Abstract Title Page
Not included in page count.

Title: Repetita Iuvant? Lessons from repeated RCTs on the effectiveness of a teacher professional development program

Authors and Affiliations:
Aline Pennisi (Italian Ministry of Economy and Finance)
Gianluca Argentin (Milan-Catholic University)
Giovanni Abbiati (IRVAPP)
Andrea Caputo (Rome La Sapienza University)

Abstract Body

Limit 4 pages single-spaced.

Background / Context:

Description of prior research and its intellectual context.

Rigorous counterfactual evaluation is now commonly acknowledged by the scientific community to be the best way to inform policy makers and practitioners about *what works* and hopefully to guide public expenditures. The need for more and more robust evidence has hence become a common claim (e.g. Duflo and Kremer, 2003; Banerjee, 2007). However, the increasing amount of such kind of evidence on a growing number of programs is not making things easier when it comes to design or to implement a new policy. Unexplored scaling-up issues of pilot projects and low external validity of randomized control trails (RCTs) represent the main plagues of putting research results into practice (Moffitt, 2004; Attanasio et al., 2003). Moreover, most evaluations are designed to detect short-term effects, while many interventions deploy their effects over longer periods of time (Ravallion, 2009).

A remedy to these obstacles is replicating RCTs over time and assessing the differential effectiveness of the programs across contexts, although this might not always be financially or ethically feasible (Ravallion, 2009). Incentives to undertake study replications are few for researchers, given that scientific journals do not reward such products (Rodrick, 2009). Replications may also provide different results depending upon the site/cohort studied, the outcome measures, alternative treatments available (e.g., Minneapolis Domestic Violence Experiment) and there is increasing attention to the fact that empirical findings may be irreproducible because of random or systematic error (Open Science Collaboration, 2015). In sum, RCTs are a bronze standard more than a gold one (Berk, 2005) and difficulties to extract lessons for practitioners, especially in cases of contradictory evidence and zero effects, must not be neglected while conducting such evaluations.

One way to improve the usability is of research results is to include thorough implementation analysis, often neglected in the context of counterfactual evaluation literature. Among the reasons behind this lack of interest is the frequent use of small pilots, where settings are highly controlled. However, when the evaluation is run when the program is at-scale, implementation data is widely available and can be crucial to inform policy makers.

Purpose / Objective / Research Question / Focus of Study:

Description of the focus of the research.

This work summarizes the results of two RCTs aimed at evaluating the effectiveness of a professional development program for lower secondary school math teachers. The program, called M@t.abel, was financed by the Ministry of Education in Southern Italy with EU funds. It lasts a full school year and it is based on formal and on-line tutoring, providing alternative methods for teaching traditional math contents. The program and its evaluation were both held at scale, a situation far from artificial settings and allowing for the observation of real constraints faced in delivering such a service. This situation, along with the replication of the RCT on two consecutive cohorts of teachers, provides several opportunities and challenges both from the policy and research practice points of view.

The main outcome variable is student math achievement, but effects of the program on student attitudes are also explored (e.g., attitudes towards math, the motivation to study, the perception of the curriculum pace and text anxiety, etc.). A big effort in the collection of monitoring data allows us to better understand how the program works. Changes in the effectiveness of the program were detected from the first to second RCT. They are interpreted in the light of changes in some features of the program.

Setting:
Description of the research location.

The program was evaluate while already at scale and financed by the Ministry of Education in particolar for teachers in four regions of Southern Italy (Campania, Calabria, Apulia, Sicily). National and international evidence (including IEA and PISA scores) has been pointing out that math performance is lagging behind in these regions, starting from middle school. Given Italian teachers are amongst the oldest worldwide [OECD 2007], the majority does not have any specific training in teaching and many math teachers are "out-of-field". Hence, in service professional training was considered crucial by the Ministry of Education which allocated a conspicuous amount of EU funding to this program.

Population / Participants / Subjects:
Description of the participants in the study: who, how many, key features, or characteristics.

