Consiglio Nazionale delle Ricerche

IRCrES

ISTITUTO DI RICERCA SULLA CRESCITA ECONOMICA SOSTENIBILE
RESEARCH INSTITUTE ON SUSTAINABLE ECONOMIC GROWTH

Working Paper

Numero 10/2018

**Evaluating social innovation: results and emerging issues
from a random-trial evaluation of a program
for the inclusion of migrant adolescents**

Valentina Lamonica, Elena Ragazzi, Lisa Sella

Evaluating social innovation: results and emerging issues from a random-trial evaluation of a program for the inclusion of migrant adolescents

VALENTINA LAMONICA, ELENA RAGAZZI AND LISA SELLA

CNR-IRCRES, National Research Council. Research Institute on Sustainable Economic Growth, Via Real Collegio 30, Moncalieri (TO) – Italy

corresponding author: lisa.sella@ircres.cnr.it

ABSTRACT

This paper presents the experimental evaluation of a pilot program for the educational inclusion of foreign adolescents with low language skills and high dropout risk in junior high school. The intervention aims at filling migrant students' gaps in both general subject knowledge and character skills. The pilot is jointly developed by a private VET (Vocational Education and Training) center for adolescents in Turin, Italy, and some local junior high schools.

During the pilot, treated students attend 290-hour classes at the VET center. VET classes aim at restoring students' interest in education by professionally-oriented teaching, inductive innovative pedagogical methods, and individual mentorship.

The pilot is evaluated through experimental counterfactual design. In the recent past, a similar preliminary project had positive results, but its evaluation was based on qualitative evidence and on teachers' subjective assessment. Now, a grid of outcome variables is proposed to monitor students in the treated and control groups regarding their educational inclusion and scholastic achievement. The aim is to assess social and behavioral skills, alongside school performance.

Previous experimental evaluations of similar programs (see Kautz et al., 2014 for a survey) provide ambiguous results: depending on program features such as length of the treatment, role and pervasiveness of training, tutors, mentors, families' and peers' influence, long-term analyses showed either positive (Durlak et al., 2011; Tierney et al., 1995; Orr et al., 1994) or negative (Rodrigues-Planas, 2012) effects on some outcome variables such as behavioral attitudes, character skills, educational attainment, future wages and other labor market outcomes. For this reason, the project adopts a randomized experimental design, so as to get sound evidence on the effectiveness of the implemented activities.

HOW TO CITE THIS ARTICLE

Lamonica, V., Ragazzi, E., & Sella, L. (2018). Evaluating social innovation: results and emerging issues from a random-trial evaluation of a program for the inclusion of migrant adolescents. *Working Paper IRCrES*, *4*(10). http://dx.doi.org/10.23760/2421-7158.2018.010

CONTENTS

Evaluating social innovation: results and emerging issues from a random-trial evaluation of a program for the inclusion of migrant adolescents

VALENTINA LAMONICA, ELENA RAGAZZI AND LISA SELLA

1 INTRODUCTION AND MOTIVATION

The present case study describes some preliminary results of an experimental evaluation design assessing the impact of a pilot VET (Vocational Education and Training) program (named "Oltre i Muri", OiM from now on) addressed to non-EU migrant adolescents at high dropout risk in junior high school[1]. The project and its evaluation are currently underway.

The intervention is supposed to reduce school dropout risk by a multifaceted treatment that concerns the whole educating community (students, teachers, and families). All students involved in the pilot attend the second junior high class and experience problems in school integration, from both a pedagogical point of view and in the social perspective as well. This is fundamental, since scarce school attendance is an important dropout predictor, most of all among young adolescents, and may be the basis for the development of bullying behavior[2]. During the pilot, treated students attend VET classes at the training center for two days a week, instead of standard school lessons, for a total of 290 hours. VET classes are characterized by strong work-oriented contents (bread-maker, chef's assistant…) and aim at restoring students' interest in education by professionally oriented teaching, innovative inductive pedagogical methods, and individual mentorship.

The impact of the pilot is evaluated by a counterfactual experimental approach. The design is intended to assess the effects of OiM on educational integration of migrant students. This is fundamental since eventually positive results could encourage an extension of the program to a large number of junior high schools. In fact, despite growing migration dynamics of kids and adolescents, Italian compulsory school is still far from being truly effective in migrants' integration, needing complimentary support, e.g. from targeted public programs or private actions from no profit organizations.

In particular, the evaluation activity is expected to give back a multiple contribution:

- Fine tuning of both the pilot general approach and the pedagogical instruments (internal mission)

---

[1] The pilot was conceived and implemented in Turin, Italy, by Piazza dei Mestieri, a private training centre for adolescents, and some State junior high schools in the city. It was funded by the Italian Home Office through FAMI funds.

