Teacher Training, Teacher Practices and Student Performance in Science: Evidence from a Randomized Study in French Primary Schools *

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

April 8, 2022

Abstract

We conduct a randomized study to assess the impact of an in-service teacher training program on inquiry-based learning in science in French primary schools. The study comprises 134 randomly assigned volunteer teachers and two cohorts of about 2,500 students. In addition to student test scores, we collect information on teacher practices and pedagogical knowledge. We find that the training program (80 hours over two years) increases the weekly hours of science instruction as well as the number of science experiments conducted in class. Yet, one year later, most of these effects disappear. Consistently, while we find small effects on students' scientific knowledge during the last year of the training program, these effects are also short-lived. These results highlight the difficulty to train experienced teachers in a way that has deep and long-lasting influence on their practices.

JEL classification: I20

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

*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 under the grant ANR-13-APPR-0004-01. The experiment received the Paris School of Economics IRB approval IN/2005009, and was registered on the AEA registry under number AEARCTR-0001864.

[†]University of Mannheim, Germany. Suzanne Bellue is supported by the German Academic Exchange Service (DAAD).

[‡]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.

[§]CNRS-Paris School of Economics, France

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

[‖]Université Paul Valery Montpellier 3, France

Teacher quality explains a great deal of student performance. Consistent findings from the teacher effect literature show that a one standard deviation (SD) improvement in teacher quality increases student performance by 0.1-0.2 SD (Rivkin et al., 2005; Rockoff, 2004; Rothstein, 2010), an effect approximately equivalent to a costly reduction of about 5 to 10 students per class.[1] While it is still unclear whether these impacts are short-lived (Kane and Staiger, 2008; Carrell and West, 2010) or have long-lasting consequences (Chetty et al., 2011), a general consensus exists in both high- and low-income countries that improving teacher quality is key to improving educational outcomes (Filmer and Rogers, 2018). However, improving teacher quality through pre-service or in-service teacher training has proven difficult. In a recent review of the experimental literature, Fryer (2017) finds that among 21 rigorous and experimental professional development studies, only five show positive impacts. Similar disappointing results were found from longitudinal studies based in the US (Harris and Sass, 2011) or from a very recent large experimental study conducted in China (Loyalka et al., 2019).

The finding that teacher quality matters while their training level might not, alongside the fact that the essence of what makes a good teacher is hard to observe (Hanushek, 1971; Hanushek and Rivkin, 2006; Rockoff et al., 2011), has led some to conclude that teacher performance is innate, stable over time, and cannot be modified. As Baumol et al. (2012) memorably puts it, education may be suffering from a cost disease that structurally restricts teacher productivity gains. Nevertheless, as mentioned by Fryer and described in Table A1, the literature also includes resounding success stories. Consistent evidence shows, for instance, that a training program based on phonological awareness, such as the Reading Recovery program, improves first graders' reading skills. These results, based on small and local experiments (Center et al., 1995; Schwartz, 2005), were recently confirmed by a large experimental study conducted among 7,000 students (Sirinides et al., 2018). Similarly, again focused on reading skills, the program Success for All has been shown to increase reading performances as well (Borman et al., 2007). As both programs include remedial education components[2], it remains unclear

[1] In a recent review of the class size literature, (Bouguen et al., 2017) find that a one-student reduction in class size increases performance by about 2pp of a SD

[2] For instance, Reading Recovery includes a 30-minute reading session targeted to low achievers in the class. Similarly, Success for All includes a comprehensive school-level reform, with changes in curriculum and classroom structure. As a result, both programs are fairly expensive in comparison with regular teacher training programs. Rightly so, Fryer (2017) regroups both programs into the sub-category "Managed Professional Development" and treats them separately in the meta-analysis. Overall, he finds that "Managed Professional Development" is effective at increasing students performance.

whether the teacher training programs *per se* would have resulted in positive results. Yet, other results based on non-experimental variation (Bouguen, 2016; Machin and McNally, 2008; Angrist and Lavy, 2001) tend to confirm that well-executed and intensive training programs, based on validated pedagogical approaches, can efficiently modify teaching performance.

In this paper, we design a large-scale experimental study, implemented in an ecological environment, to test the impact of an intensive teacher training in a specific pedagogical approach – the inquiry-based method – designed to improve primary school students' science learning. We randomize 134 volunteer French teachers, located in three regional school districts, and two cohorts of their students into a treatment and a control group. To fully track how the intervention impacts teachers, we collect teacher-level measures of their practices in science at the end of the training program and one year after the end of the intervention, so as to capture short and midterm impacts. We then measure how these potential teacher effects translate into better student scientific skills, knowledge and motivation on the two successive cohorts of about 2,500 students each.

The program is provided by a French non-for-profit organization (*La Main à la Pâte*), sponsored by the Academy of Science, and is implemented through a network of local training centers (*Maisons pour la Science* later referred to as *Maisons*). It is designed for primary school teachers in grades three to five. French primary school teachers follow a class of students throughout the school year and are entitled to teach all subjects, including two hours of science per week. Eligible teachers in the treatment group received on average 66 effective hours of science training sessions over two school years. The year before implementation, the teachers in our sample received on average only two hours of scientific training per year.

The program is directly inspired by the inquiry-based method, a pedagogical approach that emphasizes students' active role in the learning process. Inquiry-based learning in science involves conducting experiments in which students actively participate in defining the research questions, the scientific problems, and the hypothesis, and where students are involved in finding solutions. Students are encouraged to design their own hands-on science experiments that they conduct in class to test a set of scientific hypotheses. The inquiry-based learning method is considered as one of the best approaches to teaching science and has long been endorsed by the National Research Council in the U.S. (Council et al., 2000). It is expected to increase student

engagement in class, motivation in the discipline, and eventually scientific skills and knowledge.

Since Bruner (1961), the inquiry-based approach has become popular and has been the subject of an abundant theoretical literature, but the benefit of training teachers using this approach has rarely been rigorously evaluated at scale. To the best of our knowledge, among the 18 rigorous studies we identified in the literature (see Table A1[3]), only two studies specifically analyze the impact of an inquiry-based teacher training program on either science, technology or mathematics (STEM). In the first study, conducted in Alabama, Newman et al. (2012) find that a fairly intensive inquiry-based training program[4] has small but significant impacts on mathematical skills (+0.05 SD) but no significant impacts on scientific skills. The study could only capture the impacts after one year out of the two years of intervention because the randomization was compromised in the second year.[5] The second study, in Missouri (Meyers et al., 2016), finds that a very intensive program (240 hours of training over two years) impacts student performance in mathematics three year after the beginning of the program (impacts are not significantly different from zero during the first two years of intervention). The program is comprehensive (at the school level), very intensive and relies heavily on the inquiry-based principles. The study targets G7 and G8 students and finds impacts between 0.13 SD and 0.18 SD on mathematics skills. Scientific skills were not assessed. Since Newman et al. (2012) is the only inquiry-based teacher program that measures impact on scientific skills, it is the only study that can be directly compared to ours.

Our paper contributes to this fairly restricted literature by testing in an ecological context the impact of a training program on inquiry-based teaching on both teacher and student performance. First, conversely to most existing literature (see Table A1's last two columns), our design measures impacts at the end of two years of training (year 2) but also analyzes whether the trained teachers affect student performance one year after

[3]In Table A1, we selected articles that are either experimental or based on empirically rigorous methods, that have reasonable statistical power (i.e. with ex-ante minimum detectable effect below 0.3 SD) and which have been published in the last 20 years. We did not conduct a systematic review of the literature but based our review on Fryer (2017) and Yoon et al. (2007), which are the two most recent systematic reviews of the literature. We added to this literature a few additional articles based on non-experimental variations (Bouguen, 2016; Machin and McNally, 2008; Harris and Sass, 2011; Angrist and Lavy, 2001) and a recent social experiment conducted in China (Loyalka et al., 2019)

[4]The intervention is fairly similar to ours and includes 10 days of teacher training during the summer, and several follow-up training sessions during two years. The evaluation analyzes the impact on G4 to G8 on math, science and technology.

