). Can virtual replace in-person coaching? experimental evidence on teacher professional development and student learning. *Journal of Development Economics*, 155:102815.

Cilliers, J., Fleisch, B., Prinsloo, C., and Taylor, S. (2020). How to improve teaching practice? an experimental comparison of centralized training and in-classroom coaching. *Journal of Human Resources*, 55(3):926–962.

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.

Dansereau, D. F. and Dees, S. M. (2002). Mapping training: The transfer of a cognitive technology for improving counseling. *Journal of substance abuse treatment*, 22(4):219–230.

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.

Fixsen, D. L., Naoom, S. F., Blase, K. A., Friedman, R. M., Wallace, F., et al. (2005). Implementation research: a synthesis of the literature. 2005. *Tampa, FL: University of South Florida, Louis de la Parte Florida Mental Health Institute, The National Implementation Research Network (FMHI Publication# 231)*, 119.

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*.

Gentaz, E., Sprenger-Charolles, L., Colé, P., Theurel, A., Gurgan, M., et al. (2013). Évaluation quantitative d'un entrainement à la lecture à grande échelle pour des enfants de cp scolarisés en réseaux d'éducation prioritaire: apports et limites. *Approche neuropsychologique des apprentissages chez l'enfant (ANAE)*, 123:172–181.

Gersten, R., Dimino, J., Jayanthi, M., Kim, J. S., and Santoro, L. E. (2010). Teacher study group: Impact of the professional development model on reading instruction and student outcomes in first grade classrooms. *American Educational Research Journal*, 47(3):694–739.

Gersten, R., Taylor, M. J., Keys, T. D., Rolfhus, E., and Newman-Gonchar, R. (2014).

Summary of research on the effectiveness of math professional development approaches. *National Center for Education Evaluation and Regional Assistance. Washington, DC.*

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.

Guskey, T. R. (2002). Professional development and teacher change. *Teachers and teaching*, 8(3):381–391.

Harris, D. N. and Sass, T. R. (2011). Teacher training, teacher quality and student achievement. *Journal of public economics*, 95(7-8):798–812.

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

Kealey, K. A., Peterson Jr, A. V., Gaul, M. A., and Dinh, K. T. (2000). Teacher training as a behavior change process: principles and results from a longitudinal study. *Health Education & Behavior*, 27(1):64–81.

Kerwin, J. T. and Thornton, R. L. (2021). Making the grade: The sensitivity of education program effectiveness to input choices and outcome measures. *Review of Economics and Statistics*, 103(2):251–264.

Kim, J. S., Olson, C. B., Scarcella, R., Kramer, J., Pearson, M., van Dyk, D., Collins, P., and Land, R. E. (2011). A randomized experiment of a cognitive strategies approach to text-based analytical writing for mainstreamed latino english language learners in grades 6 to 12. *Journal of Research on Educational Effectiveness*, 4(3):231–263.

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.

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.

Machin, S. and McNally, S. (2008). The literacy hour. *Journal of Public Economics*, 92(5-6):1441–1462.

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*.

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.

Parsad, B., Lewis, L., and Farris, E. (2001). *Teacher preparation and professional development, 2000*. US Department of Education, Office of Educational Research and Improvement.

Piper, B., Simmons Zuilkowski, S., Dubeck, M., Jepkemei, E., and King, S. J. (2018). Identifying the essential ingredients to literacy and numeracy improvement: Teacher professional development and coaching, student textbooks, and structured teachers' guides. *World Development*, 106:324–336.

Popova, A., Evans, D. K., Breeding, M. E., and Arancibia, V. (2022). Teacher professional development around the world: The gap between evidence and practice. *The World Bank Research Observer*, 37(1):107–136.

Randel, B., Beesley, A. D., Apthorp, H., Clark, T. F., Wang, X., Cicchinelli, L. F., and Williams, J. M. (2011). Classroom assessment for student learning: Impact on elementary school mathematics in the central region. final report. ncee 2011-4005. *National Center for Education Evaluation and Regional Assistance*.

Rimm-Kaufman, S. E., Larsen, R. A., Baroody, A. E., Curby, T. W., Ko, M., Thomas, J. B., Merritt, E. G., Abry, T., and DeCoster, J. (2014). Efficacy of the responsive classroom approach: Results from a 3-year, longitudinal randomized controlled trial. *American Educational Research Journal*, 51(3):567–603.

Ross, J. G., Luepker, R. V., Nelson, G. D., Saavedra, P., and Hubbard, B. M. (1991). Teenage health teaching modules: impact of teacher training on implementation and student outcomes. *Journal of School Health*, 61(1):31–35.

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.

