Teacher Training Program, 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¶

January 15, 2021

Abstract

We conduct a large randomized study to assess the impact of an in-service teacher training program on inquiry-based teaching in sciences. The randomized study is conducted in several French primary schools, comprises randomly assigned 134 teachers and two cohorts of about 2500 students. In addition to standardized student test score, we collect information on teacher quantitative and qualitative practices and pedagogical knowledge. We find that the training program (80 hours over 2 years) significantly increase the weekly hours of science as well as the number of hands-on science experiments conducted in class. However, the program only marginally modified teachers' pedagogical approach while our data suggest that very large change would be necessary to produce substantial impacts on student performances. Consequently, our results on the student performances are disappointing: impacts are significant only on scientific knowledge and not on skills and are, in any case, fading out one year after the end of the training.

JEL classification: I20

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

*University of Mannheim
†Santa Clara University
‡Paris School of Economics
§Université de Montpellier
¶Université de Montpellier

Teachers greatly explains student performances. 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, an effect 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 high and low-income countries that improving teacher quality is key to resolve the education crisis (Filmer and Rogers, 2018). Yet, 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 studied 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 remains mostly unobserved (Hanushek, 1971), led some to conclude that teacher performances are innate, stable over time and that teacher quality 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. Yet, as mentioned by Fryer and described in Table A1, the literature also includes resounding success stories. Consistent evidence shows for instance that 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 7000 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 includes remedial education components[2], it remains unclear whether the teacher training program *per se* would have resulted in positive results. Yet, other results based on non-experimental variation (Bouguen, 2016; Machin and McNally, 2008; Angrist and

[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 SD

[2] For instance reading recovery includes a 30 minutes reading session targeted to low achievers in the class. Similarly Success for all includes a comprehensive school-level reform: change in curriculum, in classroom structure... As a result, both programs are fairly expensive in comparison with regular teacher training program. 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" are effective at increasing students performance.

Lavy, 2001) tend to confirm that well-executed and intensive training program, based on validated pedagogical approaches can modify teaching performances.

In this paper, we build on the teacher training literature to design a large scale experimental study to test the impact of a specific pedagogical approach – the inquiry-based method – designed to improve primary school students' scientific skills. The experimental study randomizes 134 volunteer teachers and two cohorts of their students (about 2500 for each cohort) into a treatment and a control group. To fully track how the intervention impacts the teacher quality, we collect teacher level quantitative[3] as well as qualitative measures[4] of the teachers' practices in science. We then measure how these potential teacher effects translate into better student scientific skills, knowledge and motivation. We measure impacts first during the last year of the training program (year 2) and then one year after the end of the intervention (year 3). We therefore measure the short and midterm impacts on scientific skills, knowledge and motivation on two cohorts of about 2500 students.[5]

The program, based on 80 hours of training sessions over two school years, is directly inspired by the inquiry-based method, a learning method that emphasizes students' role in the learning process. Inquiry-based learning involves conducting experiments in which students actively participate in defining the research questions, the scientific problems, the hypothesis and where students are involved in finding the solutions. Inquiry-based learning often involves hands-on science experiments conducted in class and designed by students to test a set of scientific hypothesis. Inquiry-based learning method is often considered as the best approach to teaching science and has long been endorsed by the National Research council (Council et al., 2000). It is expected to increase student engagement in class, motivation in the discipline and eventually scientific skills and knowledge.

Despite its popularity, the benefits of the inquiry-based 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[6]), only two studies specifically

[3]Example of quantitative items are: weekly hours of science taught, number of hands-on scientific experiment conducted, number of scientific topics covered in class...

[4] Example of qualitative items are: "during the science sessions, did students formulate the scientific problem, the hypothesis? Did they test the hypothesis?"...

[5]Some teachers may have changed grade and therefore had the same students during the first and second year. This concerns a minority of cases however see (infra).

[6]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

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) finds that a fairly intensive inquiry-based training program[7] has small but significant impacts on mathematical skills (+0.05 SD) but no significant impacts on scientific skills. The study can only capture the impacts after one year of the two years of intervention as the randomization is compromised in the second year.[8] 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 the first two years of intervention). The program is comprehensive (at the school level), very intensive and heavily relies on the inquiry-based principles. The study targets G7 and G8 students and find 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. Given the lack of systematic review of the impact of the inquiry based method on STEM skills, we do not believe the literature is conclusive on the effective of such learning method on student performance in science.

Our paper contributes to this fairly restricted literature by testing at scale the impact of an inquiry-based learning method on both teachers and students performances. Conversely to Newman et al. (2012), our design allow measuring impact at the end of the two years (year 2) and one year after the end of the intervention (year 3). The intensity of the program we investigate (80 hours over two years) 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). Our design based on 134 clusters and two cohorts of 2500 students is also more precise than most equivalent studies in the literature.[9] Finally, as we will see below, the adherence to the randomized protocol

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)

[7]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 analyses the impact on G4 to G8 on maths, sciences and technologies.

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

[9] According to our calculation based on our selected studies listed in Table A1, our study remain 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 our experiment (see Table A1)

is very high. We therefore contribute to the teacher training literature by providing a precise measure of the teacher training impact that includes impacts one year after the end of the intervention. We believe that having two years of impacts is crucial for the teacher training literature because the cost-effectiveness of such program crucially depends on whether or not the teacher practices learned during a training are preserved throughout the teachers' career. If such program affects teacher quality throughout their career, teacher training programs are key to improve the overall quality of education.

Last, in addition to measuring the impact of the program on student performances, we design students and teacher surveys to capture the different aspects of the inquiry-based method, both quantitatively and qualitatively. Every year we collect quantitative and qualitative measures at the teacher level to capture whether their practice of science have been modified. Our survey also precisely documents the content of the intervention, its intensity and the level of teacher satisfaction. Finally, we include a measure of the students' motivation in science, a mechanism that is key to the inquiry-based method but has never been precisely measured in previous work. These additional measures at the student and teacher level, conducted every year, help establishing a clearer causal route and better understanding the constraints that often make teacher training program ineffective.

Our results first highlight the inherent difficulties associated with conducting a training program at scale. We document the difficulties we experienced to recruit and motivate teachers to participate in an intensive in-service teacher training program. While our original target was to recruit 320 teachers in four school regional districts,[10] the school district managers were only able to recruit 134 teachers in three school districts.[11] We faced difficulties in finding substitute teachers, convincing teachers to commit to a two years program or to attract teachers with lower level of scientific experience.[12] Although our sample size remains sufficient to precisely estimate the impacts on the student performance, our group of volunteer teachers (134) is unsurprisingly different from the average teachers in similar schools. Compared to non-volunteer peer teachers enrolled in the same school, the 134 volunteer teachers included in the experiment are older, more experienced, have received more pre-intervention in-service

[10] *Academies* in French. There are 18 academies in France.

[11] One of the school district entirely dropped out because they could not find enough substitute teachers to replace teachers during training sessions.

[12] As we will see later, training hours were not paid and were taken from normal teaching hours.

training and practice much 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, targeting in-service teacher training programs to teachers who lack scientific background, with low level experience teaching science or low interest in science is a challenging task. We believe this is one of the likely impediment to scaling up intensive teacher training programs.

Second, we find that among the fairly selected 134 volunteer teachers, the teacher training program was almost perfectly implemented. Treatment teachers declared having received an average of 78 hours of in-service teacher training over the two school years of the program (against 11 hours in the control). The experiment therefore generated a 67 hours increase in in-service training, fairly close to the objective of 80 hours additional training. In contrast, the year before the program started, volunteer teachers declared having received about 11 hours of training overall and only 2 hours of training in specifically about science. This program therefore constitutes a very significantly increase of the regular volume of in-service training provided to teachers in France.