[5]Using non-experimental variation, the authors suggest that the effects are likely to be around +0.10 SD at the end of the intervention.

the end of the intervention (i.e. on a new cohort of students in year 3). We believe that measuring impacts over two years, and specifically one year after the end of the training program, is crucial to analyze the impact of teacher training programs as the overall effectiveness of a teacher program crucially depends on whether or not the teaching practices learned during training are preserved throughout the teacher's career. If teachers do modify their practices permanently, teacher training programs have the potential to affect several generations of students and therefore to profoundly improve the performance of education systems. Surprisingly, very few articles specifically cover this issue. As shown in Table A1, a handful of studies measures student long-term effects (i.e. the effect on student performance of having benefited from a trained teacher one year prior) and only two measure the teacher long-term impacts (i.e. the effect of a trained teacher on a new cohort of student).[6] Our article contributes to this important and under-studied dimension of teacher training programs.

Second, the intensity of the program we investigate (80 scheduled hours over two years, 66 hours effectively completed in the treatment group[7]) is 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). Besides, 66 effective hours of in-service training in science constitutes a very significant increase when compared to the amount primary school teachers usually receive in France, which is about one hour of training in science per year.[8] Our design, based on 134 clusters, two cohorts of 2,500 students and an almost perfect adherence to the randomized protocol, is also more precise than most equivalent studies in the literature.[9]

Third, in addition to measuring the impact of the program on student perfor-

[6]For instance, Newman et al. (2012) do not have effect after the first year. Similarly, Meyers et al. (2016) measures impacts after three years of intervention but not the school year after the training program is over. Similarly, in the rest of the literature, most articles measure the effect of a training program during its implementation but do not measure how the trained teachers fare after the program is over.

[7]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), to absenteeism and to maybe a recall bias.

[8]Some of the teachers we surveyed and who did not participate in the experiment but worked in the same schools reported receiving an average of about one hour of training per year before the experiment (see Table 1) to be compared to 78 hours of training for the treatment teachers and 11 hours for the control teachers during the experiment over two years (see Table 3).

[9]According to our calculation based on our selected studies listed in Table A1, our study is much more precise than most of the available experimental studies on the subject. For instance, only Loyalka et al. (2019) has a larger number of clusters than we do (see Table A1) but they do not measure teacher practices.

mance, we design student and teacher surveys to capture the different aspects of the inquiry-based method. Every year we collect teacher level measures to capture whether their science teaching practices have been modified, both in terms of the scientific intensity (e.g. number of hours of science) but also in terms of the practices teachers mobilize (e.g. are inquiry-based principles understood and implemented?). Our survey also precisely documents the content of the intervention, its intensity, the level of teacher satisfaction. Our measures at the student level includes, in addition to scientific knowledge and skills, students' motivation in science, a mechanism that is key to the inquiry-based method but has never been precisely measured in previous work. These measures at the student and teacher level, conducted every year, help establish a clearer causal route and better explain the constraints that often make teacher training programs ineffective.

Our results first highlight the inherent difficulties associated with conducting a training program in an ecological context. In this experiment, trained teachers have been replaced in class by substitute teachers during their training time, so that there is no direct negative effect on students. In spite of this, school district managers faced difficulties in convincing teachers to commit to a two-year program, and attracting teachers with a lower level of scientific experience. Our group of 134 volunteer teachers is, as a result, unsurprisingly different from the average teachers in similar schools. Compared to non-volunteer peer teachers working at the same schools, our teacher are more likely to hold a scientific degree, are older, are more experienced, have received more pre-intervention in-service training, and teach more science in class than their peer counterparts. In our experimental context, which closely mimics the context in which a typical school district manager would implement such program, it is challenging to target in-service training programs to teachers who lack scientific background, have little experience teaching science, or have low interest in science.

Second, we find that treatment teachers almost perfectly adhered to the training program. Treatment teachers said they had received an average of 78 hours of in-service training over the two school years of the program (against 11 hours in the control), including 66 hours directly provided by the *Maisons*, fairly close to the objective of 80 hours. The experiment therefore generated a 67-hour net increase in in-service training (+61 hours by the *Maisons*). By contrast, the year before the program started, volunteer teachers said they had received about 11 hours of training overall and only two hours of training specifically about science. This program therefore constitutes a

very significant increase in the regular number of hours of in-service training provided to teachers in France.

Third, we find evidence that the training program affected teachers' reported practices. At the end of the two-year training program (year 2), treatment teachers outperform control teachers by about 0.5 SD on our science intensity index of teacher practices, which aggregates quantitative measures such as number of hours of science taught per week or number of science experiments conducted in class. For instance, treated teachers reported teaching more hours of science per week (+0.18 h/week) and conducting more hands-on scientific experiments (+7 percentage points) in year 2. The effects are smaller in year 3 (one year after the end of the training program) and mostly not significantly different from zero, suggesting that teaching practices fade out quickly after the end of the training program. We also find that the science topics covered during the training affected the topics covered in class during the last year of the training (year 2) and, consistently with the impacts on teacher practices, this connection fades out in the subsequent year, when the training is over (year 3). Effects on our other measures of teacher practices (declared inquiry-based practices and normative statement about inquiry-based practices) are positive but insignificantly different from zero suggesting that while teachers may have increased the intensity of science content in year 2, the trained teachers did not modify their pedagogical approach enough to have deep and long-lasting effects on their practices.

Fourth, our measures at the student level show significant impacts on scientific knowledge (+0.1 SD) in year 2 (at the end of two-year training program), but no impacts on scientific skills or motivation. In year 3, effects on skills and knowledge are insignificant and the effect on motivation is even negative (-0.1 SD).These student level results suggest that the quality of teaching declined in year 3 when the teachers no longer benefited from the support of the training centers.

We provide two complementary interpretations to the lack of results on student performance and their apparent decline in year 3. First, the impact on teachers' practices, except for the effect on intensity in year 2, is not strong enough to generate visible effects on students' skills. This interpretation is consistent with the size of the correlations that we observe in the control group between teachers practices and student performance. Taken at face value, these correlations indicate that only a large transformation of teacher practices can lead to significant impacts on students. This interpretation is compatible with the literature that often fails to demonstrate positive

impacts on students, even when the amount of training is relatively large, as it is the case in this study. The second interpretation is that, while some of the teacher practices were effectively affected during training years, through the direct support and in-class intervention of the program, once the training was over, teacher practices became less effective at improving student performance, resulting in a decline of the teaching quality in year 3. The negative effect on motivation in year 3 may even suggests that, when poorly implemented, inquiry-based pedagogy reduces student interest in science. This interpretation is compatible with the results found by a research team in didactics of science that analyzed the training sessions qualitatively (Chesnais et al., 2017; Munier et al., 2021). Since the cost-effectiveness of teacher training programs crucially depend on whether teacher practices are persistent over time, our findings highlight the difficulty to enhance teacher quality in the long-run via teacher training programs alone. This hasn't been clearly documented before.

In the rest of the article, we will describe the teacher training program conducted by the *Maisons pour la Science*, present the experimental setting, data and compliance, and then discuss our results. We conclude with a discussion that highlights the challenges faced by education systems to improve education quality.

1 Training Program: background and content

This article studies a teacher training program delivered by the *Maisons pour la Science*, referred to as *Maisons*. The objective of the program is to train teachers in the inquiry-based learning method and to improve their students' scientific skills, knowledge and motivation.

1.1 Teaching Science in French Primary schools

In France, primary school covers first to fifth Grade and caters to children from age six to eleven. In the vast majority of schools, primary school teachers are responsible for one grade and therefore teach all subjects. Teachers' initial and in-service training in science varies greatly depending on the age of the teacher, her initial education (scientific or not) and her pedagogical training. We provide more information about the characteristics of the teachers included our sample in Tables 1 and 3 below. Suffice here to say that primary school teacher's initial education typically includes one undergrad degree and one teacher's initial certification, obtained in national certification/training

centers.[10] 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.[11]

Teachers are expected to follow a fairly precise national curriculum. The curriculum specifies which topics should be covered, which competence should be mastered by the end of each school year and also how teachers should allocate their teaching time. For instance, in third, forth and fifth grades, the curriculum provides that teachers should devote two hours per week to science and technology. The curriculum also lists eight topics that teachers can cover in class.[12] Teachers are free to use the pedagogical method they see fit.[13] In class, scientific experiments and the inquiry-based method is recommended in the curriculum. Yet, the general consensus among trainers in science is that French primary school teachers do not have sufficient training in science, rarely rely on inquiry-based methods and do not conduct many experiments in class. Besides, while the curriculum provides that teachers should devote two hours per week to science, our data shows that regular teachers report on average 1.2 hours of science per week.[14] The training program we analyse in this article is designed to provide teachers with all the tools and competences to implement the inquiry-based learning method in class.