[2] Recent VET projects on Italian bully students ("CODS" and "Peer-CODS", implemented by Piazza dei Mestieri) showed promising results, although their assessment was solely based on qualitative evidence from teachers' judgement.

- Assessment of perceived results to
  - o enforce the sharing of pedagogical practices (trans-institutional and transnational mission)
  - o strengthen the proposed model for future wider implementation (perpetuation mission).

So, the experimental approach is here conceived as an instrument for learning-by-doing in action-research experiences. It must be nevertheless underlined that its implementation interferes with the practices and interests of the different stakeholders involved. This causes difficulties that may induce subtle distortions, even hampering the accomplishment of the trial.

Previous experimental evaluations of similar programs (for a survey see Kautz et al., 2014) provide ambiguous results. Depending on some characteristics of the training, such as its duration, the role and pervasiveness of tutors and mentors, the influence of families and peers, long-term analyses showed either positive (Durlak et al., 2011; Tierney et al., 1995; Orr et al., 1994) or negative (Rodrigues-Planas, 2012) effects on outcome variables such as behavioral attitudes, character skills, educational attainment, future wages, and other labor market outcomes.

For this reason, a critical issue of OiM evaluation is the definition of the outcome variables, since the pilot aims at migrants' educational and social integration, which are multidimensional objects. The set of outcomes was fundamentally developed based on some tools successfully implemented in a previous large educational project[3] and, more in general, in psychometric analysis. The set has been monitored for both treated and control students, assessing social and behavioral skills, alongside school performance.

The rest of the paper is organized as follows. Section 2 frames up in the literature the problem of social and educational inclusion of young migrants; section 3 describes the pilot, focusing on the target population, the treatment and the specific pedagogical approach adopted; section 4 describes the experimental design, discusses some relevant issues of its implementation and the final set of outcome variables, which has been developed with the fundamental contribution of the VET operators, who represent an innovative element of the evaluation experience. Section 5 and 6 discuss some preliminary results, since the project is currently underway. Section 7 concludes.

2   MIGRANTS' EDUCATIONAL INTEGRATION IN THE LITERATURE (PRELIMINARY)

In order to prevent and tackling ELET (Early Leaving from Educational and Training), it is fundamental to detect the main factors that contribute to causing it. On the one side school-based explanations are proposed (school segregation), on the other side individual, family and socio-economic characteristics (Thibert, 2013).

Having a migrant background is one of the most relevant predictor of early leaving: the percentage of ELET among foreign students (born abroad) in Europe is still much higher than among natives, twice as high (EACEA, 2014). In Italy, this percentage is about three times higher: in 2016, the native students leaving educational and training paths were about 11.9% versus the 32.8% foreign students (Bonini & Santagati, 2016). The disadvantage of migrant students clearly emerges also when analyzing NEET (Not in Education Employment or Training) data, which refer to the age range 15-29. In Italy, this condition concerns the 34.0% foreign students, but "only" the 23.2% natives (Lodigiani & Santagati, 2017). It is important to notice that the foreign origin *per se* brings further sources of disadvantage (European Commission, 2013). Insufficient language skills and the lack of parental support (in terms of cultural and economic capital) are the most significant. In particular, the ability to master the host country language is fundamental to overcome the disadvantage in the school environment,

---

[3] The project "Provaci ancora, Sam!" has been implemented for twenty years by Fondazione per la Scuola, Italy, and it aims at preventing educational failure by specific intervention at school (http://www.provaciancorasam.it/progetto).

which represents the first step for social inclusion. While OECD countries can work on language skills, the parental support is an acquired condition, which is very difficult to improve.

Since 2009, European countries implement different measures and programs in order to reduce ELET rates by three types of action: prevention, intervention, and compensation (EACEA, 2014). Prevention policies are not directly oriented to reduce ELET rates, but to support students who are at risk (i.e., education guidance, positive discrimination measures). Intervention policies are addressed to students who have experienced school failures or difficulties, in order to engage them and prevent the dropout. These interventions are focused on individual reinforcement (i.e., language support for migrant students, scholastic support for low achiever, absenteeism management, networking with parents and other actors). Finally, the third type of policy gathers compensation measures aimed at reintegrating dropout students (i.e., second chance education system).

The OiM pilot project fits in between intervention and compensation measures, because it involves a target population at high dropout risk and it provides language support, re-motivation and new chances for remaining in the educational path. The main innovation of the pilot is its ability to transfer a VET intervention program to lower secondary education students, in order to prevent dropout events.

