Trying to raise (low) math achievement and to promote
(rigorous) policy evaluation in Italy.
Evidence from a large scale randomized trial

Gianluca Argentin (PhD), Università di Milano Bicocca
Aline Pennisi (PhD), Ministero dell’Economia e delle Finanze, aline.pennisi@tesoro.it
Daniele Vidoni (PhD), INVALSI, daniele.vidoni@invalsi.it
Giovanni Abbiati (PhD student), Università degli Studi di Milano, giovanni.abbiati@unimi.it
Andrea Caputo (Degree), INVALSI, andrea.caputo@invalsi.it

ACKNOWLEDGEMENTS
M@t.abel is an acronym for Matematica, apprendimenti di base con e-learning. We evaluate here the most recent version of this program, carried out with EU funding - PON Istruzione 2007-2013 (I-3-FSE-2009-2). Special thanks for their support and precious advice to: Alberto Martini (who contributed to design and set up the project), Piero Cipollone, Annamaria Leuzzi (Ministry of Education), Annamaria Fichera (Ministry of Education), Alessandro Cavalli, Maria Pia Perelli d’Argenzio, Enrico Rettore, Jaap Scheerens.

BOZZA PRELIMINARE,
PER FAVORE NON DIFFONDERE
1. Introduction

In many countries the drive to improve education has triggered a season of rigorous research on what kind of instructional practices, curricula and interventions work. Italy is still lagging behind for several reasons: data on student achievement are limited and only recent, there is widespread aversion to testing and little tradition of evidence-based policy evaluation. Given the relative weakness of Italian students in international assessments on mathematics and science (i.e. IEA, TIMSS and OECD PISA) and also thanks to EU funding, there has been a recent boost in the initiatives to improve student achievements and an urge to understand their effectiveness. Educational research today clearly agrees on the fact that teachers have a fundamental influence on student results (Scheerens, 2000; OECD, 2009) and are crucial to improve school achievement (Rivkin, Hanushek, & Kain, 2005). Notwithstanding the influence of factors such as socio-economic background, family and school context, student learning is influenced by what and how teachers teach (e.g. Ma & Papanastasiou, 2006).

Italian teachers are among the oldest in Europe and they do not necessarily acquire formal teaching skills. Recruitment is centrally-based; teacher mobility is driven essentially by seniority and by the position they have acquired in provincial lists more than on school’s choice. High teacher turnover leads to segregation of the more experienced teachers in their preferred schools, and usually in those with better-performing students (Barbieri, Rossetti, & Sestito, 2011). In this context there are several issues suggesting that a general improvement of the existing staff is possible through in service training. Indeed, professional development courses (PD hereafter) can activate two important levers to increase student achievement: first, through the introduction of new pedagogical skills, especially among the older and not specifically trained teachers; second, by developing teacher communities within schools, an element often associated to effective schools (Ma & Ma, 2004; Scheerens, 2000).

Improvement for in service teachers might be sought following different approaches to PD:
helping them to understand more deeply the content they teach, or the ways students learn, but also providing alternative solutions, methods and materials to present the contents.

This study investigates the effects of a specific teacher PD program called M@t.abel. The program, supported by the Italian Ministry of Education, covers a substantial fraction of the lower secondary school curriculum. It is becoming popular among teachers and spreading around the country. M@t.abel is being eagerly promoted in the four regions of Southern Italy, thanks to European Union funding.

According to international and national testing, regions of Southern Italy show the lowest levels of math achievement. The OECD-PISA 2009 findings reveal that one student out of three is unable to properly master most elementary and routine tasks in mathematics. The ratio is only one out of ten in the North, suggesting that the overall low performance is mainly due to results in the South. Moreover TIMSS proves that Italian students perform above the international average in 4th grade in math by face a significant fall in 8th grade, placing Italy among the poor performing countries. In the 2010 INVALSI national assessment on 6th grade students, the share of correct answers in Southern Italy was on average 4 percentage points less than in Italy as a whole. Available empirical evidence suggests that the differences in student performance between Northern and Southern Italy

---

1 NOTE DELETED FOR BLIND REVIEW
2 There also is an upper secondary version of the program. However we limit this study to lower secondary school, which has homogenous curriculum.
3 OECD-PISA 2009. We refer to the share of 15-year old students with at most competence level 1 in mathematics.
4 INVALSI (Istituto Nazionale per la Valutazione del Sistema di Istruzione e Formazione) is a research institute of the Ministry of education holding the specific aim to evaluate Italy schools. Among other tasks, a recent law prescribed that the INVALSI should carry out standardized measurement on students’ ability in reading and math at various levels of education.
cumulate over time. Whilst limited in primary school, the gap gets bigger in lower secondary school.

Using a randomized control trial, we seek to detect whether the M@tabel program makes a measurable difference in promoting student achievement and attitudes and in modifying teaching practices. To our knowledge, this is a new attempt in the Italian school system. The experiment is performed on a large scale program (not through a pilot study) and meets standard requirements for the identification of rigorous evidence in the field of education.

In this paper we present the effects estimated at the end of the first year of the experiment. Students and teachers involved in the experiment are being observed for two additional years and new longitudinal data will be available subsequently.

The paper is organized as follows. In section 2 we briefly describe relevant literature; in section 3 the Italian school context and the M@tabel program; in section 4 we illustrate the design of the experiment and the data collected; in section 5 we discuss the effects of the M@tabel program on student achievement and attitudes and, in section 6, on teacher practice.

2. Features of effective professional development programs

Teacher training initiatives vary widely and there is little empirical evidence on factors affecting both teaching practice and student achievement (Guskey, 2003; Fraser, Kennedy, 2003; Fraser, Kennedy, 2003).