Third, we find evidence that the training program affected teacher practices on the short-term, maybe less so one year after the end of the training. At the end of the two years training program (year 2), treatment teachers declare teaching slightly more hours of science per week (+0.175 h/week) and conduct more hands-on scientific experiment (+7 pp). Some of these effects remain detectable one year after the end of the intervention (year 3): treatment teachers still declare more hours of science (+0.225 h/week) than the control group but the number of hands-on experiment conducted in class fades out to a non-significantly +4 pp effect. Importantly, during the last year of the training, teachers benefited from in-class assistants to conduct experiments in-class which could explain in part why the number of self-declared experiments is lower in year 3. Similarly, we find that the science topics covered during the training affects the topics covered in class in year 2 but this effects fade out quickly as well. Inversely, our qualitatively measures of teacher practice does not show any sign of fade-out: teachers declare implementing the inquiry-based principles in year 2 (+0.12 SD, barely significant) and more so in year 3 (+0.288 SD). Overall, the quantitative effects on the volume of science taught per week are modest and show some evidence of fading out: they correspond to a 12% - 17% increase in the total volume of science taught. The fact that the impact on hands-on experiment and the one on the topics covered in class

fades out one year after the end of training suggests that implementing in class the inquiry-based principles on the long-run is challenging either because good practices fade out or because such practices are hard to conduct without external help.[13]

Fourth, given these mixed results at the teacher level, our measures of impact at the student level show modest and short-lived impacts on scientific knowledge, no impacts on scientific skills and even negative impacts on motivation for science. We find that students in the treatment group perform +0.11 SD better than control students in year 2 (barely significant when accounting for multi-hypothesis testing) but this effect is not significant one year after the end of the intervention in year 3. The fact that student effects rapidly fades out while the effect on the number of hours of science taught remains constant is indicative that the quality of the science sessions conducted in class has significantly declined after the end of the intervention. The negative effect on motivation in year may even suggests that, when poorly implemented, inquiry-based pedagogy reduces student interest in science. The fact that the impact on the number of hands experiments is not significant in year 3 may explain in part the decline in teaching quality. Yet, our findings suggest that this cannot be the only factor. Unfortunately, our qualitative measure at the teacher level does not provide enough information to identify what underlying mechanism is driving down the effect in year 3. Improvement in the quality of teacher level measures need to be made to better understand the dynamics behind our effects. Suffice here to say that the program's impact did not carried out beyond the duration of the teacher training, suggesting that the quality of the scientific session conducted in class has significantly declined once the intervention stopped.

Since the efficiency and cost-effectiveness of teacher training program crucially depends on whether teacher practices are persistent over time, our findings raise important doubt on the validity of the inquiry-based approach and highlight the difficulty to enhance teacher quality via teacher training programs only. Our results indicate that teacher practices among experienced teachers are hard to modify, when modified are short-lived and that this approach is unlikely to be viable strategy to fundamentally improve teacher quality in science. Our results also suggest that the prior, supported by many inquiry-based approaches, that there exists a causal chain between interest for science, skills and scientific knowledge is not demonstrated in our context.

In the rest of the article, we will first describe the teacher training program con-

[13]as mentioned above, during year 2, teachers benefited from the infrastructure of the teaching training program and in-class assistants while they were left on their own in year 3

ducted by *the Maisons pour la Science*, discuss the experimental setting, data and compliance and in a third party discuss our results on the students and on the teachers. We conclude on the contribution of this paper to the teacher training literature and the challenges faced by education system to improve education quality.

1 Maisons' training Program: background and content

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

1.1 Background of the "Maisons pour la Science" project

The *Maisons pour la Science* project was initiated by the *Grand emprunt*[14], a vast 57 billion euros 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.[15] In 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 from the *the Grand Emprunt* to finance 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 program consists in establishing local training centers, called *Maisons pour la Science*, in four academic regions.[16] 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 topic of this research.

[14] initiated under the Sarkozy presidency and later renamed *investissement d'avenir* under Holland Presidency

[15] Grand Emprunt financed for instance new universities, research centers, innovative start-up companies, green technologies...

[16] so-called *academies* in French

1.2 The *Maisons*' training program

The Maisons' training program is eligible to G3, G4 and G5 primary school teachers and is inspired by the inquiry-based learning method, a learning method that emphasizes student's role in the learning process. Inquiry-based learning often involves conducting experiments in which students actively participates in defining the research questions, the scientific problem, the hypotheses and where students are directly involved in finding the solutions. Inquiry-based learning often involves hands-on experiments conducted in class in which students designed a scientific experiment to test a scientific hypothesis. The objective of the training program is to help teachers designing their own scientific teaching sequence and facilitate the implementation of the in-class experiments with supply of material and visits of a teacher aid. The intensive program is composed of 80 hours of training and lasts two consecutive school years. The 80 hours includes training sessions at the *Maisons*, in class observations, discussion with other teachers and attendance to scientific conferences. Treatment teachers attend the training sessions during their normal teaching hours. The school district manager appointed a substitute teacher during the training hours, therefore insuring that the program did not reduce normal schooling hours.

2 Experimental Setting, Data and Compliance

2.1 Sample, Randomization and Surveys

Although the intervention was conducted in four regional school districts[17], only three took part of the experiment: Auvergne, Lorraine and Midi-Pyrénées.[18] Within the three regional school districts, school district managers were responsible to conduct the recruitment of the teachers and find substitute teachers. The objective of the recruitment was to identified teachers who would be willing to attend the training, who might benefit from it and who could be substituted for during the training hours. Before the beginning of the intervention, we specifically highlighted to the district managers the important of identifying teachers who lack training in scientific teaching method (i.e. teachers with no prior scientific degree, who expressed difficulty designing

[17] *académies*

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

scientific sequences....). Our original objective was to recruit 80 volunteer teachers per regional school district i.e., 320 teachers overall.

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 the volunteer teachers during three years: before the beginning of the intervention when teachers applied to the program in the summer 2013, 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 teaching about science, number of hours of teacher training...). We then randomly assigned volunteer teachers to treatment or control group and the teachers assigned to the treatment group started the training program the following year (year 1).[19]

We randomized at the local school district level[20] within the three regional school districts.[21] Since each school district administered Q0 at different point in time (from June to September 2014), we randomized each district individually. In most districts, we stratified using baseline teaching experience and the number of hours the teacher declared practicing science every week in class.[22] Finally, since each district had a fixed number of available teacher training available spots, each school district had a unique probability to be assigned to treatment. We account for this by including randomization weights equal to the inverse of the assignment probability.

The original sample is composed of 134 teachers across three regions and nine school districts as described in Table 1. In addition to teachers in treatment and in control group, we include in our teacher sample a group of peer teachers: these peers are non-volunteer teachers, working in the same school as the volunteer teachers and teaching in G3, G4 or G5. We use peer teachers to document how volunteer teachers and their students (self-)selected themselves into the program.

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

[20] *DASEN* or *department* in French

[21] *academies* in French

[22] In two cases where several teachers from the same school applied to the program, we stratified at the school level. In one district (54), the teachers did not fill out Q0 before randomization: in this case, we used the municipality as a variable of stratification.

Figure 1: Survey protocol

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

The figure presents the survey protocol for the main sample. Note that an additional school district (Tarn), 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. Tarn schools filled out Q0 survey and were randomized at the same time as the rest of the sample however.

Table 1: Teacher Sample Size by Survey Round

	Full Sample	Control	Treatment
Q0 survey			
Volonteer Teachers	134	62	72
... Auvergne	23	12	11
... Lorraine	37	16	21
... Midi-Pyrénées	69	32	37
Q1 survey			
Volonteer Teachers	130	61	69
Peer teachers	83	33	50
Q2 survey			
Volonteer Teachers	129	60	69
Peer teachers	89	37	52
Q3 survey			
Volonteer Teachers	114	48	66
Peer teachers	60	24	36
Q4 survey			
Volonteer Teachers	113	50	63
Peer teachers	0	0	0

The table 1 gives the number of volunteer and peer teachers by treatment the control and the peer group for each survey.

After completing 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). Because teachers do not usually change grade between two consecutive school years, the students surveyed in the first school year are generally different from the ones surveyed in second school year.[23] Consequently, our dataset is a panel at the teacher level but a cross-section at the student level.

2.2 Sample Description and Selection

Using data from the Q0 and Q1 survey among volunteer and peer teachers, we describe the sample of volunteer teachers and how they compared with their peers. We present our results in Table 2. In the first panel of Table 2, we compare the treatment and control teachers at baseline and compare 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 (second panel). Volunteer teachers are significantly older (+2 years), have more teaching experience (+4 years) and are more likely to hold a degree in science (+ 16 pp almost significant at *p-value*=10.2%). In addition to being more experienced, volunteer teachers are more likely to have benefited from in-service training (+17 pp) and benefited from more training hours (+3 hours). Likewise, prior to randomization, volunteer teachers are more likely to have benefited from the *Maisons* and from *La Main à la pâte*, a local NGO that provides training about inquiry-based learning in science. Prior to applying for the *Maisons*'s training, volunteer teachers also report teaching more sciences (+0.82 hours or +66% more hours of science) and cover in average one additional scientific topic during a school year.