Yoon, K. S., Duncan, T., Lee, S. W.-Y., Scarloss, B., and Shapley, K. L. (2007). Reviewing the evidence on how teacher professional development affects student achieve-

ment. Working Paper 033, Regional Educational Laboratory Southwest (NJ1).

Appendix A: Additional Tables

Table A1: Summary of the literature

Authors (Year)	Country	Design	topic	Grade	Sample	Intensity	ES	Sig.	Post-inter.
Angrist and Lavy (2001)	Israel	NE	m;r	prim	31/850	11 d	.3-.6	s	-
Borman et al. (2007)	USA	RCT	r	K-G2	35/4180	95h, 1y	.21-.33	s	s
Garet et al. (2008)	USA	RCT	r	prim	90/5000	108h, 2y	.03	ns	ns
Machin and McNally (2008)	UK	NE	r	G1	14000/1.6m	intens.	.09	s	-
Gersten et al. (2010)	USA	RCT	r	G1	81/468	20h, 6m	.23	ns	-
Kim et al. (2011)	USA	RCT	r	G6-12	103/2726	46 h	.05	ns	-
Randel et al. (2011)	USA	RCT	m	G4-5	67/4700	self	.10	ns	-
Allen et al. (2011)	USA	RCT	g	G4-5	76/2237	1y	.22	s	-
Harris and Sass (2011)	USA	NE	m;r	prim	1031/487000	NA	.00	ns	-
Campbell and Malkus (2011)	USA	RCT	m	G3-5	36/24759	very	.05	ns	ns
Newman et al. (2012)	USA	RCT	m;s	G4-8	82/3000	10 d, 2y	.05	s	-
Bos et al. (2012)	USA	RCT	r	G6-8	52/18180	14 days	.01	ns	-
Gentaz et al. (2013)	France	RCT	r	K	56/2398	42 h	.00	ns	-
Rimm-Kaufman et al. (2014)	USA	RCT	g	G3-5	24/2904	20 h	.00	ns	-
Sirinides et al. (2018)	USA	RCT	r	G1	9784	intens.	.47	s	-
Meyers et al. (2016)	USA	RCT	m;r	G7-8	60/3072	24 h, 2y	.13-.17	s	s
Bouguen (2016)	France	NE	r	K	118/4345	intens.	.15	s	-
Loyalka et al. (2019)	China	RCT	m	G7-9	300/16661	15 d	.00	ns	-

The table lists the rigorous (RCT or rigorous non-experimental methods) and recent teacher training studies (since 20'). Included studies are recent (post-2000), rigorous, and sufficiently powered (ex-ante MDE below .30 sd). We provide the reference to the study, the year of publication, the country, the type of design– RCT or Non-Experimental (NE)– the topic of the training program–, the topic –general (g), maths (m), reading (r), the grade, the sample size (clusters/individuals), the intensity (h, d, m y for hours, days, months and years), the effect size expressed in standard deviation and the significance (s for significant at 10%). The last column provides information on whether the results on students post-intervention is significant or not (students of teachers surveyed at least one year after their teacher has benefited from the teacher training program).

Table A2: 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 A3: 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 A4: 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 A5: 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 4	-0.055	-0.09	0.015
	[0.761]	[0.003]	[1]
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 A6: 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 A7: 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.

Appendix B: Robustness Checks on Main Results

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

			Treatment v. Control	
	Obs.	Control	(1)	(2)
Year 2				
Endline knowledge	2,427	0.742	0.090	0.071*
			(0.060)	(0.043)
			[0.680]	[0.419]
Endline skills	2,427	0.531	-0.000	0.002
			(0.051)	(0.037)
			[0.918]	[1.000]
Endline motivation	2,420	-0.050	-0.043	-0.022
			(0.043)	(0.040)
			[0.680]	[1.000]
Year 3				
Endline knowledge	2,473	0.512	0.031	0.021
			(0.063)	(0.049)
			[0.718]	[0.990]
Endline skills	2,473	0.380	-0.034	-0.015
			(0.055)	(0.045)
			[0.718]	[0.990]
Endline motivation	2,472	-0.053	-0.135***	-0.091**
			(0.046)	(0.039)
			[0.013]	[0.072]
Number of clusters			112	112
Controlling for baseline variables			N	Y

The table gives the impact of the program on student outcomes. The sample is restricted to classes that participated in both survey waves. Column *Obs.* gives the number of students surveyed, *Control* 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 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. In parentheses are standard errors clustered at the teacher level. In brackets are p-values robust to multiple testing. p<0.01 ***, p<0.05 ** p<0.1 *.