The evaluation was conducted through a repeated large-scale randomized control trial on two consecutive cohorts of teachers (2009-10, "first wave", 2010-11 "second wave"). About 250 schools, almost 900 teachers and over 15.000 students have been involved in the experiments. Both experiments lasted three years: teachers and their classes have been followed along the entire course of lower secondary school (three years in Italy, from grade 6 to grade 8).

Intervention / Program / Practice:
Description of the intervention, program, or practice, including details of administration and duration.

M@t.abel aims at providing math teachers operating in grades 6 to 10 with innovative instructional practices. It is based on the idea that math should become more appealing for the students and closer to their everyday life experience. The program lasts one full school year and it is based on formal and on-line tutorship, and on experimenting specific teaching materials in the classroom.

Teachers, organized in groups of about 15/20 participants, take part to eight face-to-face meetings with a tutor, for a total of 27 hours of work. The service location point is another (usually big) school with capacity for hosting the trainees and adequate computer and internet equipment. Participants also use virtual classrooms and online tools (for 40 hours of work), attend video-conference lectures, download specific teaching materials, chat with their tutor and peers to share their experience. Moreover, teachers can use forums and discussions groups to interact with colleagues attending the training course M@t.abel elsewhere.

The teaching units address four thematic areas (numbers and algorithms, geometry, relations and functions, data handling and statistics) and three main learning processes (measuring, posing and

solving problems, argumentation and conjecturing). Teachers are asked to implement at least four teaching units in their school classrooms during the training year and to fill out a structured diary for each unit proposed to their students. Enrollment to the course, as well as attendance, is free for teachers. There is no incentive for participation, i.e. the certificate obtained does not lead to any career or monetary advantage.

Some changes occurred between first at second wave, namely as concerns: enrolment procedures, grades involved, quantity of teaching units available, implementation instructions and acceptance of non-tenured teachers (notwithstanding the randomization procedure). Teachers and students in involved in both waves are on average similar according to the main observed variables.

Research Design:
Description of the research design.

In both cases, the RCTs were built on a delayed treatment principle (no candidate to the program was excluded, but controls were set to participate to the program one year after the treatment group). For the first wave, schools were stratified according to geographical criteria and by peer participation intensity (less than five peers, at least five peers from the same school). Schools were randomly assigned to the control group (with all their teachers), proportionally to the distribution in sample layers. The observed class was randomly chosen among grades 6-8 (stratified). For the second wave, randomization was conducted within blocks defined by the service location points. Control schools were randomly selected within each block, until the number of teachers enrolled in the same service location point reached a given threshold. This way we were able to maximize the number of control teachers but at the same time preserving the minimum required number of participants for the course to be started. The observed class was randomly chosen (but restricted to grades 6). Moreover, a pre-test on student math achievement was conducted. Finally, for the second wave, for reasons independent from the RCT the program was suspended after the first year, so control teachers remained in that condition with no contamination possible.

The RCTs both waves proved high internal validity (tested a large set of variables collected at school, class, teacher and student level); moreover, on a large set of observables, our sample is representative of schools, students and teachers of the whole Southern Italy, but not of the rest of the country (still bearing the fact that the program is voluntary, so teachers self-select into the treatment).

Data Collection and Analysis:
Description of the methods for collecting and analyzing data.

Every school year students were administered a math assessment along with a background questionnaire used by the Italian national education evaluation institute (INVALSI). Math tests were scheduled in -May, i.e. at the end of each school year. In spite of the lack of a national student register, a longitudinal panel of students was built. During the second wave an additional math assessment was set at the beginning of the first year (pre-test) in order to increase the statistical power of the experiment and to check for baseline math competences.

Teachers were surveyed via CATI. Primary data were integrated with administrative information on teachers and schools (variables on the geographical context and the socio-economic composition of their students). Finally, data were also collected on the tutors, the first time through a structured interview (via CATI for the first wave, via CAWI for the second wave) and the second time through focus groups.

Findings / Results:
Description of the main findings with specific details.