Overall, not surprisingly, volunteer teachers are more trained in science, are more exposed to in-service training pre-intervention and are more used to practicing science

[23] 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.

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 and accustom 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 lower level of scientific awareness has not produced significant results. More generally, we believe that this lack of targeting reflects an important conundrum for teacher training : targeting the program to the teachers who need it the most is challenging and may be one of the under-studied reasons why teacher training programs are generally ineffective at improving students performance.

More 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), specifically when compared with the global volume of training provided by our intervention (40 hours per school year). The comparison between our training intervention and the usual volume of training received by volunteer teacher is even more striking when looking at the training in science: volunteer teacher are indeed used to 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.

Finally, we reject the null hypothesis that the treatment and control teachers are different at baseline in two cases 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%.[24] Yet, the fact that the self-declared number of hours of sciences in the treatment group is larger than the one in the control group is concerning since this is one of the intermediary expected outcome. As a robustness test, we will add baseline hours of sciences to our regressions in the analysis of the results.

[24]Say it otherwise, the minimum Q-value is 59% for these 15 coefficients, far from standard significant level.

Table 2: Pre-Randomization Teacher Characteristics - Q0 Survey

	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	0.054
			(0.080)			(0.060)
Birth year	132	1969.97	-0.714	214	1971.27	-1.680
			(1.088)			(1.214)
Higher education in years	132	2.836	0.356**	215	3.157	-0.130
			(0.165)			(0.164)
Holds a scientific degree	132	0.637	-0.123*	212	0.396	0.174**
			(0.071)			(0.083)
Had a career in science	132	0.146	-0.018	213	0.107	0.029
			(0.050)			(0.050)
Teaching experience	132	17.456	0.516	214	14.343	3.392**
			(0.927)			(1.419)
In-service training last year						
Received some training	132	0.287	-0.011	214	0.145	0.136
			(0.056)			(0.085)
Total training hours	118	11.179	-1.753	199	3.460	6.774
			(7.753)			(4.695)
Total training hours in science	118	2.077	0.872	199	0.996	1.550
			(1.720)			(1.162)
Received Maisons training	132	0.203	-0.040	214	0.015	0.167***
			(0.063)			(0.057)
Received La Main à la Patte	132	0.174	-0.064	214	0.011	0.128**
			(0.067)			(0.053)
Teaching practices last year						
# of hours of sciences	132	1.920	0.271**	189	1.234	0.831***
			(0.120)			(0.121)
# of topics covered (max 8)	132	5.098	0.277	189	4.030	1.217***
			(0.282)			(0.284)
% of sequences with experiments	132	0.569	0.032	205	0.588	-0.001
			(0.037)			(0.046)
Practice inquiry-based	132	0.814	0.082	.	.	.
			(0.049)			.
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. Dependant variables are displayed in rows. Column *Obs.* gives the number of observations (here teachers), column *Control* the average in control group, *Peer* the average in the peer teachers and column *(1) and (2)* the result of the regression of the dependant variables against the treatment variable or the *volunteer* teacher variable. Clustered at the strata level, standard errors are given below the regression coefficients in parenthesis.
p<0.01 ***, p<0.05 ** p<0.1 *

2.3 Exposure to Training

Using data from Q1 and Q2 teacher survey, 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 to be enrolled into the training program provided by the *Maisons* (+ 72 pp in the first year +87 pp in the second year). While some control teachers benefited from the program (14% the first year and 3% the second), the difference between the experimental groups is significant. Over the two years, our results indicate that 95% of the treatment teachers received some form of training provided by the *Maisons*.[25] 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 of training the second year. Overall, treatment teachers reported approximately 65 hours of training offered by the *Maisons* and 78 hours of any training over the two years, very close to the objective to offer 80 hours of training over the two year.[26]

Last, we look at whether enrolling into the *Maisons* training program had spill-over effects on other training program's enrollment during or after the intervention. Specifically, we could be worried that the intensive training program offered by the *Maisons* comes as a substitute to other training programs provided by other institutions either on science or on 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 hour of training from Maisons" indicating no systematic patterns of substitution. Likewise, we do not find any significant spill-over on enrolling to 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 3.5 hours of training per year, not significantly different in the treatment and control group.

Overall, the training provided by the *Maisons* was remarkably well implemented: Treatment teachers received overall 78 hours of training and only a few teachers benefited from the program in control. Besides, the intervention did not come as a substitute

[25]Control group teachers were not supposed to receive any training from the Maisons.We believe that in some limited cases the *Maisons* took the liberty to authorize some control teachers to attend training

[26]These results are consistent with the one found when using 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.

Table 3: Exposure to Training

	Obs.	Control	Impact
Year 1 & Year 2			
Received any training	132	0.444	0.540***
			(0.065)
... from *Maisons*	132	0.155	0.791***
			(0.066)
# of hours of any training	132	10.751	67.57***
			(6.527)
... from *Maisons*	132	4.539	61.27***
			(6.989)
Year 1			
Received any training	129	0.275	0.660***
			(0.078)
... from *Maisons*	129	0.142	0.719***
			(0.062)
# of hours of any training	129	6.187	36.68***
			(3.892)
... from *Maisons*	129	3.747	32.11***
			(4.426)
Year 2			
Received any training	127	0.266	0.652***
			(0.065)
... from *Maisons*	127	0.031	0.873***
			(0.042)
# of hours of any training	127	4.974	32.09***
			(4.385)
... from *Maisons*	127	0.942	30.37***
			(4.049)
Year 3			
Received any training	115	0.243	0.038
			(0.088)
... from *Maisons*	115	0.111	-0.050
			(0.052)
# of hours of any training	115	5.228	0.669
			(2.250)
... from *Maisons*	115	3.579	-1.280
			(2.384)

The table shows difference between the treatment and control group (column *Impact*) in terms of the exposure to training program, both overall exposure and exposure to training program provided by the Maisons. Column *Obs.* gives the number of *volunteer* teachers surveyed, *Control* the average in the control group. Standard errors are clustered at the strata level and given below the regression coefficients in parenthesis.
p<0.01 ***, p<0.05 ** p<0.1 *

to other form of training. Finally, in results not presented here, treatment teachers expressed a high degree of satisfaction to the training program provided by the *Maisons*: 86% was rather or very satisfied after the first year, 87% after the second year. They expressed satisfaction with all the aspects of the training: in-class visits (95% satisfaction), on-site training sessions (92% satisfaction), group work (87% satisfaction).

3 Impact on Students

3.1 Measures of Student Outcomes

To estimate the effect of the intervention on students, we measure and compared scientific performance and motivation between control and treatment groups 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)[27].

The students of all volunteer teachers take a grade-specific test at the beginning and at the end of year 2 and 3.[28] These tests capture *scientific motivation* and two dimensions of students performance in science: *scientific skills* and *scientific knowledge*. We conducted a pilot before the beginning of the intervention to validate our tests and select the best items for each grade [cite Akim mimeo + add details on skills and knowlegde (MPIS, RIDS, NOS and COD)]. We built three indexes for each dimension by averaging (normalized) items within each dimension. All indexes are then standardized by pre-test scores standard-deviation in the control group and can therefore be interpreted as effect-sizes.

We present the results of the test administered at the end of year 2 and 3 in Table 4 — respectively in Table A2 for the results at the beginning of each school year. In Table 4, the control mean column provides the students' progress in standard deviation (SD) over an academic year in the control group. 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, but their motivation index did not change. Since we expect knowledge and skills to increase during a school year but not necessarily student

[27] We randomized volunteer teachers into control and treatment groups, and tracked them over the three years. At any given year, we identify 'treatment students' as students currently taught by a treatment teacher, and 'control students', students currently taught by a control teacher. The randomization strategy is explained in detail in section 2.1

[28] 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 effect in the regression.

motivation, these results suggest that our measures are informative. We illustrate this further in Appendix Table A4, where we provide a set of correlations between the different indices and some student characteristics (gender and grade level). For instance, knowledge (resp. skills) increases on average by 0.2 SD (resp. 0.4 SD) when students progress from grade 4 to grade 5[footnote although questionnaires are specific to grade level, most items are the same across grades – reason why we observe a progression.]. Inversely, motivation decreases with age[29], a well-known dynamic for motivation [idéalement référence]. Girls have higher skills and lower knowledge. Skills and knowledge are correlated, but they clearly measure different dimensions (corr = 0.501). The correlation between motivation on the one hand and skills and knowledge on the other is negligible.

