Abstract Title Page
Not included in page count.

Title: A quest for increasing student math achievement and promoting rigorous evaluation in Italy: evidence from the M@t.abel Teacher Professional Development Program

Authors and Affiliations:

Argentin, Gianluca, Università di Milano Bicocca, Italy, gianluca.argentin@unimib.it
Pennisi, Aline, Ministry of Finance, aline.pennisi@tesoro.it
Vidoni, Daniele, INVALSI, daniele.vidoni@invalsi.it
Abbiati, Giovanni, Università degli Studi di Milano, Italy, giovanni.abbiati@unimi.it
Caputo, Andrea, INVALSI, andrea.caputo@invalsi.it

Background / Context:
Research has proven that teachers have a fundamental influence on student results (Scheerens and Bosker, 1997; Scheerens, 2000; Scheerens, 2007). Moreover, effective teacher professional development is one of the key mechanisms for improving student achievement (Darling-Hammond, Holtzman, Gatlin & Heilig, 2005; Darling-Hammond, 2000; Kennedy, 1998).
By the most recent OECD definition (TALIS survey, OECD, 2009, p. 49), "Professional development is defined as activities that develop an individual's skills, knowledge, expertise and other characteristics as a teacher". The M@t.abel program suits this definition both in terms of increasing teacher subject knowledge and in terms of providing math teachers with alternative solutions and methods for presenting usual contents. Indeed, M@t.abel is focused on teaching students to solve real life problems using mathematical tools and concepts, rather than learning abstract formulas and ideas. It is addressed to math teachers in grades 6-8 (middle school) and 9-10 (first two years in high school).
Given that the effects of any professional development program on student achievement are mediated by teacher actual practices in the classroom (Cohen & Hill, 2000; Cohen, Raudenbush & Ball, 2002), it is necessary to verify whether teachers enrolled in the M@t.abel program do follow the training, whether they apply the acquired knowledge in the classroom, and whether they change their attitudes and teaching methods. The study will address these issues by collecting information through specific questionnaires and the teacher's M@t.abel log-book.
The difficulty in measuring the effectiveness of teaching is that "teachers have powerful effects on reading and mathematics achievement, though little of the variance in teacher quality is explained by observable variables, such as education or experience" (Rivkin, Hanushek and Kain, 2005). The study rules out issues resulting from hidden heterogeneity by designing ex ante a randomized experiment, which allows us to compare the results of two statistically equivalent groups of participants in the most objective way possible.

Purpose / Objective / Research Question / Focus of Study:
The weaknesses shown by Italian students in international tests on mathematics and science (IEA-TIMSS and OECD-PISA studies), has produced a flurry of initiatives to help schools and teachers improve student achievements. Among these initiatives, teacher professional development programs supported by the Ministry of Education are playing an important role. One of these programs, called M@t.abel, is aimed at training teachers in the application of mathematics to everyday life situations. M@t.abel classroom materials have recently been expanded to cover a substantial fraction of the middle-school curriculum and the program is being strongly promoted in four regions of Sothern Italy, thanks to European Union funding.
The key evaluation question is whether exposure to M@t.abel has a discernible effect on student achievement, both in terms of the whole population and with respect to specific subgroups of students identified in terms of prior ability and individual characteristics. The study attempts to answer this question through random assignment (at school and classroom level) and involves the collection of a large quantity and variety of data on teachers, students, and schools. In addition to the students' math scores in standardized tests, we measure students' attitudes towards math. Corollary to the main question question, we also investigate whether exposure to the program leads to measurable changes in teachers' attitudes towards teaching math and in their reported teaching behaviors.
Being (to our knowledge) the first large scale random assignment evaluation attempted in the Italian school system, the study also contributes to understanding how rigorous evaluation

techniques can be carried out in Italy, what the reactions of the educational and practitioner community are to practices such as random assignment, and how policy makers can be involved.

Setting:
4 out of the 20 administrative Italian regions (Campania, Calabria, Sicilia, Puglia) receive European Union Structural Funds.
These regions are all in the South and are characterized by poor socioeconomic characteristics, relatively weak infrastructures, low student achievement, and high rates of student drop-out.