In recent years, a random assignment evaluation was carried out in schools to test the effectiveness of chess extra-curricular classes on mathematics achievement; a natural experiment has been used to show that the presence of an external examiner in standardized school tests reduces the proportion of correct answers in monitored classes (Bertoni, Brunello, & Rocco, 2012). Other experiments have been run with university students (De Paola & Scoppa, 2011; De Paola, Nisticò, & Scoppa, 2012).

OECD (2005) distinguishes among many types of teacher training: pre-service teacher training (typically held in universities or in qualification courses for future teachers), induction programs (to accompany new teachers during their first year(s) of teaching) and professional development programs, designed for teachers which have already been hired. Several observers (Desimone, Porter, Garet, Yoon, & Birman, 2002; Schleicher, 2011), as well as international organizations (i.e. OECD, 1998, 2005; Eurydice, 2003), stress the importance of on-going PD aimed at teaching skills, updates on contents, curricular changes and other educational reforms. The latter point is crucial since teachers play a double role in educational reforms, being both subjects and objects of change (Villegas-Reimers, 2003) and ultimately being a key element deciding the success of educational reforms and innovations (OECD, 2005).

Previous research on in-service teacher training stresses the importance of some specific features: extended duration, content-focus and peer collaboration. With US cross-section data, Garet, Porter, Desimone, Birman, and Yoon (2001) identified focus on content knowledge, opportunities for active learning and coherence with other learning activities as key features for successful training programs. These elements are confirmed by a further analyses based on longitudinal data (Desimone, Porter, Garet, Yoon, & Birman, 2002). Other non experimental evidence suggests that student performance is likely to improve when PD is focused on academic content, based on teachers’ collective participation and administered through long-term activities rather than one-day generic workshops (e.g., Kennedy, 1998; Ingvarson, Meiers, & Beavis, 2005; Timperley, Wilson, Barrar, & Fung, 2007). These studies identify good practices on the basis of large or small-scale teacher surveys. Cohen and Hill (2000), for example, observed the math performance of students whom teachers participated in initiatives specifically targeted to improving the math curriculum in California: their students scored higher on a test of math concepts imparted by the new curriculum. The results of these studies, although consistent in showing a positive effect of PD, are seriously
challenged by self selection issues: even with longitudinal data it is not possible to fully manage endogeneity problems. To state it in other terms, it’s hard to answer the following question: are the trained teachers performing better because of the training or because the better/most motivated ones self-selected into it?

More recently, some experimental evidence on PD effectiveness has been gathered – mostly in the US – showing results less consistent than previous observational research. In their review, Yoon, Duncan, Lee, Scarloss, and Shapley (2007) highlight a generally positive effect of PD on student achievement, although not always statistically significant (given the small samples used in many studies). Such results are confirmed by further experiments showing positive effects of PD especially in reading and comprehension skills (Vaughn et al., 2011; Sailors & Price, 2010; Kim et al., 2011), even if some programs have not proven to be effective, at least in the short run (Gersten, Dimino, Jayanthi, Kim, & Santoro, 2010; Garet et al., 2008).

In general, studies have focused more on programs focused on language and science than on math. For example, as attention has been rising on English Language Learner (ELL) in the American school system, so has the interest on programs addressed to train teachers facing such challenge. Both non-effective programs (Bos, Sanchez, Tseng, Rayyes, Ortiz, & Sinicrope, 2012; Arens et al., 2012) and effective programs have been detected (Matsumura, Garnier, Correnti, Junker, & Bickel, 2010; Greenleaf et al., 2011). Longitudinal studies showed also that positive effects obtained after one year of implementation (Kim et al., 2011) can last also for the subsequent year (Olson et al., 2012), suggesting a potential long term effect of such programs.

In the field of sciences, several PD programs have been evaluated and resulted, on average, in positive effects (Heller, 2012; Heller, Daehler, Wong, Shinohara, & Miratrix, 2012). Moreover science PD interventions seem to have significant cross-disciplinary impacts on students reading comprehension (Greenleaf et al., 2011), as well as in math skills (Lee,
Maerten-Rivera, Penfield, LeRoy, & Secada, 2008). Probably this result depends on the connections existing between part of the science curriculum and other disciplines. Recent experimental evidence on math seems more mixed, suggesting that only some programs might have relevant impacts on student performance. Looking at middle school teachers, Santagata, Kersting, Givvin, and Stigler (2011) found that a series of video-based modules designed for US teachers produced a positive effect. Another large scale program focusing on teachers’ knowledge of rational number topics was found ineffective after two years of implementation (Garet et al., 2010; Garet et al., 2011). Other two recent experiments, one held in high schools (Cavalluzzo, Lowther, Mokher, & Fan, 2011) and one in primary schools (Randel et al., 2011), have also reported a lack of effects.

The core outcome of these evaluations is student learning, since it is the ultimate goal for which these programs are designed. However, also effects on teachers seem just as crucial, because the transfer of competences to students must pass by teachers’ change in instructional practice (self-reported or observed) and/or their knowledge (measured through surveys or tests). Indeed, in most studies PD programs effective on students show effects also on teachers (Heller et al., 2012; Sailors and Price, 2010; Matsumura et al., 2010; Greenleaf et al., 2011). Nonetheless, there are cases where an observed improvement on teacher practices/knowledge was not accompanied by a significant increase of student performance (Garet et al., 2008; Gersten, Dimino, Jayanthi, Kim, & Santoro, 2010).