3.2 Estimation

In Table 4, column (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, as well as in the following tables, we use OLS regressions, with robust standard errors. All observations are weighted by randomization probabilities and standard errors are clustered at the strata level. We estimate the following equation:

$$y_{it} = \beta_0 + \beta_1 \cdot T_{it} + \boldsymbol{X}_{it} \cdot \boldsymbol{\beta_2} + \nu_{it}$$

With y the outcome of interest (knowledge, skills or motivation) of student i in year t, T is the treatment status of student i in year t, and $\boldsymbol{X}$ a set of control variables at the student level. In the first column of Table 4, we only control for grade level[30], whereas in the second column, we also control for baseline scores, which increases the precision of the estimates. Also, because we are testing for a large number of treatment parameters, we provide the q-value for the false discovery rate in brackets (Anderson, 2008): this can be interpreted as a p-value, robust to multiple hypothesis testing.

By design, students are not directly randomly allocated to control and treatment groups, only teachers are.[31] This design creates the possible risk that, post randomization, treatment teachers are assigned to better or worst students than control teachers. In Table A2 in the Appendix, we verify that it is not the case by comparing results

[29] The motivation questionnaire is the same across grades.
[30] Student performance questionnaires are grade specific.
[31] See section 2.1 for more details

at the beginning of year 2 and 3 (baseline) between students of control and treatment teachers. All indexes are balanced at both baselines i.e. at beginning of each school years[32]. Anecdotally, while volunteer and peers teachers were significantly different (cf. Table 2), both teach to relatively similar students, indicating no selection at the student level.

3.3 Impacts on Students

Impacts on students are given in Table 4. At the end of year 2 – the year during which the teacher received her final year of training – students in the treatment slightly outperform students in the control group. After controlling for baseline results (column 2), the impact is modest (+ 11 SD) and significant at 5%. The result is however not significant when accounting for multi-hypothesis testing but close to acceptable standards (q-value 16.4%). Besides, scientific skills and motivation are unaffected in year 2 while both dimensions were the prime objective of the program.

One year after the end of the training in year 3 (same teachers but different students), the impact on knowledge faded out and scientific skills remained unaffected. Our results even suggest that students motivation is negatively affected (-0.1 SD). The results is negative even when controlling for multi-hypothesis testing.

To get a sense of how robust this negative motivation effect is, we create sub-components of the motivation index using a Principal Component Analysis (PCA).Following the Kaiser criterion, we retained all the components with an eigenvalue greater than one (Kaiser (1960)). This gave us three main components, from which we create a simple averaged index of the (normalized) variables strongly loaded on each factor. Those three sub-dimensions are balanced at baseline (cf. Table A6) and have a relatively high Cronbach Alpha[33]. We label the three sub-dimensions "I like science", "Scientific mindset" and "Science is easy".[34]

Table 5 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. In survey year 3 (bottom panel), the three coefficients are negative (between -0.4 SD and -0.09 SD) in both columns, and very

[32]And there is no differential attrition (cf. Table A3 in the Appendix)

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

[34]The details of those new indexes are in the Appendix Table A5.

Table 4: Impacts on students scores

			Treatment v. Control	
	Obs.	Control	(1)	(2)
Year 2				
Endline knowledge	2,694	0.737	0.123*	0.105**
			(0.070)	(0.050)
			[0.394]	[0.164]
Endline skills	2,694	0.542	0.015	0.021
			(0.053)	(0.048)
			[1.000]	[0.790]
Endline motivation	2,686	-0.071	-0.040	-0.035
			(0.049)	(0.036)
			[0.709]	[0.511]
Year 3				
Endline knowledge	2,489	0.514	0.040	0.016
			(0.086)	(0.060)
			[1.000]	[1.000]
Endline skills	2,489	0.374	-0.017	-0.024
			(0.071)	(0.049)
			[1.000]	[1.000]
Endline motivation	2,488	-0.051	-0.123***	-0.092***
			(0.032)	(0.024)
			[0.003]	[0.003]
Number of clusters			24	24
Controlling for baseline variables			N	Y

The table shows difference between the treatment and control group (column (1) and (2)) in terms of the exposure to training program by the Maisons. 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 and in column *(2)* we add baseline scores. All observations are weighted by sampling probabilities. Standard errors are clustered at the strata level and given below the regression coefficients in parenthesis. $p<0.01$ ***, $p<0.05$ ** $p<0.1$ *

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. Similarly, the negative effect on the global motivation index in year 3 is observed in each of the three districts ([I think F-test is enough – add F-tests results]).

Table 5: Impacts on the motivation of students

		Treatment v. Control		
	Obs.	Control	(1)	(2)
Year 2				
I like science	2,686	-0.095	-0.018	-0.016
			(0.034)	(0.026)
			[1.000]	[1.000]
Scientific mindset	2,685	0.055	-0.031	-0.026
			(0.028)	(0.023)
			[1.000]	[1.000]
Science is easy	2,685	0.011	-0.019	-0.025
			(0.030)	(0.027)
			[1.000]	[1.000]
Year 3				
I like science	2,488	-0.059	-0.065**	-0.043**
			(0.024)	(0.017)
			[0.009]	[0.013]
Scientific mindset	2,488	0.043	-0.076***	-0.067***
			(0.017)	(0.023)
			[0.001]	[0.009]
Science is easy	2,487	-0.036	-0.073***	-0.087***
			(0.026)	(0.022)
			[0.009]	[0.003]
Number of clusters			24	24
Controlling for baseline variables			N	Y

The table shows difference between the treatment and control group (column (1) and (2)) in terms of the exposure to training program by the Maisons. 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 and in column *(2)* we add baseline scores. All observations are weighted by sampling probabilities. Standard errors are clustered at the strata level and given below the regression coefficients in parenthesis. p<0.01 ***, p<0.05 ** p<0.1 *

Finally, we investigate the heterogeneity of the training effect. We find little. The first row in Appendix Table A7, "T*H", contains the estimate of the differential effect

of the treatment with respect to observed characteristic "H". The effectiveness of the training on students scores does not vary with student initial scores, student gender, teacher initial training in sciences nor with the teacher gender.

To summarize, we observe a small but barely significant positive 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. Besides, in year 3, we find robust and significant negative effects on students' motivation in science. To better understand the plausible mechanisms of such results on students, we now analyze our results from the teacher surveys. We first present the impacts on teacher in section 4 and then discuss their relevance to understand our results on students in section 5.

4 Impact on Teachers

4.1 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 their declared practices in class in terms of hours of science, number of topics covered, frequency of conducting (hands-on) science experiment in class[35], etc. We refer to these as our quantitative measures of teacher practices. In years 2 and 3, we also ask teachers about more qualitative aspects of their work in order to capture whether or not the teachers have understood implemented the inquiry-based principles in class. Specifically, we define five topics related to inquiry-based teaching: introduce a scientific problem, formulate hypothesis, link models and observations, frame student's vision, evaluate students. For each topic, we asked teachers what they did in class; and then, according to them, how important each of those aspects was for teaching science. For every dimension, questions are answered on a Likert scale. We build the inquiry-based learning index by taking the average of the items; this index is then centered and normalized in the control group.[36]

[35]We distinguish conducting a science experiment (i.e. the teacher conduct an experiment in front of the class) and conducting a hands-on experiment (i.e. students actively participated in the experiment). This is an important distinction for the inquiry-based method as it generally requires that students actively participate in designing and conducting the experiments in class

[36]We also submitted a questionnaire eliciting teacher's vision of science at the end of the third year. The treatment did not affect those at all, and they do not correlate with student's performance, so

There is some attrition in those surveys: out of 134 teachers initially enrolled in the experiment, 129 answered at least some of the questions related to the first year, 119 for the second year and 102 for the last year. There is no evidence of differential attrition (see table A3 and 1).

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

4.2 Training effects on topics covered

[introduce this section]

In France, the primary school science curriculum can be divided into eight topics ("Earth and the Universe", "Energy", "Technical objects", etc.). Much of the teacher training content consisted in defining and discussing a class sequence on a specific question from one of those topics. For instance, one training center used medieval machinery to illustrate levers, a sequence that belongs to the topic "Technical objects". The content of this sequence covered during the training can easily be re-used in class. Each year, we have collected information on the topics covered in each of the training center within each region (the three regions are divided into nine local centers). Information on three centers-years is missing: the sample is therefore composed of 15 center-year observations that generate variation in the topics covered during training.