Population / Participants / Subjects:
Teacher enrolment in the M@t.abel program is on annual basis. The reference population included in the experiment consisted of 877 teachers coming from 263 schools enrolled in 2009.
The actual population for the experiment consists of 581 teachers in 174 schools identified on the basis of selection criteria privileging schools with more than one teacher enrolled and no previous exposure to M@t.abel.The immediate-treatment group comprises 120 schools, for a total of 409 classes and 7,692 students. The control-by-waiting group comprises about 50 schools, for a total of 172 classes and roughly 3,372 students. Within the immediate-treatment group, roughly 40% of teachers were "compliers," while the rest were "non compliers".
Thanks to the huge amount of information collected, we were able to test the equivalence between treatment and control group across a unusually wide range of variables at school, teacher and student levels. By large, the internal validity of the experiment is verified. Controlling for the randomization variables, we found small but statistically significant differences only on a reduced set of factors, probably due to the extremely wide set of variables used in the comparison between treatment and control group.

Intervention / Program / Practice:
M@t.abel is a teacher training program offered to Mathematics teachers of grade 6 to 10.
In terms of the current research, we focus on training provided to teachers of grade 6-8 as these grades correspond to the middle school cycle. Training is offered for one school year and is not repeated. The program administration in a blended, as it includes both meetings in person (8 along the year, 30h work) and online forums.
In the program, teachers receive training on specific didactic units that allow them to use alternative methods for teaching curricular contents. Each unit involves the use of practical examples and laboratories. Of the over 200 available didactic units, teachers must experiment in class 4 units. Individual and collective reflection of the class experiences is solicited and favored by means of online forums and the discussion with the virtual class. Moreover, as a requirement for obtaining the certification, teachers must fill out an "on board diary".

Research Design:
The paper presents the design and the results after the first year of activity of an ongoing 3-years study on the effectiveness of the M@t.abel program. The basic evaluation design sees schools and teachers within schools enrolling on a voluntary basis for the 2009/10 school year. The schools are randomized into two groups: one group receives the specialized training beginning in year 2009-10, the other group is delayed admission for one year, then released into treatment.
Randomization and the delayed treatment allow credible estimate of the effect of the first year of involvement in M@t.abel, as shown in the Figure 1.

-- Please enter Figure 1 about here --

Effects on student will be measured over three years by using the standardized tests produced by INVALSI, the Italian National Institute for Educational Evaluation. Figure 2 shows the second year effect estimation procedure.

-- Please enter Figure 2 about here –

Only the results on the first year effects have been estimated at this stage, but we are now running the second year estimates for students and collecting data from teachers at the end of the third year. At the conference, we will be able to present definitive findings for students at the end of the second year and for teachers at the end of the third year.

Statistical, Measurement, or Econometric Model:
We use the Student National Assessment data (SNV – INVALSI) to estimate effects on their math skills. Considering the high non compliance rate, we estimate the program effects both using ITT and ATT, using instrumental variable regression and correcting the standard errors for data clusterization.

Usefulness / Applicability of Method:
Being (to our knowledge) M@t.abel the first large scale random assignment evaluation attempted in the Italian school system, we spent a lot of time collecting and sharing information with all the figures involved in the program: teachers, on line tutors, contents providers, etc. Moreover we promoted a strict connection with the Ministry of Education to make results available step by step to the decision makers.
We are now able to integrate quantitative and qualitative data about the implementation process, to better explain positive and null results of the intervention. Hence, we can develop reflections about the implications of using a randomized control trial in the Italian educational system.

Data Collection and Analysis:
With the support of INVALSI, the project team is collecting a large amount of primary data on both students and their teachers. The data have been analyzed following these steps: a) using student results to estimate student ability by means of IRT modeling; b)Investigating the effect and comparing it among different grades (6, 7, 8); c) developing comparison of the effect size among students' subgroups (gender, social background, previous math's school mark); d) estimating the potential effect on students' attitude toward math after the first and second year; e) estimating the potential effect on teachers' beliefs and reported teaching behaviors after the first and third year.