1.2 Background of the *Maisons pour la Science* project

The *Maisons pour la Science* project 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*, a very influential and experienced actor in the field of scientific awareness at school, received a grant to support a vast project aiming at improving scientific knowledge and motivation at school. The project mainly revolves around training programs offered to primary and secondary school teachers. The pro-

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

[11] Again depending of the age of the teacher this varies greatly.

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

[13] This is the principle of "pedagogical freedom".

[14] This is pre-experiment from data collected among teachers who did not volunteer to participate to the experiment (see Table 1). Professional trainers anecdotally confirmed to us that very few primary teachers respect the two-hours recommendation.

gram consists of establishing local training centers, called *Maisons pour la Science*, in four regional school districts. Established within local universities, the *Maisons* then designed an intensive training program on inquiry-based learning targeted toward primary school teachers. This training program is the object of this research.

1.3 The *Maisons*' training program

The *Maisons*' training program is available to G3, G4 and G5 primary school teachers and is inspired by the inquiry-based learning method, which emphasizes students' role in the learning process. Inquiry-based learning involves conducting hands-on experiments in which students actively participate in defining the research questions, the scientific problem, and the hypotheses, and are directly involved in finding the solutions. The objective of the training program is to help teachers design their own scientific learning sequence and to facilitate the implementation of in-class experiments by supplying ideas, material and providing visits of a teacher aid. The program is composed of 80 hours of training and lasts two consecutive school years. The 80 hours include training sessions at the *Maisons*, discussion with other teachers, attendance at scientific conferences, and in-class educational support. Each session covers one of the eight possible topics of the G3-G5 curriculum: the training program therefore closely follows the national teaching requirements.

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 by three different kinds of contributors: professional trainers from the national certification and training centers, field trainers who observed and assisted the teachers during the science sequences in her own classrooms, and scientists who gave lectures on a specific topic of the curriculum. 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 that was designed at the training centers, with the support of a field trainer.

Importantly, the training sessions conducted at the training centers occurred during normal teaching hours. The school district managers appointed a substitute teacher, ensuring that the program did not reduce regular 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.

2 Experimental Setting, Data and Compliance

2.1 Sample, Randomization and Surveys

Although the intervention was conducted in four regional school districts (*Académies*), only three took part in the experiment: Auvergne, Lorraine and Midi-Pyrénées.[15] Within the three regional school districts, the Maisons worked in close relationship with the school district managers who authorized the training program and were responsible for recruiting teachers and finding substitute teachers. The objective of the recruitment was to identify teachers who would be willing to attend the training, who might benefit from it, and who could be substituted for during the training hours. The initial aim was to identify teachers who lack training in science (i.e., teachers with no prior scientific degree, who expressed difficulty designing scientific sequences, etc). Given the intensity of the program (80 hours over two years), teachers who were contacted by their district manager had to voluntary enroll. As a result, the voluntary teachers are likely to have an inclination for science or for the inquiry-based method (see section 2.2 below). Their students, however, did not self-select into the program and are therefore likely to be similar to other students in the same districts (see section 2.3 below).

To measure the effect of the intervention on both the teachers and their students, we used the teacher as the unit of randomization and tracked volunteer teachers during three years: before the beginning of the intervention when teachers applied to the program in 2014, during the two years of the intervention (year 1 and year 2) and one year after the end of the treatment (year 3). As described in Figure 1, *volunteer teachers* expressed their willingness to participate to the experiment by filling out an online questionnaire (Q0 survey) in 2014. The survey includes basic socio-economic data (gender, education, experience), tracking information (name, phone number, school name) and questions about teaching practices and exposure to science (number of hours of science teaching, number of hours of teacher training, etc.). We then randomly assigned volunteer teachers to the treatment or control group, and the teachers assigned

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

to the treatment group started the training program the following year (year 1).[16]

We randomized teachers within eight local school districts (*Département*) located in the three regional school districts. Since 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 (5), we stratified the randomization using baseline teaching experience. In some cases (three school districts), we additionally used both experience and an index of teaching practices. In the districts where several teachers from the same school applied to the program, we stratified at the school level. 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 treatment. We account for this in regressions by including strata fixed effects and sampling weights proportional to the inverse of the assignment probability.

The original sample is composed of 134 teachers across three regions and nine local school districts. In addition to teachers in treatment and control groups, we surveyed a group of peer teachers; these peers are non-volunteer teachers, working in the same schools as the volunteer teachers and teaching in G3, G4 or G5. We use peer teachers to document how volunteer teachers (self-)selected themselves into the program.

After completing the Q0 survey, treatment teachers started the first year of the training program. We started measuring the impact of the program during year 2, i.e., during the second year of the intervention. As shown in Figure 1, we surveyed teachers (and their students) at the beginning of the second year of training (Q1 survey), at the end of the same school year (Q2 survey) and at the beginning (Q3) and the end (Q4) of the following school year (year 3).[17] The Q2 survey therefore captures the effect of the program immediately after having completed two years of training while the Q4 survey measures the "mid-term" impacts as it measures the effect one year after the end of the program. Note that students in year 2 and in year 3 are generally different students. Indeed, since teachers do not usually change grades between two consecutive school years, the students surveyed in the first school year are generally different from the ones surveyed in second and the third school years.[18] Consequently, our dataset is

[16]Due to implementation difficulties, teachers who belonged to one local school district started their training one year later in 2015/2016 and were surveyed in 2016/2017 and 2017/2018.

[17]In this paper we do not use the teacher questionnaire in Q3 but only the Q3 student test scores.

[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.

a panel at the teacher level but a repeated cross-section at the student level.

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 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 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 school year 2016-2017 (Q3): since we do not rely on this survey in this paper we do not report it here.

2.2 Teacher Sample Description and Selection

Using data from the Q0 and Q1 survey among volunteer and peer teachers respectively, we describe the sample of volunteer teachers and how they compare with their peers.[19] We present our results in Table 1. In the first panel of Table 1, we compare the treatment and control teachers at baseline and the volunteer teachers (treatment or control) with the peer teachers in the second panel.

Several important features characterize the group of volunteer teachers. First, in terms of socio-economics characteristics and compared to the national average (DEPP, 2018), volunteer teachers are less likely to be female (74% against 81.6%) and are older (45 years old against 42). Volunteer teachers are also strikingly different from their peers in terms of observable characteristics (panel "Volunteer vs Peer"). Volunteer 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). Prior to applying for the *Maisons*' training, volunteer teachers also report teaching more science (+0.82 hours or +66%). In addition, prior to randomization, volunteer teachers are more likely to have benefited from any in-service training (+17 pp), have completed more hours (+10 hours) and are more trained in sciences (+3.2 hours). They also were more likely to have been trained by the *Maisons*[20] or from other *La Main à la pâte* interventions before the beginning of the Maisons' training program.[21]

Overall, teachers who applied for training in the *Maisons* are better trained in science, more exposed to in-service training pre-intervention, and more used to practicing science in class. These findings are particularly meaningful to interpret our results: a very intensive teacher training program is likely to attract 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 limitation, their effort to attract teachers with a lower level of scientific awareness mostly failed. More generally, we believe that this lack of targeting reflects an important conundrum for teacher training: targeting the program to teachers who need it the most is challenging and may be one of the under-studied reasons why teacher training programs are generally

[19] Peer teachers were not interviewed in Q0, as only volunteer teachers registered. They answer some questions that belong to the Q0 questionnaire in Q1.

[20] 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 incomparable with the the training program we study in this experiment.

[21] The *Main à la pâte* foundation has other interventions about science in primary schools.

ineffective at improving student performance.