Despite the presence of promising results on student attitudes (Argentin et al. 2014), in the first wave M@t.abel had no impact on students' math achievement neither in the short nor in the medium run in the first wave: effects are little in size and not significant. The situation slightly differs for the second wave: we observe a significant (although small) effect at the end of the first year (Table 1).

As concerns student attitudes, the first wave showed at the end of the first year more favorable attitudes towards math, a higher sense of responsibility among students for their own learning, a lower tendency to attribute failures to bad luck and a faster perception of the curriculum pace. These effects were not detected in the second wave experiment, while test anxiety proved stronger among treated students all three years and in both waves.

A main difference between waves is teachers' take-up rates, increasing from 39 to 45% between waves. Self-selection determinants of compliance are similar between waves (except age; probit models) and organizational features rather than individual features seem to have counted in ensuring more compliance in the second wave.

The higher teacher take-up rate in the second wave contributes to reduce the dilution of the treatment. Moreover, teachers report more satisfaction for the program in the second wave and, according to various measures, they interact more with their peers. There are hints that teachers were in the second wave more aware of how to implement the M@t.abel approach in the classroom and more capable of integrating it into the regular curriculum. On the other hand, the increased availability of teaching units in the second wave does not seem to count: more than 75% of the materials used in classroom are made of only 3 different teaching units and most of the preferred units are the same in both waves.

Conclusions:
Description of conclusions, recommendations, and limitations based on findings.

In this evaluation, researchers managed to engage with policymakers and practitioners in the planning stages of the program and before even starting, the evaluation contributed to program design (streamlining the treatment, focusing on contents and ensuring collaboration among the different institutions). Randomization through delayed treatment resulted acceptable under technical and ethical profiles and response rates were high from both treatment and control groups, and from both teachers and students. The repeated RCTs bearing different results made the task for researchers to communicate results particularly difficult and requested an increase effort in collecting and analyzing implementation data.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Argentin, G., Pennisi, A., Vidoni, D., Abbiati, G., & Caputo, A. (2014). Trying to Raise (Low) Math Achievement and to Promote (Rigorous) Policy Evaluation in Italy Evidence From a Large-Scale Randomized Trial. *Evaluation review*, *38*(2), 99-132.

Attanasio, O. P., Meghir, C., & Santiago, A. (2012). Education choices in Mexico: using a structural model and a randomized experiment to evaluate Progresa. *The Review of Economic Studies*, *79*(1), 37-66.

Banerjee, A. V., & He, R. (2007). *Making aid work* (pp. 91-97). Cambridge: MIT press.

Duflo, E., & Kremer, M. (2005). Use of randomization in the evaluation of development effectiveness. *Evaluating development effectiveness*, *7*, 205-231.

Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American journal of community psychology*,*41*(3-4), 327-350.

Guskey, T. R. (2009). Closing the knowledge gap on effective professional development. *Educational Horizons*, 224-233.

Lendrum, A., & Humphrey, N. (2012). The importance of studying the implementation of interventions in school settings. *Oxford Review of Education*,*38*(5), 635-652.

Mofiitt, R. A. (2006). Forecasting the effects of scaling up social programs: An economics perspective. *Scale-up in education: ideas in principle*, *1*, 173.

Ravallion, M. (2009). Evaluation in the Practice of Development. *The World Bank Research Observer*, *24*(1), 29-53.

Rodrik, D. (2009). The New Development Economics: We Shall Experiment, but How Shall We Learn?. In Jessica Cohen, and William Easterly, eds., *What Works in Development? Thinking Big and Thinking Small*, Washington: Brookings Institution Press

Appendix B. Tables and Figures

Not included in page count.

Table 1 - Average Impact on Student Math Performance

	First wave		Second wave	
Year	ITT	ATE	ITT	ATE
I	0,01	0,02	0,13*	0,24
II	-0,08	-0,19	0,06	0,12
III	-0,02	-0,01	0,11	0,25

Note: standard errors clustered at the class level; coefficients expressed in terms of effect size