Findings / Results:
In table 1, short-term effects of M@t.abel on three different measures of student performances (overall math score, frequency of skipping at least one item, frequency of not completing the assessment) are reported, both in terms of ITT and in terms of ATT.

-- Please enter Table 1. About here --

This results show that the treatment has no significant impact on the average math competence score difference. Students in the classes where teachers participated to the M@t.abel professional development program actually present, ceteris paribus, an average slight advantage in the performance, although not statistically significant. However, the program seems to have an undesired impact on increasing the propensity of the treated to skip at least one item. Interpreting this result is not straightforward as no significant effect is detected on connected variables, such as the average number of items skipped or double-marked, nor on the propensity of not completing the assessment. Further research in the next years should help clarify whether this is a short-term behavior due for example to the fact that treatment students were more concerned by doing well during the test and avoiding just guessing the answer.

-- Please enter Table 2. About here --

At the same time, M@t.abel seems to have considerably improved students' attitudes towards math, favoring in particular a deeper interest for the subject and promoting a stronger engagement along the learning process (table 2). Students of treated teachers, moreover, report less frequently the causes of academic failure to chance or to bad luck. Treated students report a higher level of anxiety while taking the test. Given that specific sub-groups could have benefited most from the program, we explored heterogeneity of the effects among different groups of schools, teachers and student. There are some hints that teachers' age group could be a factor affecting the effectiveness of the intervention. 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: 15,2). The small size of teacher and school subsamples suggests cautiousness in deriving conclusions.
This result of null effect is quite usual in the evaluation of teacher professional development initiatives; it does not mean that the program would be not effective at all. That is the reason why we are observing teachers and students also in the 2nd and 3rd year following the treatment. Preliminary results on student achievement at the end of the second year are confirming the 0 impact of the intervention. We are still working on this part of the dataset and definitive results will be available at the conference.

Conclusions:
Although findings at the end of first year and preliminary findings at the end of second year show there is no significant impact of the M@t.abel program on student math performance, some effects on students' attitudes and teachers' practice appeared. In particular, the effects on teachers' attitudes and practices suggested a change towards a more innovative way of leading the classroom. We need to understand why teachers change did not affect student skills. Effects on teachers at the end of the third year have not been explored yet, but whether teachers abandoned the use of the M@t.abel approach and materials we would have an explanation for the lack of effects on student math achievement.
Anyway, the simple planning and implementation of the evaluation has already achieved some effects: a substantial redesign of the program design itself; an increase of the interest in evidence-based research in the school context; the development of a network of researchers focused on this topic, promoting initiatives in the country.

Appendices

Not included in page count.

Appendix A. References

Baker, F. B. (2001). The basics of item response theory. Washington, DC : ERIC, Clearinghouse on Assessment and Evaluation.

Barzelay, M. (2001). "The New Public Management: Improving Research and Policy Dialogue", Journal of Comparative Policy Analysis, Vol. 4, No.1, March 2002, 95-105

Bates, D.M. and Pinheiro, J.C. (1998) "Computational methods for multilevel models" http://franz.stat.wisc.edu/pub/NLME/

Bloom H.S. (2007), "The Core Analytics of Randomized Experiments for Social Research" in Handbook of Social Research, Sage Publications.

Brookover, W.B., Beady, C., Flood, P., Schweitzer, J. and Wisenbaker, J. (1979). *School Social Systems and Student Achievement: Schools make a difference*, New York: Praeger.

Cohen, D. K., & Hill, H. C. (2000). Instructional policy and classroom performance: The mathematics reform in California. Teachers College Record, 102(2), 294–343.

Cohen, D. K., Raudenbush, S., & Ball, D. L. (2002). Resources, instruction, and research. In F. Mosteller & R. Boruch (Eds.), Evidence matters: Randomized trials in education research, (pp. 80–119). Washington, DC: Brookings Institution Press.

Darling-Hammond, L. (2000). Teacher Quality and Student Achievement. Education policy analysis archives, North America, 8, jan. 2000. http://epaa.asu.edu/ojs/article/view/392.