Anecdotally, even though volunteer teachers are more exposed to in-service training than their peer counterparts, the average number of training hours they declared receiving per year remains relatively small (11 hours per school year), especially when compared with the number of hours of training provided by our intervention (40 hours per school year). The comparison between our training intervention and the usual number of hours of training received by volunteer teachers is even more striking when looking at training in science: volunteer teachers typically receive about 2 hours of training in science per year. The intervention therefore constitutes a significant increase in in-service training exposure, even for volunteer teachers who already benefit from more training than their peers.

Still using Table 1, we reject the null hypothesis that the treatment and control teachers are different at baseline in only one case out of 15 regressions conducted. Using the multi-hypothesis testing step-up method developed by Benjamini and Hochberg (1995), we find that none of the coefficients is significant at 10%. The minimum value to reject at least one of the 15 outcomes tested is 59%.[22] Yet, we observe that the treatment group had declared teaching science slightly more than the control group at baseline (one year before the intervention started). This is a source of concern, since this is one of the intermediary expected outcome. As a robustness test, we will add baseline hours of science teaching to our regressions in forthcoming analysis.

[22]In other words, the minimum Q-value is 59% for these 15 coefficients, far from standard significant levels.

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	118	11.179	-1.105	199	3.460	10.32*
			(6.702)			(5.852)
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 at 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 *volunteer* 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 *

Last, in Table 2, we investigate whether attrition affects our measurements at the teacher level. Teacher attrition increases with time, from 1.3% in Q1 to 24.6% in Q4 among control teachers. Attrition in the treatment group follows a similar pattern, with no significant differential attrition, although treatment teachers respond slightly more to our questionnaires. Teacher attrition is due to three different factors: (i) teachers not teaching science any more, (ii) teachers not teaching in G3-G5, (iii) teachers refusing to answer. Very few teachers refused to answer, three in year 2 (Q2) and five in year 3 (Q4). As mentioned in section 1.1, in France primary school teachers usually teach all subjects. Only two teachers in year 2 (zero in treatment) and seven in year 3 (five in treatment) did not teach science anymore. The bulk of the attrition is actually caused by teachers not teaching G3-G5 anymore : their number went from 10 in Q2 to 20 in Q3. Overall, treatment teachers have a slight tendency to stay in G3-G5 longer compared to control teachers, therefore explaining the slight (non-significant) differential attrition between treatment and control teachers. Nevertheless, attrition rates in Q2 and Q4 are high and may have distorted the sample. We look at this issue in Appendix Table A6 where we compare the baseline characteristics of the treatment and the control group on teachers who responded at Q1 (year 1), Q2 (year 2) and Q4 (year 3). As already noticed in the full sample (Table 1), on these sub-samples of respondents, treatment teachers declared teaching slightly more hours of science than control teachers at baseline. In addition, in year 2 only, responding treatment teachers are significantly more likely to have used inquiry-based learning than control teachers at baseline. We therefore add the baseline variable "practices inquiry-based" as a control variable in forthcoming analysis.

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 rounds and at the student level in year 2 and year 3. Column Obs. gives the number of observation, 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 *

2.3 Student Sample Description and Selection

Using data from the Q1 survey (student baseline in year 2) and Q3 survey (student baseline in year 3), we describe the sample of students in year 2 and in year 3. As mentioned, because teachers generally teach the same grade year after year, the student questionnaires were administered to two different cohorts of students. We present the results of the baseline questionnaire in year 2 (first cohort of students) and in year 3 (second cohort of students) in Appendix Table A3. Note that by design, students are not directly randomly allocated to control and treatment groups, only teachers are (see section 2.1 above). This design generates the risk that, post randomization, treatment teachers were assigned to better or worse students than control teachers. In Table A3 in the Appendix, we verify that this is not the case by comparing results at the beginning of year 2 and 3 (baseline) between students of control and treatment teachers. All indices are balanced at both baselines, i.e., at the beginning of each school year. Anecdotally, although volunteer and peer teachers were significantly different from each other (cf. Table 1), both teach relatively similar students, indicating no selection at the student level. While volunteer teachers are very specific teachers, the cohorts of students who took the tests are likely comparable to the average students in similar schools and districts.

We also use the second panel of Table 2 to look at attrition to endline surveys at the student level. We first look at whether the class of the volunteer teachers responded to the student survey (class attrition): a class did not respond to the survey if the teacher refused (happened only once) or did not teach in G3-G5 anymore. Class attrition is caused by teachers but could affect the validity of the student assessments as well if, for instance, treatment teachers, who do not teach G3-G5 anymore, would have taught to better students than their control counterfactual. Attrition in the control group increases with time as does teacher attrition. In year 2, we could not survey the students of 10.6% of control teachers, and in year 3, this fraction reached a fairly substantial 18.5%. Not surprisingly, class attrition is slightly lower in the treatment group but not significantly so. Table 2 finally provides attrition level at the student level i.e. students not filling out endline student tests. Attrition at the student level is fairly small (8.3 %) and is not differential.

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 *

2.4 Exposure to Training

Using data from the Q1 and Q2 teacher surveys, we present the exposure to the training program provided by the *Maisons* in Table 3. Being assigned to the treatment group significantly increases the teacher's probability of being enrolled into 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 year and about 66 hours offered by the *Maisons*, close to the objective of offering 80 hours.[23]

Importantly, 15.5% of control teachers report having benefited from a training conducted by a *Maison* although they were not supposed to receive any training from them. One of the *Maisons* authorized a few control teachers to attend some hours of training sessions (for instance some of the lectures offered by scientists, see Section 1.3). Yet, these control teachers did not receive the full 80 hours of intervention: they only received about four hours of training from the *Maisons* while they would have received 80*0.155 = 12 hours if they had benefited from the full training program.[24] These treated control teachers therefore received a relatively weak treatment, incomparable with the intensity received by treatment teachers. The fact that our intervention cannot be characterized by a simple dichotomous treatment variable (treated or not) has one important consequence for the estimation, however: We prefer not to provide a local average treatment effect of the program and restrict our analysis to ITT estimates.[25]

Last, we look at whether enrolling into the *Maisons* training program had spill-

[23] 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.

[24] This number is consistent with the monitoring data that we obtained from the *Maisons*.

[25] Alternatively, we could have calculated the LATE using the differential number of training hours as a first stage but this would have meant expressing our estimates in units of training hours which is both hard to interpret and incomparable with the available literature.

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 patterns of substitution. 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 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.

Overall, take-up of the training provided by the *Maisons* is high: Treatment teachers received 66 hours of training from the *Maisons*, while control teachers received less than five hours of training. Besides, the intervention did not act as a substitute for other forms of training. Finally, in results not presented here, treatment teachers expressed a high degree of satisfaction with the program: 86% were rather or very satisfied after the first year, and 87% after the second year. They expressed satisfaction with all the aspects of the training: in-class visits (95% are satisfied), on-site training sessions (92% are satisfied), and group work (87% are satisfied).

3 Results for Teachers

3.1 Measures of teacher outcomes

In this section, we leverage the rich data obtained from teacher questionnaires covering three years: first training year (year 1), second training year (year 2) and post-training year (year 3). Each year, we first measure teacher's self-declared number of hours of science taught per week, the number of scientific experiments conducted in class, whether these experiments were hands-on or not[26] and whether they feel they teach enough science. We aggregate each of these items into one index by standardizing each item and taking its average. We call it the *Science intensity* index. In years 2 and 3, we also administered two additional surveys. The first one lists five dimensions related to inquiry-based learning: introducing a scientific problem, formulating a hypothesis,

[26]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.

linking models and observations, framing student's vision, and evaluating students. For each dimension, through sub-items, we ask teachers if they implemented them in class. The second survey uses the same five dimensions, with subset items, but asks teachers to provide normative statements about their importance for teaching science.[27] For each dimension in each of the two surveys, we test the consistency of the sub-items using Cronbach alphas and keep only the dimensions that are consistent. With those, we aggregate three of the five practice dimensions into a *Declared practices* index; and we aggregate four of the five normative dimensions into a *Normative statements* index.[28] Arguably, the *Science intensity* index is the most objective and quantitative measure of practices, whereas the *Normative statements* index is the most subjective in nature.

To better understand how the training affected practices, we analyse, in the following, the relationship between the topics covered during the training and those covered in class. We then analyze the impact of the training on our measures of reported practices.