Best Evidence Encyclopedia researchers draw up systematic reviews about PD effectiveness across different subjects, trying to identify the key elements making a PD program effective (Slavin & Lake, 2008; Slavin, Cheung, Groff, & Lake, 2008; Slavin, Lake, & Groff, 2009; Slavin, Lake, Chambers, Cheung, & Davis, 2009; Slavin, Lake, Cheung, & Davis, 2009). These works conclude that the most effective classroom interventions are those designed to change teacher daily practices. Other types of interventions, such as computer-assisted instructions and curricular or textbooks modifications seem less effective. However, which
features make PD effective is still quite debated and the existing evidence is not sufficient to clearly identify them (Romano, 2012).

Even when designed according to the criteria suggested in the literature, PD programs do not lead \textit{per se} to effects on student achievement or on teacher instructional practices. More needs to be explored in order to understand how to shape teacher training to be effective and how to involve teachers in attending the programs. Little is known about whether a program has a similar impact when delivered across a range of typical settings or by multiple trainers/tutors. Finally, most studies concern the US providing little indication on specific context-based features that might condition the European (and Italian) school systems.

3. The M@t.abel program

Promoting effective professional development programs for teachers is challenging all around the world. In Italy this effort faces additional challenges, including having to address the oldest lower secondary teachers in the world\textsuperscript{7}, low wage differentiation not linked to teacher performance (OECD, 2007) and a recruitment system which does not require specific training in teaching. Moreover, in the Italian educational system, there is no differentiation in teaching career pathways and no formal or informal teacher assessment: the Italian teachers declare more frequently than in other countries that they miss feedback about their job (OECD, 2009). Most math teachers did not graduate in math or physics and should be probably considered as out-of field teachers\textsuperscript{8}.

\textsuperscript{7} In our sample, the average age is 52.

\textsuperscript{8} In our sample, 83\% graduated in other disciplines (64\% in biology).
Although in-service training is formally indicated as a professional duty, schools have little resources to actually carry out such programs. Incentives to attend in-service training opportunities are few at school and individual level, since there is no link to any form of career advancement or salary increases (Eurydice, 2003). Training supply is fragmented and delivered mainly through one-day-seminars based on frontal lectures (Moscati, 2000).

The M@tabel program evaluated in this study presents some novelty. It aims at increasing lower and upper secondary school math achievement, providing teachers with alternative solutions and methods for presenting traditional contents. The main idea is that students, rather than learning abstract formulas and ideas, should be engaged in solving real life problems through mathematical tools and concepts. The program is addressed to tenured math teachers in grades 6-8 (middle school) and 9-10 (first two years in high school). It is based on formal and on-line tutoring and it lasts a full school year. There is a repository of teaching materials facing different curriculum math concepts by adopting a problem-solving perspective. Teachers are required to use at least four of these teaching materials (precisely one per major math content area) in their classrooms and to report on the experience to their tutor and peers through a structured diary. Moreover, the program encourages a virtual community of teachers to exchange views through on-line forums and discussion groups, also from home.

Schools and single teachers within schools enroll on a voluntary basis to the program. While registering, they also indicate their preferred location, among the ones available in their areas and delivering the formal training sessions of the program. The course takes place through selected schools (called “presidii”) with proper facilities for tutors and teachers meetings. Although a substantial part of the training is done at distance, through an e-learning platform,

\[\text{9}\] The high mobility rate of non-tenured teachers generates uncertainty and hinders the development of professional communities - which is part of what the training program tries to do.

\[\text{10}\] The M@tabel platform and tools are made available only to the teachers following the training program.
M@t.abel actually starts off if at least 12 teachers sign up for the same location to attend the 7 formal lectures, for a total of 26 hours of training.

Considering the existing evidence, M@t.abel seems a promising training program, as it hold features considered of success according to international literature. It works both on teacher practice and on math content, offering specific teaching materials and promoting group work in the classroom. It lasts the whole school year, and, finally, it involves teachers and their continuous interaction through an online virtual community.

4. Experiment design, implementation and data

To evaluate the effects of M@t.abel we designed a large scale randomized controlled trial, involving 175 schools, 666 teachers and roughly 11,000 students in four Southern regions. It was designed as a three-year experiment starting in 2009/10 and addressed only to lower secondary school teachers (grades 6-8). A large amount of primary and secondary data was collected both on teachers and their students.

Our main target measure for student performance is the students’ INVALSI math competence score; however, to investigate more thoroughly the process underlying the program’s

---

11 The experiment involved about 2.2 percent of the overall 6-8 grade student population in the four regions under study.

12 The four regions are: Sicily, Apulia, Calabria and Campania. These regions cover about 85% of the Southern lower secondary student population.

13 While tests for the 6th grade are those already adopted by INVALSI for the yearly national assessment of student achievement carried out every year in May, additional items were developed for 7th and 8th graders and then linked, using Rasch analysis, to measure all students on a common scale. See Appendix for further information on the 7th and 8th grade tests.
impact we also seek effects on (self-reported) students’ attitudes and teachers’ beliefs and classroom practice\textsuperscript{14}. 

Enrolment to M@t.abel is basically voluntary, so we have to control for self selection at school and individual levels. Moreover, an additional selection process is due to the program’s extended duration: after the enrolment, teachers could drop out or not implement all the scheduled activities. We kept these considerations in mind while designing the experiment and we will discuss their implications further, while dealing with compliance issues.