In our teacher surveys, we asked each teacher to list the topics that they covered with their students during the year, so that we have information for each of the three years of the experiment. A simple measure of how much the training sessions influenced teaching consist in estimating if a topic covered during the training was subsequently covered in class more often than other topics. To that aim, we re-organize the dataset so that each observation is at the teacher, topic and year level. For each observation, we create a dummy variable taking one if the topic was covered in the teacher's training center in a given year (training topic) and zero otherwise, and another dummy taking one if the topic was covered in class in a given year (class topic) and zero otherwise. We then use this data to measure how training topics affects class topics in a given year and in subsequent years. The training topic dummy variable of the topics covered in training are defined for control teachers too, but training topic should only affect

we do not present this data here.

the practices of treated teachers.

Table 6 presents the relationship between topics covered in class and topics covered during the training using the following difference-in-differences model:

where *class topic* is the dummy variable taking value 1 when the ...

The training topic dummy captures the correlation between topics covered in class and during the training in the control group. It should be zero since the control group did not benefit from the training program. The interaction term captures the same correlation in the treatment group which may be positive if teachers use the training material in their class. Columns (1) and (3) show same-year relationship, whereas the other columns verify if training topics from a year influenced class topics in subsequent years.

The constant in Column (1) of Table 6 indicates that slightly more than half the topics were covered in class in year one, on average in the control group (some more than others – the most popular being "Matter", "Health and Human body" and "Technical objects"). The interaction terms indicate that the probability that any topic is covered in class increases by 23 percentage points if it has been covered during training that same year. As expected, the training topics covered during the training do not affect the ones 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 the same year.[37] This is a very clear indication that participating to the training influenced the teacher practices: they likely used the material developed during the training and translated them into their own science sessions in class.

Interestingly, this relationship 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 increases the likelihood that they are covered in class by 12 and 9 points respectively (less significantly so). This finding is compatible with a model where training only affects teacher practices in the short run: they adopt training material when trained, but do not use it much afterwards. Given the high level of reported satisfaction over the training that we measured, it is unlikely to happen because teachers would be unconvinced by the material provided to them. It is more likely that settled practices are difficult to change in the medium-run.

[37] In year 2, the coefficient on the treatment dummy also indicates that fewer different topics were covered in the treatment group.

Table 6: Effects of training topics on class topics

	Class topic Y1	Class topic Y2		Class topic Y3	
	(1)	(2)	(3)	(4)	(5)
Training topic Y1 × Treatment	0.232***	0.150		0.118**	
	(0.048)	(0.089)		(0.047)	
Training topic Y2 × Treatment			0.251***		0.088
			(0.060)		(0.066)
Training topic Y1	0.031	-0.035		-0.050	
	(0.038)	(0.077)		(0.034)	
Training topic Y2			-0.083*		-0.048
			(0.042)		(0.086)
Treatment	-0.026	-0.079**	-0.088**	-0.024	-0.009
	(0.019)	(0.076)	(0.036)	(0.029)	(0.032)
Constant	0.548***	0.562***	0.568***	0.547***	0.543***
	(0.025)	(0.027)	(0.033)	(0.030)	(0.036)
Obs.	1,024	944	944	808	808

The table shows difference between the treatment and control group (column (1) and (2)) in terms of the exposure to training program by the Maisons. 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 and in column *(2)* we add baseline scores. All observations are weighted by sampling probabilities. Standard errors are clustered at the strata level and given below the regression coefficients in parenthesis.
p<0.01 ***, p<0.05 ** p<0.1 *

4.3 Training effects on reported practices

We separate the analysis of practices into quantifiable measures (such as hours of science) in Table 7 and qualitative practices in Table 8. Table 7 shows separate regressions, for each of the three years, of several reported practices on a treatment dummy. Column (1) includes no controls. To account for the fact that treatment and control teachers are imbalanced with respect to their hours of science reported at baseline (i.e. referring to the school year before the start of the training), we include that variable as a control in column (2). Specification is otherwise the same as for students and we also provide q-values for the false discovery rate in brackets.

The training seems to influence the weekly of hours of science in each of the three years, but more strongly so in years 2 and 3. The related item ("do you feel you teach enough science") is also affected. Based on column (2), the treatment increases weekly hours of science by 0.175 and 0.225 hours in each year respectively, starting from a control group average of almost 1.5 hours per week, thus a 12-17% increase. Trained teachers are also more likely to conduct hands-on science experiment (+7pp, barely significant), as expected from a inquiry-based intervention but only so in the second year. We will interpret this further in section 5 below.

However, with a starting sample of 134 teachers, we do not have much power, especially if we account for multiple testing with such a long list of outcomes. In the uncontrolled specification, the effect on hours is clearly significant; but with the baseline control for hours of science at baseline, the coefficient drops and no longer passes the multiple testing adjustment.

Table 7: Impacts on teacher quantitative practice

		Treatment v. Control		
	Obs.	Control	(1)	(2)
Year 1				
Feels teach enough sci.	127	-0.041	0.283	0.201
			(0.176)	(0.191)
			[0.212]	[0.512]
Weekly h. of sci.	121	1.537	0.124*	0.009
			(0.070)	(0.075)
			[0.171]	[1.000]
No. of sci. subj.	129	4.455	0.231	0.047
			(0.244)	(0.237)
			[0.356]	[1.000]
Freq. of pupils manip.	129	0.642	0.058	0.055
			(0.050)	(0.050)
			[0.356]	[0.512]
Freq. of teach manip.	128	0.290	0.024	0.042
			(0.054)	(0.050)
			[0.679]	[0.530]
Year 2				
Feels teach enough sci.	119	-0.051	0.483**	0.367
			(0.217)	(0.216)
			[0.121]	[0.367]
Weekly h. of sci.	119	1.429	0.229**	0.175**
			(0.084)	(0.077)
			[0.081]	[0.347]
No. of sci. subj.	119	4.420	-0.339	-0.438
			(0.309)	(0.285)
			[0.356]	[0.429]
Freq. of pupils manip.	119	0.652	0.083**	0.073*
			(0.040)	(0.038)
			[0.132]	[0.367]
Freq. of teach manip.	117	0.349	0.016	0.035
			(0.065)	(0.063)
			[0.752]	[0.769]
Year 3				
Feels teach enough sci.	101	0.003	0.420**	0.288*
			(0.159)	(0.158)
			[0.081]	[0.367]
Weekly h. of sci.	102	1.349	0.302***	0.225**
			(0.105)	(0.100)
			[0.081]	[0.347]
No. of sci. subj.	102	4.260	0.066	0.001
			(0.225)	(0.210)
			[0.752]	[1.000]
Freq. of pupils manip.	102	0.644	0.042	0.036
			(0.044)	(0.041)
			[0.356]	[0.530]
Freq. of teach manip.	102	0.382	-0.064	-0.065
			(0.063)	(0.061)
			[0.356]	[0.512]
Number of clusters	28		24	24
Controlling for baseline variables			N	Y

The table shows difference between the treatment and control group (column (1) and (2)) in terms of the exposure to training program by the Maisons. Column *Obs.* gives the number of students surveyed, *Control* the

Similarly, we find small treatment effects on our the Inquiry-based learning index, which captures how teachers have integrated inquiry-based principles in their science sessions. In the Appendix Table A8, we report regressions of each of the practice categories ("introduce a scientific problem", "formulate hypothesis", etc.) on a treatment dummy, separately for years two and three. Few effects are statistically significant; however, they are almost systematically positive, with relatively high point estimates. In order to leverage that feature, we aggregate all qualitative items into one index, separately per year. The conceptual meaning of such an aggregate of different pedagogical dimensions is questionable. From a statistical point of view, however, the aggregation helps determining if the intervention affected anything, even though we lack statistical power to tell which one(s) precisely. Note that this index of practices is correlated with the number of hours of science (correlation 0.25) and with the frequency of student manipulation (correlation 0.31) (see supra Table A4): the same teachers that report following the inquiry-based method teach more science and conduct more hands-on science experiments. However, again from Table A4, observed teacher characteristics are uncorrelated with the inquiry-based learning index.