3.2 Training effects on topics covered

Before analyzing teacher practices *per se*, we first look at the relationship between science topics covered during the training sessions and science topics covered in class. While establishing such relationships does not imply that the intervention influenced teachers' pedagogical practices, it does show that the training had a first step influence on the way teachers designed their science sequence in class.

In France, the primary school science curriculum 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

[27] 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.*

[28] Specifically, we keep only the dimensions for which the items have a Cronbach alpha above 0.7. As a result, for the *Declared practices* index, we are left with "introducing scientific problems", "framing students' vision" and "evaluating students". For the *Normative statements* index, we only keep the teachers' normative statements on "importance of introducing a scientific problem", "formulating a hypothesis", "linking model and observation", and "evaluating students" (see Table A5). We also submitted a questionnaire eliciting teacher's vision of science at the end of the third year. The treatment did not affect this at all, and answers do not correlate with student performance, so we do not present this data here.

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 of the training centers within each regional school district (the three regions are then divided into nine local school districts).[29] The sample is therefore composed of 15 local district-year observations that generate variation in the topics covered during training.

In our teacher surveys, we list all possible topics and ask each teacher to list the ones that 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 in the following years).

We define W_{jpt} as a dummy variable that takes value 1 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 defined for the control teachers as well: 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, and 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 treated teachers simply 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.[30]

Table 4 gives the results of the above regression. Columns (1) and (3) show same-year relationships, whereas the other columns verify whether training topics from a year influenced the class topics in subsequent years. The constant coefficient in Column (1) of Table 4 indicates that slightly more than half of the existing topics were covered in class in year one, on average, in the control group. The interaction terms (β_2) indicates

[29] 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.

[30] Of note, in this analysis, we have a maximum of 134 teachers * 8 topics = 1072 data points which mechanically reduces our statistical power in comparison to our level of precision at the student level.

that this probability increases by 22 percentage points in the treatment group for those topics that have been covered during training that same year. As expected, the training topics covered during the training do not affect the topics covered in class in the control group (training topic coefficient is close to 0). Similarly, in column (3), topics covered during the second training year also increase by 25 percentage points the likelihood that they will be covered in class that same year. Interestingly, in year 2, the negative coefficient on the treatment dummy also indicates that fewer different topics were covered in the treatment group. In Appendix Table A4 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, although not in years 1 or 3.

The findings so far are a very clear indication that participating in the training influenced the teachers: they used the material developed during the training and translated it into their own science sequences in class. This result is explained by the fact that preparing and implementing the sequence in class was an aspect of the training itself, for which they could receive trainer assistance in their class. We also see that in year 2, teachers concentrated their effort on a lower number of activities, but likely with more preparation and a stronger focus on hands-on science experiments.

The relationship between topics in the training and in class seems to fade out rapidly. Training topics from year 1 do not influence class topics in year 2 (column (2)). In year 3, the training is over, and training topics covered in years 1 or 2 are covered in class more often in the treatment group by only 13 and 10 percentage points respectively (not significantly so). This finding is compatible with a model where training affects teacher practices only in the short run. Teachers adopt training material when trained, but do not use it much afterward. Given the high level of reported satisfaction with the training, this decay is apparently not caused by teacher dissatisfaction with the material provided to them. It is more likely that settled practices are difficult to change in the medium run. An additional possibility is that some of the activities presented during the training are difficult to implement in class without the support of the training center, as they often imply specific material or frequent outdoor activities (such as activities on wood or biodiversity by the river).

Table 4: Effects of training topics on class topics

	Year 1	Year 2		Year 3	
	Class topic	Class topic		Class topic	
	(1)	(2)	(3)	(4)	(5)
Training topic Year 1 × Treatment	0.219***	0.147**		0.129	
	(0.067)	(0.073)		(0.08)	
Training topic Year 2 × Treatment			0.254***		0.107
			(0.068)		(0.1)
Training topic Year 1	0.045	-0.028		-0.058	
	(0.051)	(0.053)		(0.054)	
Training topic Year 2			-0.079*		-0.076
			(0.047)		(0.077)
Treatment	-0.022	-0.07*	-0.08**	-0.022	-0.009
	(0.033)	(0.035)	(0.033)	(0.039)	(0.036)
Constant	0.545***	0.554***	0.562***	0.546***	0.544***
	(0.023)	(0.026)	(0.023)	(0.03)	(0.029)
	1,032	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 parenthesis. $p<0.01$ ***, $p<0.05$ ** $p<0.1$ *

3.3 Training effects on reported practices

In this section, we estimate the impact of the training program on teacher practices. For any measure y, we run an OLS regression, with robust standard errors. All observations are weighted by randomization probabilities and regressions include strata fixed effects. We run the following intention-to-treat[31] regressions separately per year:

$$y_j = \alpha_0 + \alpha_1 T_j + \boldsymbol{X}_j \boldsymbol{\alpha_2} + \nu_j$$

where j indices teachers, T is the assignment status of teacher j, and $\boldsymbol{X}$ a set of control including strata fixed effect. 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. Note that while the experiment has good statistical power at the student level (see section below), our analysis is based on 134 data points at the teacher level and therefore suffers from limited statistical power.

We separate the analysis of practices into our three indices (intensity, declared practices and normative statements), the exact composition of each being detailed in Appendix Table A5. Given our limited statistical power at the teacher level, our three aggregate indices leverage the fact that, although few dimensions show significant effects independently, they generally point to the same positive direction. Column (1) only includes strata fixed effects as control. To account for the fact that treatment and control teachers are imbalanced with respect to their hours of science taught reported at baseline (see Table 1) and practicing inquiry-based in year 2 (on respondent teachers in year 2 as shown in Table A6), we include both variables as a control in column (2).

The training program affects the *Science intensity index* every year, more strongly and significantly so in years 2 when the impact reaches almost 50% of a standard deviation. The effects in years 1 and 3 are significant only without controlling for baseline imbalance and are in any case of lower magnitude. In year 2, when the training is entirely completed, the effect size is large (around 0.65 SD) even after controlling for baseline initial imbalance (0.49 SD). Yet, these impacts are significantly smaller than the ones found in the literature (around 0.7 and 1.4 SD) when positive impacts on students are reported (see infra Section 5). Appendix Table A5 shows that the impacts

[31] We restrict our analysis to ITT estimates because we do not have a dichotomous treatment variable as explained in section 2.3 above.

Table 5: Impacts on Teacher Practice Indices

	Obs.	(1)	(2)
Year 1			
Science intensity	129	0.331**	0.186
		(0.157)	(0.153)
Year 2			
Science intensity	119	0.647***	0.482***
		(0.167)	(0.166)
		[0.001]	[0.014]
Declared practices	119	0.102	-0.044
		(0.171)	(0.171)
		[0.583]	[1.000]
Normative statements	119	0.141	0.102
		(0.175)	(0.190)
		[0.583]	[1.000]
Year 3			
Science intensity	102	0.360**	0.224
		(0.173)	(0.167)
		[0.065]	[0.383]
Declared practices	101	0.432**	0.291
		(0.199)	(0.205)
		[0.065]	[0.383]
Normative statements	100	0.284	0.233
		(0.200)	(0.223)
		[0.065]	[0.383]
Baseline covariates		N	Y

The table gives the impacts of the program on the teacher 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 brackets, we provide the p-values robust to multiple testing, where indices are regrouped per year.
p<0.01 ***, p<0.05 ** p<0.1 *

on the *Science intensity index* are mostly driven by increased hours of science per week (+0.18 hours in year 2 for instance), from a control group average of about 1.5 hours per week, thus a 12% increase. Trained teachers are also more likely to conduct hands-on science experiments, as expected from an inquiry-based intervention (+7 pp, from a control group average of 65%), but only in the second year (and not significantly so when controlling for baseline imbalance). Interestingly, while the effect on the number of hours remains significant in year 3, the number of scientific experiments conducted is unaffected when the teachers design their learning sequences on their own, without the support of the *Maisons*. We will discuss this result further in Section 5 below.