**Randomization and validity**

The identification strategy is based on a typical treatment-control comparison between students clustered by classes (and therefore by their teachers) and schools\textsuperscript{15}. Given the importance of peer collaboration in the M@t.abel approach, only schools showing at least two enrolled teachers were considered for this experiment. We ended up with 175 schools involved in the randomized control trial. Schools were randomly assigned to two groups: the enrolled teachers belonging to the first group of schools received the specialized training in year 2009/10 (treatment group), those belonging to the second were delayed admission for one year (control group), thus admitted to the program in year 2010/11. We stratified the 175 schools according to geographical criteria (namely by province, isolating the city of Naples and Palermo as specific strata) and by peer participation at the school level (schools with less than 5 teachers enrolled and schools with 5 or more). We obtained 31 non-null sample layers,

\textsuperscript{14} Teachers went through extensive computer-assisted telephone interviews (CATI) during the first and second school-years of the M@t.abel program (January/February 2010 and December/January 2011). The first CATI was conducted after the beginning of the PD and cannot be considered as a genuine pre-intervention observation; we use the information coming from the second CATI, developing a post treatment comparison on teachers.

\textsuperscript{15} Given the high teacher mobility across schools, a high dropout rate from the three-year observation should be expected, but we rely on a large sample to avoid this drawback.
then 55 schools were randomly assigned to the control group, proportionally to the
distribution of the schools in each sample layer. The remaining 120 schools were assigned to
the treatment group\textsuperscript{16}. From the 175 schools, we obtained a sample of 666 teachers: 409 from
the 120 schools assigned to treatment group were invited to attend the program immediately,
172 from 55 schools were considered as control group (and their teachers were invited to
attend the program during the following school year) and 85 dropped out before the
beginning of the school year but after the randomization\textsuperscript{17}. We finally ended up with 174
schools (120 assigned to treatment and 54 controls) because in one control school all teachers
enrolled dropped out.

Only one class, randomly chosen by the research team among the many in which each
teacher works, was assigned to every teacher for the evaluation purposes\textsuperscript{18}. This random
selection was stratified across 6\textsuperscript{th}, 7\textsuperscript{th} and 8\textsuperscript{th} grade classes (for both treatment and control
group), so that students being observed throughout this experiment are about equally
distributed across the three grades. Teachers were asked to implement the \texttt{M@table} teaching
materials in the assigned class (treatment group) or were informed about the involvement of
their class in the experiment and the fact it would be tested at the end of the school year
(control group).

Thanks to the large amount of information collected, we were able to test the equivalence
between treatment and control group across an unusually wide range of variables at school,

\textsuperscript{16} The number of schools belonging to the control group is smaller than the treatment group for two reasons: to
reduce the inconvenience of treatment delay as much as possible and to ensure enough room for the experiment
to hold even in case of drop-outs from the treatment during the school year.

\textsuperscript{17} The drop out is due to uncontrollable factors (such as teachers deciding to retire or change of school at the
beginning of the school year) and not related in any way to the experiment. The proportion of drop-out teachers
was roughly the same in the treatment and control groups and their classes were not observed.

\textsuperscript{18} The \texttt{M@table} teaching materials are engaging for teachers not familiar with the approach. It was not realistic
to expect that teachers would properly implement \texttt{M@table} in all their classes.
teacher and student levels (about fifty\textsuperscript{19}). The internal validity of the experiment is verified. Indeed, controlling for the randomization variables, we found only few and very small (although statistically significant) differences between treatment and control groups\textsuperscript{20}. However, the extremely wide set of variables used in the comparison should be acknowledged when looking at this result. In any case, we run several models estimating the program effects by adding control variables to take into account even these small differences and results are robust to this sensitivity analyses.

Looking at the external validity of the experiment, we detected that figures of our sample (based on self selected schools in four Southern regions) generally compare to those of the population of other schools, teachers and students in the whole of the eight regions of Southern Italy\textsuperscript{21}, but not to the rest of the country. In any case, considering the self selection

\textsuperscript{19}To provide some examples, we checked the equivalence on schools using data on student population (average SES, number of immigrants, math competence in 2008-09 test for 8\textsuperscript{th} graders), school characteristics (number of students, average class size, size of the town, geographical variables). While checking the equivalence on schools and teachers, we adopted a very cautious definition of statistical significance (p value<.10).

For the equivalence of teachers we relied on a survey administered at the beginning of the experiment, with information on personal characteristics (age, gender), as well as work and academic history (i.e.: field of study, previous training, year of work experience, tenure). For what concern students, we exploited the questionnaires to obtain background information (mainly: gender, age, migration status, parental involvement in homework, parental education). While checking the equivalence on students, we adopted standard definition of statistical significance (p value<.10).

\textsuperscript{20}Namely, at school level: the presence of primary and secondary schools in the same institute; at class/teacher level: the presence of an external observer during the national math assessment; at student level: treated and control slightly differed on help at home with homework, the rate of parents holding only primary education and age of the 6\textsuperscript{th} graders. The differences, it should be noted, were never big in size, even when statistically significant.

\textsuperscript{21}The only relevant difference is that the observed schools are located more frequently in urban areas and less in the mountains. Only Southern Italy is considered in our external validity tests; the divide with the rest of the country is striking and our results could be not applied to the other areas of Italy.
of schools and teachers during the enrolment, our experiment does not provide external validity, despite its large sample-size.

**Non compliance**

The estimated effects could be diluted by the fact that some teachers assigned to treatment did not actually complete the M@t@bel program or did not participate for the length of the whole school year. Complier teachers are those who fulfilled the treatment protocol as follows: *a)* to receive the final certificate given by the tutor, proving their full attendance; *b)* to experiment at least 4 teaching units in their observed class; *c)* these teaching units must come from 4 different math content areas. The first two criteria were adopted in order to ensure a sufficient intensity of the treatment; given the availability of teaching material, we chose to add also the latter criterion in order to guarantee homogeneity of the treatment.