Table 8: Impacts on teacher qualitative practice

			Treatment v. Control	
	Obs.	Control	(1)	(2)
Year 2				
Index of practice and judgement items	-0.025	-0.054	0.120	0.120
			(0.143)	(0.147)
			[0.258]	[0.269]
Year 3				
Index of practice and judgement items	0.022	-0.054	0.304**	0.288**
			(0.116)	(0.130)
			[0.032]	[0.081]
Number of clusters			24	24
Controlling for baseline variables			N	Y

The table shows difference between the treatment and control group (column (1) and (2)) in terms of the exposure to training program by the Maisons. 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 and in column *(2)* we add baseline scores. All observations are weighted by sampling probabilities. Standard errors are clustered at the strata level and given below the regression coefficients in parenthesis. $p<0.01$ ***, $p<0.05$ ** $p<0.1$ *

Table 8 shows the treatment impact on this index of reported practices. The training program does not significantly increase the index in year 2, although the point estimate is positive (+0.12 SD). In year 3, one year after teachers completed the training program, the aggregate index's impact is larger (+0.3 SD), and statistically significant even when accounting for multi-hypothesis adjustment (q-value = 8.1%).

5 Discussion

Summing up, we find evidence that the training program affects teachers practices: treatment teachers re-use in class the topics covered during the training sessions, they declare spending more time teaching science and they conduct more scientific experience in class. Further, our qualitative measures show that teachers are more likely to use the inquiry-based pedagogical approaches in class. However, in spite of significant statistical power, these improved practices translate into only small impacts on student knowledge, limited to year 2; and in year 3 (after the training was over), student motivation was lower in the treatment group.

One likely explanation is that the change in pedagogy induced by the program is not strong enough to have very visible impacts on student performance. To investigate the relationship between teacher practices and student performance, Table 9 provides the correlation between some of our teaching measures of practices and student performances in the control group. The upper panel of Table 9 shows the correlations between our measures of students' progress over the year (change between baseline and endline measure) and our measures of teaching practices (pooling years 2 and 3). Although they tend to be positive, the correlations are very small and generally not significant. Taken at face value, these correlation would suggest that the +0.288 SD impact on qualitative index found in year 3 (cf. Table 8) would only translate into a +0.024 SD effect on student knowledge, far from our detection power. Similarly, the 0.175 additional hours of science found in year 2 would translate into a +0.013 SD in terms of student performance, again far from our detection power. To put it otherwise, only massive transformations of our measures of teaching practices—typically a 1 SD impacts on the qualitative index or a 1 hour increase in weekly science—could have generated detectable impacts on student performances. Interestingly, in one of the experiment that can be directly compared to our study and that finds positive impacts on student performances, Meyers et al. (2016) do find much larger impacts on

self-declared teacher practices e.g., +0.73 SD on an their inquiry-based learning index or + 1.441 SD on their technology integration index.

Of course, our measures of teaching practices could be very noisy, as often found when assessing the external validity of self-reported measures of teacher practices (Seidel and Shavelson, 2007).[38] This would mechanically attenuate the relationship between teaching practices and student performances. Yet, the middle panel of Table 9 shows that our measures are generally well correlated with each other, which would not happen if they mostly contained noise.[39] Further, the qualitative practice index is based on sub-scales that all have high Cronbach alphas. Most importantly, we do find that our measures of practices are influenced by the treatment, which again should not happen if they did not capture a signal.

Another interpretation is that the index of reported practices is improved among treated teachers as a result of desirability bias. Teachers may not change their actual teaching content or methodology, but answer our questions based on what they heard in the training sessions. A more charitable version of this interpretation would be that teachers do change their *intellectual* understanding of how to teach science (hence our qualitative impacts), but do not modify their *actual* teaching practices. This would therefore mean that the principles of the inquiry-based pedagogy were understood but the training program was not influential. This interpretation might be particularly meaningful for the inquiry-based index whose impact is larger in year 3 while the impact on student performance is null, suggesting a decorrelation between declared qualitative practices and actual practices. [Topics ? Student manipulation ? revoir practice vs judgement. + Y2 vs Y3]

A final possibility is that the inquiry-based pedagogy does not structurally improve student performance in science. While it may affect teacher practices in the short-run (through more science, more hands-on experiment and better self-declared qualitative measures), those change are short-lived and have little impacts on the student learning process. The comparison between year 2 and year 3's impacts give credibility to this interpretation. The small positive effect on student knowledge is present in year 2,

[38] Stronger correlations are found when teaching practices are assessed through classroom observation. When assessed through (arguably more reliable) classroom observations, the correlation of normalized practices with normalized student test scores is to the order of 0.2-0.5, somewhat higher than in Table 9 (e.g. Kane et al. (2011), Allen et al. (2013), Grossman et al. (2014)). Yet, to our knowledge such instruments have not been developed for inquiry-based teaching in science.

[39] We can note (bottom panel) that they are not correlated with teacher characteristics, but this is not surprising, just as much as teacher fixed effects in general are hard to predict with such characteristics.

while the teachers are still attending the training sessions, but disappears in year 3, whereas motivation that was hardly affected initially, decreases strongly in the treatment group in year 3. Since the impact on the weekly hours of science is constant in year 2 (+0.175 h/week) and year 3 (+0.225 h/week) (cf. Table 7), these results suggest that the quality of the sessions have significantly decline in year 3 when the teachers ceased to benefit from the support of the Maisons. Although we lack precise measures to fully account for this deterioration of the teaching quality in year 3 among treatment teachers, the frequency of hands-on experiment, a fundamental component of the inquiry-based pedagogy, does decline in year 3 to a non significant +3.6 pp, while it was twice larger in year 2 (cf. Table 7.) Similarly, we also noted (cf. Table 6) that the material and topics covered during the training were then used in class in year 2, but much less so in year 3. This indicates that actual pedagogical change may have happened during the training, but it faded out rapidly when support disappeared. As a matter of fact, some research suggest that inquiry-based teaching may be inefficient 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 that third year also, student motivation for science decreased in treatment classes by a small but significant 10% of a S.D. approximately. One interpretation is that there is something in inquiry-based teaching that lowers students' interest for science, especially if it is not accompanied by the kind student activities that can be setup with the support of professional trainers. This would be even more likely if the training was not sufficiently well structured for teachers to implement this pedagogical approach in a complete and consistent way when on their own. The qualitative analysis of the video footage of some of the training sessions points to this possibility: Our analysis suggests that the sessions are not explicitly organized in terms of the different blocks of pedagogical knowledge, and are almost never related to the teaching practices of the trained teachers. This may explain our findings. It remains unknown, however, if actual day-to-day practices would be affected, even if pedagogical content was transmitted more systematically.

These three explanations of our results are probably not mutually exclusive. For instance, while there is no doubt that the changes in teacher practices were too modest to generate impacts on student performance (explication 1), it is also true that some of the new practices adopted by teachers were detrimental when implemented without the support of the Maisons (explication 3). Similarly, it is likely that part of the

self-reported measures of inquiry-based quality only capture *declared* practices and not *actual* practices (explication 2), like the year 3 results on teachers and students suggest. In any case, in our context which closely mimic the one that we would have observed in absence of the experiment, the training program failed to provide tools that can be readily used by teachers to improve the quality of their teaching.

Table 9: Correlations between teaching practices and student outcomes

	Qual. practice index	Freq. of pupils manip.	Weekly hours of science
Student scores			
Progress/knowledge	0.085	0.138	0.074
	[0.411]	[0.181]	[0.474]
Progress/skills	0.003	-0.009	0.095
	[0.975]	[0.930]	[0.357]
Progress/motivation	0.117	0.002	0.102
	[0.257]	[0.832]	[0.324]
Teaching practices			
Qual. practice index	1	0.308	0.246
	.	[0.002]	[0.016]
Freq. of pupils manip.	0.308	1	0.146
	[0.002]	.	[0.156]
Weekly h. of science	0.246	0.146	1
	[0.016]	[0.156]	.
Teacher characteristics			
Holds science degree	0.031	0.021	-0.029
	[0.765]	[0.839]	[0.784]
Teaching exp.	-0.189	-0.029	-0.061
	[0.067]	[0.778]	[0.555]
Gender	0.164	0.119	0.036
	[0.111]	[0.249]	[0.729]
Observations	96	96	96

6 Conclusion

Despite the potential large social benefits of teacher training programs on teacher productivity, the literature on training program remain scarce, inclusive and heterogeneous. Among the 14 relevant experimental research we identified to be fairly comparable to our setting, only five studies found some positive results. Two of these training

programs are inspired by inquiry-based principles, while three are about phonological awareness in primary school or kindergarten. These few success stories in the literature suggest that intensive and well-executed training program, based on validated pedagogical approaches, might be instrumental in improving student performances. In this paper, we propose to test this claim using a large randomized experiment.