Table 5 also reports effects on the *Declared practices* index, which captures how teachers integrate inquiry-based principles in their science sequences as well as teachers' opinion about these practices (*Normative statement index*). According to Table 5, the training program does not significantly increase neither the *Declared practices* nor the *Normative statements* index in year 2. Despite large effects on the overall amount of science taught, it seems that the program did not significantly modify the way the science was taught in class in year 2. In year 3, one year after teachers completed the training program, the impact on the *Declared practices* or the *Normative statements* are insignificant when we control for baseline imbalances but remain close to acceptance standards. Given the low precision at the teacher level, these point estimates are consistent with possibly large effects at the teacher level. Still, in year 3, teacher level effect sizes are lower than in year 2 and none of them are significantly different from zero, suggesting that without the support of the *Maisons*, teachers do not implement most of the practices they learned during the training.

Of course, these results are self-declared and may be subject to desirability bias. Teachers may not change their actual teaching content or methodology, but may answer our questions based on what they heard in the training sessions. Typically, a classroom observation conducted by enumerators would have alleviated some of these concerns, but not all of them. Indeed, observing a science sequence of the 134 teachers included in our sample would have meant taking appointments long in advance, leaving enough time for teachers to adapt their science sequences to the training content. In our opinion, classroom observation of a specific teaching sequence would also raise desirability bias of its own. In any case, we have reasons to believe that desirability bias is not the main driver of our teacher effects. Indeed, one would expect desirability to be strongly reflected in the *Normative statements* index, as teachers' opinions about

inquiry-based practices can easily be altered during the training or possibly on the *Declared practices index* but less so on the *Science intensity* index which includes very quantifiable and objective practices (hours of science, number of science experiments conducted in class). This is not what we observe in Table 5: our results on teachers are mostly driven by the *Science intensity* index while the *Normative statements* and *Declared practices index* are generally insignificant.

4 Results for Students

4.1 Measures of Student Outcomes

To estimate the effect of the intervention on students, we measure and compare students test scores in science between control and treatment students at the end of survey year 2 (the end of the second year of training) and survey year 3 (the first school year post-training).[32] The students of all volunteer teachers take a grade-specific science test at the beginning and at the end of year 2 and 3.[33] Our tests assess three dimensions of students science achievement: *scientific motivation* and two dimensions of student performance in science: *scientific skills* developed with inquiry-based learning and *scientific knowledge*.

To construct our instruments, we relied on 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:[34]

- *Scientific knowledge*: This index assesses the scientific knowledge of students. All the questions are based on the French curriculum in science for Grade 3 to 5, and, in accordance with it, a few questions are specific to a given grade, while others are common to all grade levels (i.e., when a given subject must be covered at all three 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

[32]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.

[33]Most items are similar across grades. Yet, some items are modified to account for the age level. To account for this, we use grade fixed effects in the regression.

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

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.

- *Scientific motivation*: This index captures students' attitudes toward science. The inquiry-based science method seeks to engage students in investigating scientific questions, and the level of student engagement is used to assess teaching success. This instrument is largely taken from Kind et al. (2007) which develops measures of students' attitudes towards science. This questionnaire is common to the three grade levels.

The scores from the three tests are standardized using the control group's *pre-test* scores and can therefore be interpreted as effect sizes. The observed correlations between test scores, over time and with student characteristics support the validity of our student assessment. First, as shown in Appendix Table A7, 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 ($\rho = 0.05$ but significant) while the one between skills and motivation is null. Second, the three tests are correlated over time between baseline and endline (correlations well above 0.5). Because our indices are normalized by pre-test moments, the averages in *control mean* column of Table 6 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, while students' motivation tends to decline during the school year, a well-known dynamic for this measure (Gillet et al., 2012; Opdenakker et al., 2012). Last, Table A7 provides a set of correlations between the different test scores and some student characteristics: Our tests are properly correlated with the grade level[35] and *late students* (i.e. students who were held back at least once) perform about 0.15 SD below the rest of the class in skills and knowledge.

[35] Although questionnaires are specific to grade level, some items are the same across grades, and therefore we expect to observe a progression between grades. The motivation questionnaire is the same across grades.

4.2 Estimation

In Table 6, columns (1) and (2) present the difference between scientific performance and motivation of students in treatment and control groups at the end of survey years 2 and 3. In this table, we run OLS regressions, with robust standard errors clustered at the teacher level. All observations are weighted by randomization probabilities and regressions include strata fixed effects. We run the following intention-to-treat[36] regression separately by year:

$$y_i = \gamma_0 + \gamma_1 T_i + \boldsymbol{X}_i \boldsymbol{\gamma_2} + \nu_i$$

With y the outcome of interest (knowledge, skills or motivation) of student i, T is the treatment status of the teacher of student i, and $\boldsymbol{X}$ a set of control variables at the student level. In the first column of Table 6, we control for 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.[37] As for teachers, we provide the q-value for the false discovery rate in brackets (Anderson, 2008).

4.3 Impacts on Students

Impacts on students are given in Table 6. 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), the impact is positive (+ 0.1 SD) and significant at 5%. The result is not significant when accounting for multi-hypothesis testing but it is close to acceptable standards (q-value 5.8%). Further, scientific skills and motivation are unaffected in year 2, although both dimensions were the prime objective of the program. 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).

One year after the end of the training in year 3 (same teachers but different students), the impact on knowledge fades out and scientific skills remains unaffected. Our

[36] We restrict our analysis to ITT estimates because we do not have a dichotomous treatment variable as explained in section 2.3 above

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

results even suggest that students motivation is negatively affected (-0.10 SD). This result remains significant at 1% even when controlling for multi-hypothesis testing. This somewhat surprising result is very robust (see robustness test in Appendix Table B1). It 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). Furthermore, in the Appendix Table B1, we decompose the motivation index into three sub-indices. All of them are negatively impacted by the treatment in year 3.

Finally, we investigate the heterogeneity of the training effect. We find little. In Appendix Table A10 and A11, the interaction coefficient between the treatment and different set of baseline characteristics (student scores, student gender, teacher initial training in science or teacher gender) is never significant.

To summarize, we observe a positive and significant effect of the *Maisons'* training in year 2 – at the end of the last training year – on students' scientific knowledge, but this effect quickly fades out in year 3 – one year after the end of the training. In addition, in year 3, we find a significant negative and robust effect on students' motivation in science.

Table 6: Impacts on student 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 level fixed effects. In column *(2)* we add baseline scores, 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. The coefficients in brakets p-values robust to multiple testing. $p<0.01$ ***, $p<0.05$ ** $p<0.1$ *

5 Discussion

Summing up, we find evidence that the training program affects some teacher practices, mostly in year 2: treatment teachers re-use in class the topics covered during the training sessions, they spend more time teaching science, and they tend to conduct more hands-on scientific experiments in class with the students during the second training year. Yet, these effects are short-lived: they are lower and mostly insignificant in year 3. Besides, our measures of inquiry-based teaching show no robust modification, neither through the *Declared practices* nor the *Normative statement* index in years 2 and 3. These mixed results at the teacher level translate into small impacts on student knowledge, limited to year 2. In year 3 (when the training is over), student performance is unaffected and motivation is even lower in the treatment group.

The teacher and student effects are quite consistent. To illustrate this, Table 7 provides the correlation between our measures of teaching practices and student performance *in the control group*. The upper panel of this table shows the correlations between our measures of students' progress over the year (the change between the baseline and the endline measure) and our measures of teaching practices (pooling years 2 and 3 to gain power). Although they tend to be positive, the correlations are small and generally not significant, except for the correlation between the *Science intensity* index and the knowledge score which is both larger and significant (at 10%). Taken at face value, this correlation would imply that the +0.48 SD impact on the *Science intensity* index in year 2 (cf. Table 5) would translate into +0.082 SD on student knowledge—very close to the point estimate we find in Table 6 (+0.097 SD). In year 3, the increase in *Science intensity* of only +0.224 SD (and not significant) would only translate into +0.038 SD effect, far from our detection power, but again consistent with our student results.

This back-of-the-envelope calculation suggests that only large transformations of teaching practices can generate systematically detectable impacts on student performance. Interestingly, in the one experiment that can be directly compared to our study and that finds positive impacts on student performance, Meyers et al. (2016) do find very large impacts on self-declared teacher practices (e.g., +0.73 SD on their inquiry-based learning index, or + 1.441 SD on their technology integration index).