Only 39% of the teachers were compliant to the treatment protocol (Table 1). Non compliance is characterized by different behaviors: 34% quit at the very beginning of the program; 4% did not get the end-of-training certificate (typically because they experimented less than four teaching units); 10% did not implement the program properly (they did not use teaching units from different math content areas or not in the assigned class); 13% had intended to participate but the training course actually never started in their area. The rate of compliance is similarly distributed across the three school grades, ranging from 40.7 in 6th grade, to 44.0 in 7th grade and 39.6 in 8th grade. We did not observe crossover among the control group of schools and teachers.

**TABLE 1 HERE**

---

22 Data collected directly from schools allowed us to verify that there were no close substitutes to M@t@bel available to those in control group. Schools all benefited from a wide range of other of activities or projects involving math (such as after-school homework assistance, math software, etc.) but there was no significant difference between the two groups.
Features of non compliance have been explored using multivariate binary logistic models, pointing out to self-selection issues in the subsample of teachers fulfilling the treatment protocol. Compliance is associated to younger teachers, their ICT familiarity, participation to previous in-service training opportunities and personal motivation to enroll to the program (instead of having the school principal register them for M@t.abel).

Probability to complete M@t.abel is greater when teachers were located in more urbanized and less mountainous areas. Given that in several cases teachers were not assigned to their preferred location for treatment, these factors could have counted on their willingness to follow the formal training sessions or on how easy it was for them to actually get there\textsuperscript{23}.

Indeed, the main reason reported by non-compliers to justify their dropping out is the distance from the course location, followed by time-constraints. The program requires time to reach and attend formal lectures; time in the classroom to use the materials with students; time to report about those experiences; digital skills to download materials and exchange comments with colleagues. This should be kept in mind while recalling that that the large majority of teachers in the Italian lower secondary schools are middle-age females\textsuperscript{24} (OECD, 2012) and they have to face work-family reconciliation difficulties.

5. Short term effects on student math performance and attitudes

The effects of one-year of the M@t.abel program on student and teachers are estimated in terms of intention-to-treat (ITT) by OLS models. Considering the high non compliance rate, we also estimated the average-treatment-effect-on-the-treated (ATT), instrumenting

\textsuperscript{23} We also asked teachers how much time it took them to get to the formal training location. We did not use this information, because there were too many missing answers to this question. Teachers enrolled but not attending the first PD lecture were not able to answer.

\textsuperscript{24} 84\% of cases in our sample are female.
compliance with the assignment to the treatment. The ATT estimates are displayed despite the self-selection among compliers. The relevance of observed characteristics associated to compliance suggests that it might not be reasonable to expect that effects estimated on compliers can be extended to non compliers.

In the following sections, we present the base models on students and teachers, where we control only for stratification variables considered in the randomization process and for the presence of an external observer during the math assessment. We run several models to check for robustness, using different sets of control variables and the results of our experiment do not change.

Looking at student math performance, we consider three target variables:

- the overall math score measured by the INVALSI national assessment and estimated using a Rasch model (scaled to an average of 500 and standard deviation of 100 for the 7th grade scores);

- the frequency of skipping/double-marking at least one item, proving a strong hesitance in responding (with the exception of those skipped because the student did not reach the end of the assessment);

---

25 As robustness check, we considered also ATT models instrumenting partial compliance (teachers who attended at least frontal lectures). Results hold even with this specification. Both in OLS and IV regression models we take into account the data clustering.

26 In order to avoid cheating or manipulation, for the sake of this experiment, external observers were hired to guarantee the regularity of the national math assessment administration (covering about 83% of the observed classes). Considering the relevance of this contextual factor (Bertoni, Brunello, & Rocco, 2012) and the not perfect equivalence between treated and control group on this feature, we always use a dummy variable in the model. We could not rely on data statistically corrected by INVALSI for cheating, because this procedure is applied only to the 6th grade scores.

27 See note 21.
- the frequency of not completing the assessment (i.e., not reaching one or more of the last items in order of presentation).

TABLE 2 HERE

The results reported in Table 2 show that the treatment has no significant impact on the main outcome, the average math competence score\(^{28}\). Students in the classes where teachers participated to the M@t.abel program actually present, \textit{ceteris paribus}, an average slight advantage in the performance, although neither statistically significant nor substantially relevant. However, the program seems to have an undesired impact on increasing the propensity of the treated students to skip at least one item during the assessment. However, this behavior did, not affect the overall number of items skipped or double-marked, nor the propensity of not completing the assessment\(^{29}\). It could be due to the treated students’ concern about doing well in the test and, therefore, avoiding to just guess the answer.

Indeed, several hints from looking at the effects of M@t.abel on students’ attitudes, reinforce this interpretation (Table 3). Students of treated teachers show stronger self confidence in math, report less frequently that the causes of academic failure are due to bad luck, but they also feel a higher level of anxiety while taking the test. By combining these elements, we draw a picture of greater responsibility perceived by students in determining their achievement outcomes (Zeidner, 1998), revealing that no answer could be preferred to guessing an answer (Hagtvet & Benson, 1997). Hence, the treatment could have produced a more “perfectionist” attitude (Mills & Blankstein, 2000), characterized by conscientious and a goal-oriented behavior (Tsui & Mazzocco, 2007).

\(^{28}\) The result does not change considering the median score through a quantile regression model.

\(^{29}\) The ITT is marginally significant (p value: 0.132).
Another unexpected effect lies in students’ perception of time constraints in learning: treated students report more frequently than controls that they feel they have not had enough time on a given subject. This result matches the treated teachers complaints about the little time available to actually implement the new approach in the classroom.

Although negative in the short-term, these effects on students’ attitudes could be promising for future improvements: feeling responsible for their achievement, more interested in math, confident about their skills and being perfectionist in solving problems could become an advantage in building math competences, especially as the teachers become more familiar with the M@t.abel approach.