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

		Treatment v. Control		
	Obs.	Control	(1)	(2)
Year 2				
Endline knowledge	2,070	0.717	0.121*	0.086*
			(0.062)	(0.044)
			[0.194]	[0.193]
Endline skills	2,070	0.488	0.022	0.001
			(0.055)	(0.039)
			[0.864]	[1.000]
Endline motivation	2,067	-0.030	-0.035	-0.025
			(0.046)	(0.044)
			[0.800]	[1.000]
Year 3				
Endline knowledge	2,107	0.481	0.016	0.022
			(0.070)	(0.052)
			[1.000]	[0.835]
Endline skills	2,107	0.355	-0.017	0.021
			(0.065)	(0.051)
			[1.000]	[0.835]
Endline motivation	2,106	-0.012	-0.153***	-0.098**
			(0.049)	(0.043)
			[0.008]	[0.081]
Number of clusters			95	95
Controlling for baseline variables			N	Y

The table gives the impact of the program on student outcomes. The sample is restricted to teachers that participated in both teacher survey waves. Column *Obs.* gives the number of students surveyed, *Control* 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 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. In parentheses are standard errors clustered at the teacher level. In brackets are p-values robust to multiple testing. $p<0.01$ ***, $p<0.05$ ** $p<0.1$ *.

Table B3: Effects of Training Topics on Class Topics - Controlling for Baseline Covariates

	Year 1 Class topic		Year 2 Class topic		Year 3 Class topic	
	(1)	(2)	(3)	(4)	(5)	(6)
Training topic Year 1 x Treatment	0.325*** (0.081)		0.080 (0.082)		0.164* (0.083)	
Training topic Year 2 x Treatment		-0.041 (0.095)		0.205*** (0.076)		0.117 (0.102)
Training topic Year 1	-0.010 (0.066)		-0.020 (0.063)		-0.084 (0.058)	
Training topic Year 2		0.091 (0.079)		-0.045 (0.051)		-0.078 (0.079)
Treatment	-0.083* (0.042)	0.004 (0.040)	-0.100** (0.042)	-0.114*** (0.040)	-0.025 (0.045)	-0.004 (0.042)
Constant	0.289*** (0.055)	0.267*** (0.057)	0.245*** (0.053)	0.262*** (0.051)	0.452*** (0.061)	0.452*** (0.063)
	744	744	768	768	768	768

The 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 in the local training center. Each column is a different regression. The estimated coefficients are conditional on baseline hours of science taught and baseline practices of inquiry-based learning. 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.
$p<0.01$ ***, $p<0.05$ ** $p<0.1$ *

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

	Year 1 Class topic		Year 2 Class topic		Year 3 Class topic	
	(1)	(2)	(3)	(4)	(5)	(6)
Training topic Year 1 x Treatment	0.219*** (0.067)		0.147** (0.073)		0.129 (0.080)	
Training topic Year 2 x Treatment		0.032 (0.080)		0.254*** (0.068)		0.107 (0.100)
Training topic Year 1	0.045 (0.051)		-0.028 (0.053)		-0.058 (0.054)	
Training topic Year 2		0.056 (0.064)		-0.079* (0.047)		-0.076 (0.077)
Treatment	-0.022 (0.033)	0.025 (0.030)	-0.070* (0.035)	-0.080** (0.033)	-0.022 (0.039)	-0.009 (0.036)
Constant	0.545*** (0.023)	0.546*** (0.022)	0.554*** (0.026)	0.562*** (0.023)	0.546*** (0.030)	0.544*** (0.029)
	1032	1032	952	952	816	816

The 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 in the local training center. Each column is a different regression. 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. $p<0.01$ ***, $p<0.05$ ** $p<0.1$ *

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)	-0.046 (0.176)	101	-0.000	0.440** (0.202)	0.297 (0.208)	0.03
Normative statements	119	0.000	0.146 (0.180)	0.105 (0.195)	100	0.000	0.295 (0.208)	0.242 (0.231)	0.44
Science intensity									
Weekly hours	119	1.429	0.252*** (0.093)	0.180* (0.095)	102	1.349	0.252** (0.107)	0.186* (0.108)	0.95
Number of topics	119	4.420	-0.272 (0.267)	-0.467* (0.269)	102	4.260	0.075 (0.283)	0.038 (0.309)	0.06
% hands-on experiments	119	0.652	0.088** (0.044)	0.068 (0.046)	102	0.644	0.025 (0.049)	-0.004 (0.050)	0.17
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 *Obs.* gives the number of observations, *C.* 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[49]. We label the three sub-dimensions "I like science", "Scientific mindset" and "Science is easy".[50]

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.

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

[50]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. $p<0.01$ ***, $p<0.05$ ** $p<0.1$ *