In our case, we observe strong impacts on our *Science intensity* index only. This implies that the training program encouraged teachers to teach more science, with more hands-on experiments, with effects on learning, but did not have large enough transfor-

mative effects on the pedagogy, as the small effect-sizes on the *Declared practices* and *Normative statement* indexes suggest. Indeed, doing more science may improve knowledge, as we see, but, above all, one would have expected well implemented inquiry-based learning methods to improve skills and motivation, which we do not observe.

Given the lack of effects on the inquiry-based practices, we cannot say whether the inquiry-based pedagogy structurally improves student performance in science. The findings by Meyers et al. (2016) do provide experimental evidence that inquiry-based pedagogy *can* have strong effects. But, some research suggests that inquiry-based learning may be ineffective if not implemented with a strong command of it, in particular with adequate guidance of the students (Kirschner et al., 2006; Crawford, 2007; Lazonder and Harmsen, 2016). In our case, the qualitative analysis of the video footage of the training sessions stresses some shortcomings of the training provided by the *Maisons*: Chesnais et al. (2017) and Munier et al. (2021) describe the program as mainly designed in terms of transmission of teaching activities but not structured in terms of development of a new pedagogical approach.

The most important finding is the fade-out at the teacher and student levels: it illustrates the difficulty for a training program to transform teaching practices in a deep and meaningful way. The results, particularly the striking decline in student motivation in the treated classes, suggest that the quality of the in-class sessions have significantly declined in year 3 when teachers ceased to benefit from the direct support of the training centers. Teachers may have lacked the help and encouragement provided to conduct the scientific experiments. Consistently, hands-on experiments, a fundamental component of the inquiry-based pedagogy, were not more frequent in the treatment group in year 3, while they were in year 2 (Table A5). Similarly, we noted (cf. Table 4) that the material and topics covered during the training were re-used in class in year 2, but not in year 3. This indicates that actual change has happened during the training, but it faded out rapidly when direct support disappeared.

Table 7: Correlations between teaching practices and student outcomes (control group)

	Intensity Index	Declared practices index	Normative statements index	# of hands-on exper.	Holds a degree in science
Student scores					
Δ knowledge	0.171	0.067	0.078	0.138	-0.029
	[0.095]	[0.517]	[0.448]	[0.181]	[0.779]
Δ skills	0.119	0.003	-0.001	-0.009	-0.082
	[0.249]	[0.977]	[0.99]	[0.93]	[0.428]
Δ motivation	0.12	0.073	0.126	0.022	0.049
	[0.243]	[0.482]	[0.22]	[0.832]	[0.636]
Teaching practices					
Science intensity	1	0.283	0.205	0.631	-0.078
	[0.000]	[0.005]	[0.045]	[0.000]	[0.451]
Declared practices	0.283	1	0.532	0.359	-0.124
	[0.005]	[0.000]	[0.000]	[0.000]	[0.23]
Normative statements	0.205	0.532	1	0.191	-0.049
	[0.045]	[0.000]	[0.000]	[0.062]	[0.64]
# of hands-on exper.	0.631	0.359	0.191	1	-0.044
	[0.000]	[0.000]	[0.062]	[0.000]	[0.67]
Teacher traits					
Holds a degree in science	-0.078	-0.124	-0.049	-0.044	1
	[0.451]	[0.23]	[0.64]	[0.67]	[0.000]
Teaching experience	-0.155	-0.105	-0.215	-0.029	-0.193
	[0.135]	[0.31]	[0.037]	[0.778]	[0.061]
Gender	0.028	0.089	0.188	0.119	-0.256
	[0.789]	[0.388]	[0.066]	[0.249]	[0.012]
Observations	96	96	96	96	95

The table provides the correlations between our measures of teacher practices and student increase in performance over the school year and teacher characteristics, all computed in the control group, years 2 and 3 pooled. In square brackets, we provide the p-values. *Observations* is the number of year-control teachers observations.

6 Conclusion

Despite the potentially large social benefits of teacher training programs in terms of teacher quality, the literature on training program remains scarce and inconclusive. Among the 18 experimental studies we identified as fairly comparable to our setting, only five studies found some positive results. Two of these training programs were inspired by inquiry-based principles, while three studied phonological awareness in primary school or kindergarten. These few success stories in the literature suggest that intensive and well-executed training programs, based on validated pedagogical approaches, might be instrumental in improving student performance. In this paper, we test this claim using a large randomized experiment.

Our study is (i) well powered , (ii) based on a widely recognized pedagogical strategy (i.e., inquiry-based learning), (iii) based on an intensive training program—2 years, 80 hours, comparable to studies that found positive impacts on student performance (Meyers et al., 2016; Newman et al., 2012), and large in comparison to the usual number of hours of training in science received by our sample before randomization ($\approx$ 2 hours). Further, to fully understand the causal pathways, our study also includes precise information on both teachers and students. Last, our analysis extends beyond the school years during which the program is rolled out, in order to measure whether good practices carried on one year after the end of the training program, a fundamental feature to consider when assessing such interventions.

We find evidence that the training programs modified some teaching practices but these effects are short-lived : teachers spent slightly more time teaching science (about a fifth or an hour more or about a 12% increase), they conduct more experiments, and those are more likely to be hands-on experiments as instructed by the inquiry-based approach. Yet, in spite of these impacts at the teacher level in year 2, we find no convincing evidence that these translate into positive impacts on student performance. While we do find positive and significant impacts on knowledge in year 2 (the last year of the training), the impact disappears in year 3, one year after the end of the training program. In addition, we find no impacts on skills, which is the prime objective of the inquiry-based method. Our results even indicate negative impacts on motivation one year after the end of the training program (year 3). The comparison between year 2 and year 3 impacts on both teachers and students suggests that the relative quality of teaching declined one year after the end of training program. The training program did have an effect on some quantifiable practices (more time spent teaching science,

more experiments) but probably did not profoundly changed middle- and long-term practices.

Does this conclusion apply to any training program inspired by inquiry-based principles? Probably not, as other similar papers did find more positive results (Meyers et al., 2016; Newman et al., 2012). Yet, our paper confirms some of the shortcomings of this pedagogy and of training programs in general. First, teacher practices are hard to change and, as a result, only intensive and specific training programs are likely to translate into better student performance. Interestingly, Meyers et al. (2016) do find impacts of their training program but the unusual intensity of their training program (240 hours of training over two years) makes it non-comparable to most other training programs and possibly non-implementable in most contexts. Second, as a direct result of point 1, such intensive and specific training programs are likely to attract teachers who are interested in the topic and have sufficient time to devote to the training. We do find evidence of such selection in our context and we believe that this is an under-reported issue in the literature. We cannot say whether or not such selection has affected our results but this is an important concern for the overall effectiveness of training programs. Third, even when teacher practices are considerably modified—as, for instance, in (Meyers et al., 2016)—the treatment effect remains modest in relationship to the organizational cost (such as finding and paying for substitute teachers), which has proven to be high in our case. Our results, together with related work, highlight the challenges in terms of targeting, intensity and quality that teacher training programs still need to address to be considered as viable strategies to improve overall teacher productivity.

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.

Baumol, W. J., de Ferranti, D., Malach, M., Pablos-Méndez, A., Tabish, H., and Wu, L. G. (2012). *The Cost Disease: Why Computers Get Cheaper and Health Care Doesn't*. Yale University Press.

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.

Bouguen, A., Gurgand, M., and Grenet, J. (2017). Does class size influence student achievement? Technical Report 28, PSE.

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.

Carrell, S. E. and West, J. E. (2010). Does professor quality matter? evidence from random assignment of students to professors. *Journal of Political Economy*, 118(3):409–432.

Center, Y., Wheldall, K., Freeman, L., Outhred, L., and McNaught, M. (1995). An evaluation of reading recovery. *Reading research quarterly*, pages 240–263.

Chesnais, A., Cross, D., and Munier, V. (2017). Etudier l'effet de formations sur les pratiques en termes de connaissances : réflexion sur les liens entre connaissances et pratiques. *Recherches en Didactique des Sciences et des Technologies*, 15:97–132.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., and Yagan, D. (2011). How does your kindergarten classroom affect your earnings? evidence from project star. *The Quarterly Journal of Economics*, 126(4):1593–1660.