TABLE 3 HERE

Given that specific sub-groups could have benefited more from the program than others and thanks to the large sample-size, we explored heterogeneity of the effects among different groups of schools, teachers and students. Hints for differential program effectiveness appear only as concerns teachers’ age, as shown in Table 4. Students of middle aged teachers (in our sample 50 to 55 years old) show a significant positive effect of M@t.abel on their average math score (ITT: 14.8; ATT: 41.6), while the effect is absent (if not even negative) for other age groups. Looking at attitudes, treated students of middle aged teachers show some additional difference respect to their peers: they are less likely to consider math more difficult for them than for their peers, they report learning difficulties due to the curriculum’s fast pace less frequently, they attribute failures to bad luck or chance less frequently and they feel less nervous while taking the test. These clues seem to suggest that middle aged teachers managed the new teaching approach better than the others, providing all students with the opportunity to be fully involved and actively learn. This could also have favored a higher cooperative environment which improved students’ perception of ability in comparison with their peers,
rather than self-confidence or interest in math. In this perspective, a possible explanation for the lack of the program effectiveness among the other teachers’ groups could be linked to time constraints in implementing the activities. This pressure could have made it difficult to adapt the intervention to both good and weak students’ needs, by stimulating higher competition and performance anxiety in the classroom. However, the small number of teachers for each age subgroup and their distribution in a not random subsample of schools suggest cautiousness in deriving strong conclusions from this result.

6. Short term effects on teacher attitudes and classroom practice

Even before affecting students, M@t.abel should have triggered changes in teachers’ knowledge, practices and attitudes. Thus, it is relevant to examine whether there are changes in the way teachers are teaching.

Relevant information was collected through a CATI interview based on an ad hoc questionnaire, including items on attitudes (towards math teaching, self-efficacy perception, job satisfaction) and instructional practices (classroom activities, evaluation of students, materials and instruments used to teach, interaction with colleagues). Post-treatment interviews were administered in December 2010/January 2011, once the professional development for teachers in the treated group was finished and a new school year had already started.

30 Unluckily, it was not possible to assess the improvement in teacher content knowledge, although M@t.abel is a highly content oriented program.

31 The treatment group had concluded the professional development program and started a new school year; instead, (part of the) the control group was just starting to attend the program. 57% of the control group actually...
We estimated the effects on teachers’ instructional practices and attitudes using OLS regression models on pseudo-continuous variables\textsuperscript{32}, and linear probability models on dichotomous variables\textsuperscript{33}.

Table 5 shows the questionnaire’s items where we found effects robust to different model specifications (controlling also for non-equivalence covariates) for both ITT and ATT estimates\textsuperscript{34}.

\textbf{TABLE 5 HERE}

The observed effects regard dimensions relevant in everyday teaching and strictly linked to the program: collaboration with peers and attitudes towards teaching (namely, the way they think about teaching math and the way they can do it)\textsuperscript{35}. In the school year following the treatment, treated teachers became significantly more eager to collaborate with peers at work, both in the preparation of didactic materials and in discussing the better way to present a concept to the students.

The size of these effects is relevant. Since collaboration among teachers can be considered one of the effectiveness-enhancing factors (Scheerens, 2000) and since M@tabel explicitly enrolled in M@tabel during the second observed school year and at the moment of the interview, most of them (about 80%) only attended the first introductory lecture and can be considered as not treated at all.

\textsuperscript{32} We standardized the values of the variables, originally on a 1-10 points scale.

\textsuperscript{33} Variables expressed on an ordered, non-continuous scale were dichotomized in order to estimate a linear probability model. We performed this change because routines in statistical packages do not allow for instrumental variables models with ordered outcomes.

\textsuperscript{34} As concerns categorical variables, we considered robust only the effects present both on the average (after rescaling the values of each variable) and on the dummy variables derived dichotomizing original variables.

\textsuperscript{35} A large number of other effects slightly above the threshold of significance reinforce this interpretation, showing that teachers do act in a more innovative and collaborative way and, at the same time, the treatment gave them more resources enabling them to re-think their role and their way to teach.
promoted it, this result seems particularly promising. The program seems to have had an impact also on the way teachers feel about teaching math: those in the treatment group agree less with the statement that math ability is a sort of “fate” depending on students’ own endowment (which cannot be substantially affected by teacher effort). Finally, the treatment seems to have led also teachers to perceive their limits, making them feel less confident about their effectiveness in promoting collaboration among students. This could be the consequence of more frequent group-work practiced in the classroom while using the M@tabel teaching units and understanding more profoundly the challenges of a more interactive teaching style. In brief, the way teachers conceive teaching and the way to reach their goal has been influenced by some features of the M@tabel program. These results seem particularly promising for its future effects on student achievement.

6. Conclusions and lessons learned
This paper studies the effects of the M@tabel teacher training program on student math performance and attitudes, and on teacher behavior and instructional practices. We run the one of the first ever large scaled randomized experiments in the Italian school system, based on a sample of schools in four Southern regions. We learned some lessons about the evaluated PD program, but also about the implementation of rigorous policy evaluation in the Italian educational system.

Although preliminary findings show no significant impact of the M@tabel program on student math performance, some effects on student attitudes and teacher practices and beliefs do appear. Students of treated teachers show a more positive attitude towards math and there are signals of a greater student responsibility in learning and facing the national math assessment. Their higher attention versus math performance could be an advantage in building math competences. Looking at the effects on teachers, we observe that there has been a change towards a more innovative way of leading the classroom, especially promoting
more frequent exchanges with colleagues. At the same time, we detect that complier teachers suffered harder time constraints in combining the ordinary school lessons with the implementation of M@t.abel and the training attendance.