Our study is (i) well powered—second most powered study after the recent large scale RCT conducted in China (Loyalka et al., 2019), (ii) based on a widely recognized pedagogical strategy (i.e, inquiry-based), (iii) based on a very 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 compared to the usual volume of training in science received by our sample before randomization ($\approx$ 2 hours), and (iv) almost perfectly executed—68 additional hours of training was provided to the treatment group. high level of compliance. 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 if good practices carried out one year after the end of the training program, a fundamental feature to consider when calculating the cost-benefit of such intervention.

We find evidence that the training programs slightly modified teacher practices: teachers spend slightly more time teaching science (about a fifth or an hours more or about a 15% increase), they conduct more hands-on experiment and the self-declared quality index suggests that they are more likely to implement the inquiry-based principles in class. Yet, in spite of these impacts at the teacher level, we find no convincing evidence that these translate into positive impacts on student performance. While we do find small and barely significant impacts on knowledge in year 2 (last year of the training), the impact disappear in year 3, one year after the end of training program. In addition, we find no impacts on skills while this was 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 quality of the session sessions have declined one year after the end of training program supports. Yet, our measure at the teachers level fails to fully capture this deterioration. We do observe a decline in the number of hands-on experiment but our measure of inquiry-based quality still reports positive impacts in year 3. This suggests that other mechanisms pertaining

to the teacher practices and unobserved by our teacher level measures are driving the results on the students. Progress in capturing these hidden teacher practices need to be made to fully account for the impacts of such program.

Does this conclusion apply to any training program inspired by inquired-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 incomparable to most other training programs. Second, as a direct result of point 1, such intensive and specific training are likely to attract teachers that are interested in the topic (here science) and have sufficient time to devote to the training. We do find such selection in our context and we believe that this is an under-reported issue in the literature. To be sure, we cannot know whether or not such selection has affected our results but this is an important concern for the overall efficiency of training programs. Third, even when teacher practices are modified considerably—like for instance in (Meyers et al., 2016)—the treatment effect remain modest for an organizational cost (finding and paying for substitute teachers) that has proven to be high in our case. Taken together, our results, together with a myriad of other related works, raise important doubts that training programs are viable tools to improve overall teacher productivity.

References

Allen, J., Gregory, A., Mikami, A., Lun, J., Hamre, B., and Pianta, R. (2013). Observations of effective teacher-student interactions in secondary school classrooms: Predicting student achievement with the classroom assessment scoring system-secondary. *School psychology review*, 42(1):76–98.

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.

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.

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.

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.

Grossman, P., Cohen, J., Ronfeldt, M., and Brown, L. (2014). The test matters: The relationship between classroom observation scores and teacher value added on multiple types of assessment. *Educational Researcher*, 43(6):293–303.

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

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., Taylor, E., Tayler, J., and A., W. (2011). Identifying effective classroom practices using student achievement data. *The Journal of Human Resources*, 46(3):587–613.

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

Kirschner, P., Sweller, J., and R., C. (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.

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*.

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.

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.

Seidel, T. and Shavelson, R. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research*, 77:454–499.

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	Content	Level	Sample	Intensity	ES (sd)	Sig.
Angrist and Lavy (2001)	2001	Israel	NE	m; r	primary	31/850	11 d	.3-.6	s
Borman et al. (2007)	2007	USA	RCT	r	K-G2	35/4180	95 h, 1 y	.21-.33	s
Garet et al. (2008)	2008	USA	RCT	r	primary	90/5000	48 h+60 h, 2y	0.03	ns
Machin and McNally (2008)	2008	UK	NE	r	G1	14000/1.6m	intensive	0.086	s
Gersten et al. (2010)	2010	USA	RCT	r	G1	81/468	20 h, 6 m	0.225	ns
Kim et al. (2011)	2011	USA	RCT	r	G6-G12	103/2726	46 h	0.046	ns
Randel et al. (2011)	2011	USA	RCT	m	G4-G5	67/4700	self-taught	0.096	ns
Allen et al. (2011)	2011	USA	RCT	g	G4-G5	76/2237	1y	0.22	s
Harris and Sass (2011)	2011	USA	NE	m;r	primary	1,031/487,000	NA	0	ns
Campbell and Malkus (2011)	2011	USA	RCT	m	G3-G5	36/24759	very	0.049	ns
Newman et al. (2012)	2012	USA	RCT	m;s	G4-G8	82/3000	10 d, 2y	0.05	s
Bos et al. (2012)	2012	USA	RCT	r	G6-G8	52/18180	14 days	0.01	ns
Sirinides et al. (2018)	2018	USA	RCT	r	G1	7000	Very	0.47	s
Meyers et al. (2016)	2013	USA	RCT	m;r	G7-G8	60/3072	240 h,2 y	0.13-0.17	s
Gentaz et al. (2013)	2013	France	RCT	r	K	56/2398	42 h	0	ns
Rimm-Kaufman et al. (2014)	2014	USA	RCT	g	G3-G5	24/2904	20 h	0	ns
Bouguen (2016)	2017	France	NE	r	K	118/4345	intensive	0.153	s
Loyalka et al. (2019)	2019	China	RCT	m	G7-G9	300/16661	15 d	0	ns

The tables list the rigorous (RCT or rigorous non-experimental methods) and recent studies (in last 20 years) on teacher training program we identified in the literature. To be included, a study should be relatively recent (post 2000), have a rigorous causal approach – experimental (RCT) or Non-experiemntal (NE) method – and be sufficiently powered (>.30 sd ex-ante power). In the table, we provide the reference to the study (Author), the year of publication of the article or report, the country, the type of design (either RCT or NE), the content of the training program – general (g), maths (m), reading (r) – the grade of the target teachers/students (level,) the sample size (number of clusters/number of individuals), the intensity – h for hours, d for days, m for months and y for years– the effect size (ES) expressed in standard deviation and the significance (*s* for significant at minimum 10% and *ns* for not significant).

Table A2: 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	0.009	4,495	-0.064	0.035
			(0.063)			(0.052)
Baseline skills	2,689	-0.020	-0.060	4,495	-0.132	0.078
			(0.072)			(0.068)
Baseline motivation	2,672	-0.009	0.012	4,461	0.079	-0.081**
			(0.046)			(0.032)
Grade 3	3,053	0.223	0.070	5,207	0.248	0.014
			(0.058)			(0.058)
Grade 4	3,053	0.330	-0.047	5,207	0.394	-0.091*
			(0.052)			(0.047)
Grade 5	3,053	0.447	-0.023	5,207	0.358	0.077
			(0.069)			(0.074)
Late student	3,053	0.087	-0.024*	5,207	0.073	-0.000
			(0.013)			(0.007)
Female student	3,053	0.474	0.025	5,207	0.498	-0.010
			(0.020)			(0.013)
Year 3						
Baseline knowledge	2,529	-0.027	-0.010	3,971	-0.039	0.007
			(0.076)			(0.061)
Baseline skills	2,529	-0.039	-0.105	3,971	-0.143	0.045
			(0.093)			(0.090)
Baseline motivation	2,516	0.012	-0.057	3,951	-0.001	-0.018
			(0.061)			(0.032)
Grade 3	2,883	0.235	0.102	4,408	0.260	0.031
			(0.066)			(0.066)
Grade 4	2,883	0.316	-0.019	4,408	0.423	-0.117*
			(0.066)			(0.067)
Grade 5	2,883	0.448	-0.083	4,408	0.317	0.086
			(0.072)			(0.079)
Late student	2,826	0.083	-0.023	4,351	0.110	-0.039**
			(0.014)			(0.019)
Female student	2,826	0.476	0.013	4,351	0.511	-0.027*
			(0.021)			(0.013)
Number of clusters			24			24

The table shows

Table A3: Attrition - student level

		Treatment v. Control		
	Obs.	Control	(1)	(2)
Year 2				
Student attrition in endline	2,935	0.083	-0.004	-0.006
			(0.009)	(0.009)
... absent students	2,935	0.046	-0.000	-0.001
			(0.008)	(0.008)
... non-participating students	2,935	0.037	-0.004	-0.005
			(0.008)	(0.008)
Year 3				
Student attrition in endline	2,724	0.083	0.006	0.003
			(0.015)	(0.015)
... absent students	2,724	0.047	-0.000	0.000
			(0.011)	(0.011)
... non-participating students	2,724	0.036	0.007	0.003
			(0.010)	(0.010)
Number of clusters			24	24
Controlling for level fixed effects			N	Y