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

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

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

Djeriouat, H. (2015). Validation du questionnaire d'évaluation des connaissances et attitudes vis à vis de la science. *mimeo*.

Filmer, D. and Rogers, H. (2018). Learning to realize education's promise. *World Development Report. The World Bank*.

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.

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.

Hanushek, E. (1971). Teacher characteristics and gains in student achievement: Estimation using micro data. *The American Economic Review*, 61(2):280–288.

Hanushek, E. A. and Rivkin, S. G. (2006). Teacher quality. *Handbook of the Economics of Education*, 2:1051–1078.

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.

Kane, T. J. and Staiger, D. O. (2008). Estimating teacher impacts on student achieve-

ment: An experimental evaluation. Technical report, National Bureau of Economic Research.

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.

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.

Rivkin, S. G., Hanushek, E. A., and Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, 73(2):417–458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American economic review*, 94(2):247–252.

Rockoff, J. E., Jacob, B. A., Kane, T. J., and Staiger, D. O. (2011). Can you recognize an effective teacher when you recruit one? *Education finance and Policy*, 6(1):43–74.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics*, 125(1):175–214.

Schwartz, R. M. (2005). Literacy learning of at-risk first-grade students in the reading recovery early intervention. *Journal of Educational Psychology*, 97(2):257.

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 achievement. *Issues & answers. REL*, (033).

Appendix

Table A1: Summary of the literature

Authors	Year	Country	Design	topic	Grade	Sample	Intensity	ES	Sig.	LT stud.	LT teach.
Angrist and Lavy (2001)	01	Israel	NE	m;r	prim	31/850	11 d	.3-.6	s	-	-
Borman et al. (2007)	07	USA	RCT	r	K-G2	35/4180	95h, 1y	.21-.33	s	s	-
Garet et al. (2008)	08	USA	RCT	r	prim	90/5000	108h, 2y	.03	ns	ns	ns
Machin and McNally (2008)	08	UK	NE	r	G1	14000/1.6m	intens.	.09	s	-	-
Gersten et al. (2010)	10	USA	RCT	r	G1	81/468	20h, 6m	.23	ns	-	-
Kim et al. (2011)	11	USA	RCT	r	G6-12	103/2726	46 h	.05	ns	-	-
Randel et al. (2011)	11	USA	RCT	m	G4-5	67/4700	self	.10	ns	-	-
Allen et al. (2011)	11	USA	RCT	g	G4-5	76/2237	1y	.22	s	-	-
Harris and Sass (2011)	11	USA	NE	m;r	prim	1031/487000	NA	.00	ns	-	-
Campbell and Malkus (2011)	11	USA	RCT	m	G3-5	36/24759	very	.05	ns	ns	-
Newman et al. (2012)	12	USA	RCT	m;s	G4-8	82/3000	10 d, 2y	.05	s	-	-
Bos et al. (2012)	12	USA	RCT	r	G6-8	52/18180	14 days	.01	ns	ns	ns
Sirinides et al. (2018)	18	USA	RCT	r	G1	9784	intens.	.47	s	-	-
Meyers et al. (2016)	13	USA	RCT	m;r	G7-8	60/3072	24 h, 2y	.13-.17	s	-	-
Gentaz et al. (2013)	13	France	RCT	r	K	56/2398	42 h	.00	ns	-	-
Rimm-Kaufman et al. (2014)	14	USA	RCT	g	G3-5	24/2904	20 h	.00	ns	-	-
Bouguen (2016)	17	France	NE	r	K	118/4345	intens.	.15	s	-	-
Loyalka et al. (2019)	19	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), are rigorous and are sufficiently powered (>.30 sd ex-ante MDE). 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 two columns provide the results from the student long-term measures (students tracked at least one year after having benefited from a trained teachers) and teacher long-term measures (students benefiting from a teacher who has been trained at least one year prior).

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 trainer 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 from this table are taken from Munier et al. (2021).

Table A3: Balance checks - students' outcomes and characteristics

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

The table provides the baseline difference, in year 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 parenthesis.
p<0.01 ***, p<0.05 ** p<0.1 *

Table A4: Impacts on the number of science topics covered in class

	Obs.	(1)	(2)
Year 1	129	0.247	0.020
		(0.234)	(0.233)
Year 2	119	-0.272	-0.467*
		(0.267)	(0.269)
Year 3	102	0.075	0.038
		(0.283)	(0.309)
Baseline covariates		N	Y

The table gives the impacts of the program on the number of different science topics covered in class each year. Column *Obs.* gives the number of teachers, *(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. Standard errors are given below the regression coefficients in parentheses.
p<0.01 ***, p<0.05 ** p<0.1 *

Table A5: Impacts on teacher qualitative practices, detailed

	Year 1			Year 2			Year 3		
	C	(1)	(2)	C	(1)	(2)	C	(1)	(2)
Science intensity index									
Feels teach enough sci.	-0.041	0.262	0.153	-0.051	0.488**	0.355*	0.003	0.378*	0.292
		(0.175)	(0.178)		(0.203)	(0.199)		(0.196)	(0.207)
Weekly hours of science	1.537	0.137	0.030	1.429	0.252***	0.180*	1.349	0.252**	0.186*
		(0.089)	(0.083)		(0.093)	(0.095)		(0.107)	(0.108)
# hands-on science expe.	0.642	0.053	0.042	0.652	0.088**	0.068	0.644	0.025	-0.004
		(0.042)	(0.043)		(0.044)	(0.046)		(0.049)	(0.050)
# science expe.	0.290	0.029	0.057	0.349	0.013	0.023	0.382	-0.064	-0.054
		(0.056)	(0.060)		(0.062)	(0.066)		(0.069)	(0.070)
Declared practices index									
Introduces sci. problem				-0.054	0.364**	0.294	0.067	0.472**	0.327
					(0.181)	(0.191)		(0.200)	(0.209)
Works on students vision				0.001	-0.001	-0.147	0.018	0.343*	0.268
					(0.189)	(0.188)		(0.206)	(0.220)
Evaluates students				-0.045	-0.106	-0.264	0.000	0.228	0.109
					(0.173)	(0.170)		(0.188)	(0.195)
Normative statements index									
Importance of ...									
... Introducing sci. pb.				-0.016	0.159	0.193	-0.002	0.141	0.086
					(0.180)	(0.183)		(0.202)	(0.222)
... formulating hyp.				-0.022	0.185	0.148	-0.009	0.165	0.118
					(0.188)	(0.202)		(0.188)	(0.204)
... linking model to obs.				-0.005	0.014	-0.006	0.028	0.160	0.124
					(0.181)	(0.210)		(0.177)	(0.179)
... evaluating students				0.011	0.066	-0.059	-0.003	0.391*	0.376
					(0.166)	(0.171)		(0.233)	(0.270)
Baseline covariates					N	Y		N	Y

The table provides the impact on teacher quantitative practices. Column *Obs.* gives the number of observation, *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 are given below the regression coefficients in parenthesis.
p<0.01 ***, p<0.05 ** p<0.1 *

Table A6: 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, 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 A7: 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 and 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 A8: 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 A9: 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 observation, *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 parenthesis.
p<0.01 ***, p<0.05 ** p<0.1 *

Table A10: 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 treatment variable and H 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 analyse the heterogeneity by baseline teacher characteristics (having a diploma in science and being a woman). Columns (1) gives the result of the regression without baseline covariate while columns (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 parenthesis.
p<0.01 ***, p<0.05 ** p<0.1 *

Table A11: 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 previous note but for year 3.

Appendix B : Robustness of results on students motivation

To get a sense of how robust the negative effect on motivation effect, 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 A9) and have a relatively high Cronbach Alpha[38]. We label the three sub-dimensions "I like science", "Scientific mindset" and "Science is easy".[39]

Table B1 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 one driven by a few items, or happening by mere chance.

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

[39]The details of those new indexes are in the Appendix Table A8.

Table B1: Impacts on the motivation of students

		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 observation, *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 parenthesis. In square brackets, we provide the p-values robust to multiple testing. $p<0.01$ ***, $p<0.05$ ** $p<0.1$ *