Overall, the program has not been effective in the short run, but the occurred changes in teachers and students attitudes could activate mechanisms leading to future positive effects on math achievement. We also detected signs of heterogeneity of the treatment’s effectiveness: it is plausible that middle aged teachers managed time for the implementation of the new teaching approach better, providing all students with the opportunity to be fully involved and actively learn. The consequence is that M@t.abel significantly increased their students’ math performance and their perception of ability in comparison with their peers.

These results suggest that the lack of overall effectiveness could be due to the additional effort required to teachers during the first year of the policy. The study is continuing and building up a longitudinal sample of the same students for another two years. Thanks to this approach we will be able to detect possible effects at a distance, or as teachers gain familiarity with the M@t.abel program.

The evaluation also suggests how to modify the program, to better retain in the future those who enroll. For example, given that age is a relevant factor for full compliance, addressing the program only to tenured – and usually older – teachers could be questioned. Similarly, one might find it useful to ensure that participants have sufficient ICT skills before enrolling for this kind of intervention.

Although a considerable amount of work remains to be done, by developing ex ante the evaluation design in cooperation with the Ministry of Education and the M@t.abel Scientific Committee the program schedule was affected. Once a randomized experiment was agreed upon, the institutional actors leading the program were driven to streamline the activities and actually reinforce their key features in terms of duration, content-focus and peer collaboration. In practice this has meant, insuring that the program actually started in early
autumn so as to allow its completion by the time theINVALSI standard math assessment
would be held in May. It also implied fixing more precise and homogeneous requirements on
the training (protocol), such as requiring four teaching units to be used in the classroom and
encouraging school-level rather than teacher-level participation. The evaluation produced
deeper collaboration among the different institutions, because they had to collect and
exchange data about the intervention and its features. This process was successful, leading
the research group to obtain an unusually rich set of data in the Italian context. More could be
done in the future: indeed, intervention platforms are still not designed to easily provide
information about the implementation of the program, such as time spent on line by teachers,
number of units downloaded by teachers, etc.

Before randomizing, there was substantial fear among institutions and researchers that
teachers would complain for being excluded from the program, disobey the recommendation
on which class was to practice M@tabel and/or be unwilling to collaborate in the data
collection. The actual picture was very different. Many teachers asked for explanations on the
logistics of the evaluation through the web-site contact e-mail, but almost all participated to
pre-post treatment interviews and allowed for their students to be assessed. Within the
treatment group only a minor share of the teachers decided to practice M@tabel in a class
other than the randomly assigned one. This evaluation has proven that randomized
experiments can be conducted in Italy and be useful in a context where there is a lack of
culture and some resistance to such an approach (see also CITATION DELETED FOR
BLIND REVIEW). We see this experience as an encouraging note for future randomized
trials and evidence based policies in the Italian education system.

36 Less than 4% refused to answer the pre-treatment interview and less than 8% the post-treatment interview
Denials were not necessarily related to “disappointed” control teachers.
References


De Paola, M., Nisticò, R., & Scoppa, V. (2012). Monetary Incentives and Student Achievement in a Depressed Labour Market: Results from a Randomized Experiment. *Journal of Human Capital, 6*(1), 56-85. doi:10.1086/664795


26


Slavin, R. E., Lake, C., Cheung, A., & Davis, S. (2009). *Beyond the basics: Effective reading programs for the upper elementary grades*. Baltimore, MD: Johns Hopkins University, Center for Data-Driven Reform in Education.


Table 1

*Compliance in terms of teachers and students*

<table>
<thead>
<tr>
<th></th>
<th>Teachers</th>
<th>Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intention to treat</td>
<td>409</td>
<td>7,692</td>
</tr>
<tr>
<td>Treated (compliers)</td>
<td>160</td>
<td>3,053</td>
</tr>
<tr>
<td>Control</td>
<td>172</td>
<td>3,372</td>
</tr>
<tr>
<td>Total</td>
<td>581</td>
<td>11,064</td>
</tr>
</tbody>
</table>
### Table 2
*Average impact on student math performance*

<table>
<thead>
<tr>
<th>Descriptive Statistics</th>
<th>Effect estimates and standard error (OLS and IV regression)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Treatment</td>
</tr>
<tr>
<td>Math score (mean)</td>
<td>493</td>
</tr>
<tr>
<td>Skipping items (%)</td>
<td>65.8</td>
</tr>
<tr>
<td>Not completing the test (%)</td>
<td>6.8</td>
</tr>
</tbody>
</table>

Note. The symbols ***, **, * indicate that coefficients are statistically significant at the 1, 5, and 10 percent level.
Table 3

*Average impact on student attitudes towards math and school*

<table>
<thead>
<tr>
<th>Dimensions</th>
<th>Value controls</th>
<th>ITT</th>
<th>ATT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Attitudes towards math</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 items factor (std score)</td>
<td>-0.05</td>
<td>+0.06**</td>
<td>+0.16**</td>
</tr>
<tr>
<td>4 items factor (std score)</td>
<td>-0.05</td>
<td>+0.06**</td>
<td>+0.16**</td>
</tr>
<tr>
<td>Item “In math I’m good” (1-4 points scale)</td>
<td>+2.78</td>
<td>+0.05**</td>
<td>+0.14**</td>
</tr>
<tr>
<td>Item “Math is more difficult for me than the majority of my classmates” (1-4 points scale)</td>
<td>+2.92</td>
<td>-0.04*</td>
<td>-0.10*</td>
</tr>
<tr>
<td>Item “I enjoy doing Math” (1-4 points scale)</td>
<td>+2.57</td>
<td>0.06*</td>
<td>0.16*</td>
</tr>
<tr>
<td>Item “I would like to take more Math in school” (1-4 points scale)</td>
<td>+2.22</td>
<td>0.06*</td>
<td>0.16*</td>
</tr>
<tr>
<td><strong>Curriculum pace</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>We proceeded even if some classmates did not understand the topic (1-4 points scale)</td>
<td>+1.55</td>
<td>+0.07***</td>
<td>+0.18***</td>
</tr>
<tr>
<td><strong>Failure attributions</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attribution of failures to bad luck (0-6 points scale factor)</td>
<td>+0.19</td>
<td>-0.03**</td>
<td>-0.07**</td>
</tr>
<tr>
<td><strong>Test Anxiety</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 items factor (std score)</td>
<td>-0.04</td>
<td>+0.06*</td>
<td>+0.15*</td>
</tr>
<tr>
<td>Item “I was so nervous that I couldn’t find the right answers” (1-4 points scale)</td>
<td>+1.89</td>
<td>+0.07***</td>
<td>+0.18***</td>
</tr>
</tbody>
</table>