The table shows

Table A4: Correlations between indexes and student characteristics

	Baseline knowledge	Baseline skills	Baseline motivation
Baseline scores			
Baseline knowledge	1	0.501	0.059
	[0.000]	[0.000]	[0.452]
Baseline skills	0.501	1	-0.006
	[0.000]	[0.000]	[1]
Baseline motivation	0.059	-0.006	1
	[0.452]	[1]	[0.000]
Endline scores			
Endline knowledge	0.551	0.463	0.076
	[0.000]	[0.000]	[0.034]
Endline skills	0.439	0.603	0.022
	[0.000]	[0.000]	[1]
Endline 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 shows

Table A5: 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

Table A6: Balance checks - sub-components of the motivation index

	Treatment v. Control			**Volunteer v. Peer**		
	Obs.	Control	(1)	Obs.	Control	(2)
Year 2						
Baseline: I like science	2,670	0.002	0.022	4,459	0.064	-0.050*
			(0.030)			(0.024)
Baseline: Scientific mindset	2,658	-0.012	-0.055	4,439	-0.002	-0.041
			(0.036)			(0.027)
Baseline: Science is easy	2,660	-0.008	0.036	4,437	0.043	-0.031
			(0.024)			(0.021)
Year 3						
Baseline: I like science	2,516	0.026	-0.035	3,951	0.031	-0.024
			(0.042)			(0.023)
Baseline: Scientific mindset	2,507	-0.030	-0.065*	3,938	-0.046	-0.020
			(0.033)			(0.027)
Baseline: Science is easy	2,511	-0.008	0.011	3,943	-0.033	0.032
			(0.039)			(0.020)
Number of clusters			24			24

The table shows

Table A7: Treatment heterogeneity

		Heterogeneity:											
		(A)			(B)			(C)			(D)		
		Student's knowledge above median			Female student			Teacher's main diploma is in science			Female teacher		
		Obs.	(1)	(2)	Obs.	(1)	(2)	Obs.	(1)	(2)	Obs.	(1)	(2)
Year 2													
Endline knowledge	T*H	2,415	-0.128*	-0.061	2,694	0.102	0.009	2,685	0.100	0.106	2,694	-0.022	-0.028
			(0.070)	(0.063)		(0.102)	(0.098)		(0.106)	(0.085)		(0.138)	(0.128)
	T		0.223***	0.143**		0.076	0.102		0.055	0.033		0.139	0.126
			(0.074)	(0.062)		(0.082)	(0.066)		(0.104)	(0.075)		(0.108)	(0.112)
	H		0.896***	0.115		-0.138**	-0.023		-0.160	-0.156*		0.093	0.067
			(0.051)	(0.080)		(0.064)	(0.059)		(0.097)	(0.081)		(0.099)	(0.082)
Endline skills	T*H	2,415	-0.037	-0.014	2,694	0.105	0.032	2,685	0.046	0.065	2,694	-0.110	-0.060
			(0.062)	(0.054)		(0.075)	(0.073)		(0.107)	(0.072)		(0.120)	(0.093)
	T		0.055	0.036		-0.040	0.002		-0.011	-0.016		0.093	0.064
			(0.057)	(0.058)		(0.053)	(0.049)		(0.074)	(0.057)		(0.101)	(0.084)
	H		0.600***	0.294***		0.126*	0.116*		-0.069	-0.057		0.107	0.038
			(0.057)	(0.062)		(0.064)	(0.057)		(0.077)	(0.058)		(0.092)	(0.072)
Endline motivation	T*H	2,409	-0.017	-0.060	2,686	-0.087	-0.080	2,677	-0.062	-0.066	2,686	0.163*	0.141*
			(0.077)	(0.064)		(0.088)	(0.076)		(0.097)	(0.077)		(0.089)	(0.078)
	T		-0.026	-0.001		0.005	0.006		-0.000	0.005		-0.155**	-0.135**
			(0.070)	(0.060)		(0.059)	(0.046)		(0.083)	(0.064)		(0.071)	(0.062)
	H		0.165**	0.103*		-0.055	-0.044		0.060	0.054		-0.122*	-0.133**
			(0.065)	(0.050)		(0.067)	(0.054)		(0.077)	(0.060)		(0.068)	(0.058)
Year 3													
Endline knowledge	T*H	2,216	0.012	0.013	2,435	-0.208**	-0.116*	2,477	-0.160	-0.117	2,489	0.039	0.044
			(0.112)	(0.099)		(0.075)	(0.060)		(0.142)	(0.091)		(0.200)	(0.126)
	T		0.063	0.026		0.151	0.085		0.135	0.085		0.013	-0.015
			(0.095)	(0.077)		(0.112)	(0.077)		(0.119)	(0.089)		(0.166)	(0.089)
	H		0.794***	-0.020		0.077	0.054		0.064	0.041		-0.048	-0.009
			(0.079)	(0.081)		(0.051)	(0.036)		(0.111)	(0.085)		(0.119)	(0.071)
Endline skills	T*H	2,216	-0.016	-0.035	2,435	-0.032	-0.010	2,477	-0.077	-0.011	2,489	-0.038	0.027
			(0.084)	(0.060)		(0.086)	(0.072)		(0.132)	(0.105)		(0.162)	(0.104)
	T		0.039	0.037		0.013	-0.001		0.031	-0.020		0.009	-0.043
			(0.088)	(0.062)		(0.099)	(0.067)		(0.110)	(0.076)		(0.131)	(0.074)
	H		0.612***	0.280***		0.188***	0.082		0.014	-0.039		0.025	0.024
			(0.059)	(0.046)		(0.057)	(0.048)		(0.099)	(0.083)		(0.103)	(0.059)
Endline motivation	T*H	2,216	-0.005	-0.066	2,434	-0.070	-0.083	2,476	-0.140	-0.075	2,488	0.019	0.007
			(0.087)	(0.080)		(0.092)	(0.063)		(0.118)	(0.088)		(0.126)	(0.093)
	T		-0.129**	-0.067		-0.093	-0.058		-0.038	-0.045		-0.137	-0.097
			(0.049)	(0.046)		(0.065)	(0.041)		(0.089)	(0.061)		(0.090)	(0.061)
	H		0.211***	0.134***		-0.030	0.024		0.092	0.057		-0.057	-0.004
			(0.056)	(0.047)		(0.077)	(0.057)		(0.081)	(0.053)		(0.091)	(0.060)
Number of clusters			24	24		24	24		24	24		24	24
Controlling for baseline variables			N	Y		N	Y		N	Y		N	Y

The table shows

Table A8: Impacts on teacher qualitative practice, detailed

		Treatment v. Control		
	Obs.	Control	(1)	(2)
Year 2				
Introduce sci. problem	119	-0.054	0.341**	0.337*
			(0.155)	(0.163)
			[0.466]	[1.000]
Work on students vision	119	0.001	-0.015	-0.028
			(0.169)	(0.171)
			[1.000]	[1.000]
Evaluate students	118	-0.045	-0.085	-0.136
			(0.175)	(0.178)
			[1.000]	[1.000]
Importance of introducing sci. problem	119	-0.016	0.140	0.207
			(0.162)	(0.149)
			[0.946]	[1.000]
Importance of formulating hypothesis	119	-0.022	0.174	0.189
			(0.151)	(0.168)
			[0.946]	[1.000]
Importance of linking model and obs.	115	-0.005	-0.020	-0.038
			(0.176)	(0.169)
			[1.000]	[1.000]
Importance of student evaluation	116	0.011	0.084	0.092
			(0.146)	(0.160)
			[1.000]	[1.000]
Year 3				
Introduce sci. problem	101	0.067	0.368*	0.371*
			(0.192)	(0.214)
			[0.466]	[1.000]
Work on students vision	101	0.018	0.254*	0.230
			(0.127)	(0.138)
			[0.466]	[1.000]
Evaluate students	101	0.000	0.201	0.162
			(0.169)	(0.177)
			[0.946]	[1.000]
Importance of introducing sci. problem	100	-0.002	0.065	0.064
			(0.121)	(0.126)
			[1.000]	[1.000]
Importance of formulating hypothesis	100	-0.009	0.147	0.148
			(0.157)	(0.162)
			[0.946]	[1.000]
Importance of linking model and obs.	98	0.028	0.128	0.092
			(0.129)	(0.128)
			[0.946]	[1.000]
Importance of student evaluation	100	-0.003	0.245	0.266
			(0.227)	(0.242)
			[0.946]	[1.000]
Number of clusters			24	24
Controlling for baseline variables			N	Y