Darling-Hammond, L., Holtzman, D., Gatlin, S.J., & Heilig, J.V. (2005). Does teacher preparation matter? Evidence about teacher certification, Teach for America, and teacher effectiveness. Education Policy Analysis Archives, 13 (42). http://epaa.asu.edu/epa//v13n42/.

Davies, R. S., Williams, D. D. and Yanchar, S. (2008) The Use of Randomization in Educational Research and Evaluation: A Critical Analysis of Underlying Assumptions, Evaluation & Research in Education, 21:4,303-317

Kennedy, M. (1998). Form and substance of in-service teacher education (Research Monograph No. 13). Madison, WI: National Institute for Science Education, University of Wisconsin–Madison.

Mata, J., Machado, J. (1996). Firm start-up size: a conditional quantile approach. European Economic Review 40,1305–1323.

Mortimore, P. (2000b). "Does educational research matter?", British Educational Research Journal, Vol. 26 No. 1, 5-24.

OECD (2009), Creating Effective Teaching and Learning Environments - First Results from TALIS, Paris: OECD

Scheerens, J. (2000), “Improving School Effectiveness”, in UNESCO Fundamentals of Educational Planning, No. 68, UNESCO: Paris.

Scheerens, J. (2007). *Review And Meta-Analyses Of School And Teaching Effectiveness*. www.iqb.hu-berlin.de/lehre/dateien/rapportScherens.pdf

Scheerens, J. and Bosker, R.J. (1997). *The Foundations of Educational Effectiveness*. Oxford: Elsevier Science Ltd.

Towne L. and Hilton M. (2004), Eds. Implementing Randomized Field Trials in Education: Report of a Workshop, Committee on Research in Education, National Research Council, ISBN: 0-309-09192-6

Yu, K., Lu, Z., Stander, J. (2003). Quantile regression: applications and current research areas. The Statistician 52, 331–350.

Appendix B. Tables and Figures

Not included in page count.

Figure 1. M@t.abel – first year effect estimation

	Immediate exposure to Matabel			Delayed exposure to Matabel		
	2009/10	2010/11	2011/12	2009/10	2010/11	2011/12
1° cohort	3° media			3° media		
2° cohort	2° media	3° media		2° media	3° media	
3° cohort	1° media	2° media	3° media	1° media	2° media	3° media
4° cohort		1° media	2° media		1° media	2° media
5° cohort			1° media			1° media

Exposed to Matabel

Not yet exposed to Matabel

Figure 2. M@t.abel – second year effect estimation

	Immediatly treated			Treatment delay		
	2009/10	2010/11	2011/12	2009/10	2010/11	2011/12
1° cohort				8th grade		
2° cohort	7th grade	8th grade		7th grade		
3° cohort	6th grade	7th grade				

testing the assumption of cohort equivalence:

	Immediatly treated			Delayed treatment		
	2009/10	2010/11	2011/12	2009/10	2010/11	2011/12
1° cohort	8th grade					
2° cohort	7th grade				8th grade	
3° cohort					7th grade	

Table 1 – Average impact on student math performance

	Descriptive statistics		Effect estimates (OLS and IV regression)	
	Treatment	**Control**	**ITT**	**ATT**
Math score (mean)	493	496	1,8	4,8
Skipping items (%)	65,8	59,9	6,3**	16,7**
Not completing the test (%)	6,8	5,4	1,2	3,1

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

Table 2 – Average impact on student attitudes towards math and school

Dimensions	**Value controls**	**ITT**	**ATT**
Attitudes towards math			
5 items factor (std score)	-0,05	*+0,05*	*+0,12*
4 items factor (std score)	-0,05	*+0,05*	*+0,12*
In math I'm good (1-4 points scale)	+2,78	+0,05**	+0,12**
Curriculum pace			
We proceeded even if some classmates did not understand the topic (1-4 points scale)	+1,55	+0,07***	+0,17**
Causal Attributions			
Attribution of failures to bad luck (0-6 points scale factor)	+0,19	-0,04***	-0,09***
Test Anxiety			
4 items factor (std score)	-0,04	+0,05*	+0,13*
I was so nervous I could not find the answers (1-4 points scale)	+1,89	+0,06***	+0,16***

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