Note. The symbols ***, **, * indicate that coefficients are statistically significant at the 1, 5, and 10 percent level.
Table 4

Average impact on student performance and attitudes towards math and school by teachers’ age group

<table>
<thead>
<tr>
<th></th>
<th>ITT</th>
<th>ATT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&lt;50 aged</td>
<td>50-55 aged</td>
</tr>
<tr>
<td></td>
<td>&lt;50 aged</td>
<td>50-55 aged</td>
</tr>
<tr>
<td>Math score</td>
<td>-6.0</td>
<td>+14.8**</td>
</tr>
<tr>
<td></td>
<td>-10.3</td>
<td>41.6**</td>
</tr>
</tbody>
</table>

Attitudes towards math

General attitude

<table>
<thead>
<tr>
<th></th>
<th>ITT</th>
<th>ATT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&lt;50 aged</td>
<td>50-55 aged</td>
</tr>
<tr>
<td></td>
<td>&lt;50 aged</td>
<td>50-55 aged</td>
</tr>
<tr>
<td>Item “In math I’m good” (at least “I quite agree”)</td>
<td>0.00</td>
<td>+0.03</td>
</tr>
<tr>
<td>Item “Math is more difficult for me than the majority of my classmates” (at least ”I quite agree”)</td>
<td>0.01</td>
<td>-0.04***</td>
</tr>
<tr>
<td>Item “I enjoy doing Math” (at least ”I quite agree”)</td>
<td>0.00</td>
<td>+0.02</td>
</tr>
<tr>
<td>Item “I would like to”</td>
<td>+0.05*</td>
<td>+0.02</td>
</tr>
</tbody>
</table>
take more Math in school”
(at least "I quite agree")

**Curriculum pace**

We proceeded even if some classmates did not understand the topic (at least "Seldom")

<table>
<thead>
<tr>
<th></th>
<th>+0.05*</th>
<th>+0.02</th>
<th>+0.08***</th>
<th>+0.08*</th>
<th>+0.06</th>
<th>+0.31***</th>
</tr>
</thead>
</table>

**Failure attributions**

Attribution of failures to bad luck (standardized score)

<table>
<thead>
<tr>
<th></th>
<th>-0.04</th>
<th>-0.07*</th>
<th>-0.04</th>
<th>-0.07</th>
<th>-0.20*</th>
<th>-0.15</th>
</tr>
</thead>
</table>

**Test Anxiety**

General anxiety index (4 items factor analyses standardized score)

<table>
<thead>
<tr>
<th></th>
<th>+0.07</th>
<th>+0.02</th>
<th>+0.08</th>
<th>+0.12</th>
<th>+0.07</th>
<th>+0.28</th>
</tr>
</thead>
</table>

Item “I was so nervous that I couldn’t find the right answers” (at least "I quite agree")

<table>
<thead>
<tr>
<th></th>
<th>+0.03*</th>
<th>+0.02</th>
<th>+0.06***</th>
<th>+0.05*</th>
<th>+0.07</th>
<th>+0.22***</th>
</tr>
</thead>
</table>
Note: The symbols ***, **, * indicate that coefficients are statistically significant at the 1, 5, and 10 percent level.

<table>
<thead>
<tr>
<th>Item</th>
<th>Value controls</th>
<th>ITT</th>
<th>ATT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>β</td>
<td>SE</td>
<td>β</td>
</tr>
<tr>
<td>Interactions with colleagues</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Preparation of teaching materials jointly with colleagues. 1-3 times a week (Ref: less frequently)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>47.2%</td>
<td>+12.3***</td>
<td>4.5</td>
</tr>
<tr>
<td>Exchange of point of views on how to teach a particular topic. 1-3 times a week (Ref: less frequently)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td>+7.2*</td>
<td>3.8</td>
</tr>
</tbody>
</table>

Table 5

Differences across treatment and control groups on teachers’ attitudes and practices (ITT and ATT)
Many students show difficulties while doing logical abstract reasoning

<table>
<thead>
<tr>
<th>Variable</th>
<th>Value</th>
<th>p-value</th>
<th>Coefficient</th>
<th>p-value</th>
<th>Coefficient</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-efficacy perception (standardized score)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Make the students work together</td>
<td>7.45</td>
<td>0.17</td>
<td>0.09</td>
<td>0.43</td>
<td>0.22</td>
<td></td>
</tr>
</tbody>
</table>

Note: The symbols ***, **, * indicate that coefficients are statistically significant at the 1, 5, and 10 percent level; "Coefficients of the battery "Interaction with colleagues" are expressed in percentage points. The value of controls for the attitudes towards math teaching and self-efficacy perception refer to the original scaling of the variable (ranging from 1 to 10).