In this regard, the OiM pedagogical approach is supported by literature from projects in several countries that mainly show two kinds of evidence. First of all, the importance of small class size to reduce students' disadvantage, specifically for migrants, minorities, and pupils with poor socio-economical background (Schanzenbach, 2014; Andersson, 2007; Nusche, 2009; Björklund *et al.*, 2005). Secondly, the relevance of including teachers with migrant background in the classroom. In fact, they can positively influence educational outcomes but also students' engagement.

Sharing with a teacher or tutor (as in OiM specific case) the same migration path can develop students' self-confidence and increase their motivation (Carrington & Skelton, 2003). Furthermore, if students show poor linguistic skills, the tutor mainly becomes a linguistic mediator. This approach aims to overcome the methodological debate between the idea that migrants' educational achievement passes through their mother language and the opposite theory promoting no use of mother language in the educational context to include foreign students (Christensen & Stanat, 2007).

3 THE "OLTRE I MURI" PROJECT

3.1 The target population

The evaluation exercise concerns a pilot program promoted by a non-profit foundation dealing with weak young students and early school leavers (ELET). In particular, the OiM project targets young migrant students included in State classes of the lower secondary school[4].

Attending school is compulsory until the age of 16, even though one can absolve this duty also in training or apprenticeship paths rather than in the upper secondary school. It is worthwhile observing that high ELET rates are experienced in Italy, mostly in the passage from lower to upper secondary school. For this reason, some innovative projects aiming at preventing dropout start in the lower secondary school. In the context of students at high ELET risk, young migrants having recently arrived from non-EU countries suffer even greater difficulties, due to strong language barriers.

The Italian approach to educational integration of young migrants is to include them in a standard educational path, so as to maximize their social inclusion from the beginning. The main drawback with this approach is represented by the poor knowledge of the Italian language. On their arrival, young foreigners should attend non-curricular courses of Italian language (for

---

[4] In Italy, education is structured as follows: Primary school starting at the age of 6 and lasting 5 years; Lower secondary school, lasting 3 years; Upper secondary school, lasting 5 years, with different paths.

instance, Italian law sets that classes of Italian ought to be offered in refugees' reception centers), but their effectiveness is low. At the same time, the competencies of school teachers' are not adequate to teach Italian to foreigners, and most of all they have not enough time to supply target services at school, above all when the share of foreign students in the class is high.

In all these cases, teachers acknowledge reduced participation to school activities, non-satisfactory improvements in knowledge and skills, and the natural appearance of ethnic subgroups (mostly in schools with high migrants' share). These in turn represent a problem, both because they hinder the desired melting of cultures, and because these groups are generally outbreaks of violence and bullying behavior. Hence, the main paradox is that poor language skills represent a high barrier for both learning and social inclusion, which represents the main aim of migrants' immediate enrolment in general education.

In this regard, it is interesting observing that the initial project idea addressed bullying behavior of foreign students (regardless their date of arrival and Italian skills), but the involvement of the stakeholders convinced the project manager to change the target. In fact, teachers agreed that low language-skilled students represent a major problem, that cannot be properly addressed by the school team alone, and that requires an innovative approach to prevent future problems of early school leaving and deviance.

3.2   The evaluated pilot program

The OiM project is financed by the FAMI (Asylum, Migration and Integration Fund – EU funding) program in the context of an action intended to ease the national and transnational diffusion of best practices in the field of migrant hospitality and inclusion. For this reason, the project has a twofold aim. The first objective concerns the target population, and consists in the inclusion of young migrants at the social and school levels to prevent dropout. Beyond that, the project aims to reinforce the local (schoolteachers' empowerment) and transnational community (exchange of best practices and design of common tools).

The evaluation exercise focuses on the effectiveness of an innovative dual path alternating classes at school and at the training center. OiM main objective is the social inclusion of young migrants in the school environment, in order to prevent the ELET. As previously described, the project targets non EU-students with manifested behavioral disorders (for instance the tendency to be excluded by their classmates, low language skills, etc.) as indicated by their teachers. The specificity of the project consists in the alternation of school and training courses during the week. Through this alternation, students are stimulated to break their stalemate and they are (re)motivated by the new and dynamic learning context.

The intervention model elaborated by Piazza dei Mestieri is characterized by three steps. First of all, the VET center takes in charge the student and his family. Then, a personal project (school guidance and specific needs) is developed by the training center in agreement with both the provenance public school and the family. Finally, innovative pedagogical activities are carried out to promote student's inclusion.

Concretely, students are involved in a 290-hour path at the training center Piazza dei Mestieri. In particular, students attend courses for a total of 90 hours devoted to basic school subjects (Italian, Mathematics, and English) and 200 hours of professional laboratories (cooking, graphic design and bartender). This educative model is based on a work-related and inductive notion of knowledge and learning, which means that students are stimulated to learn theoretical subjects through laboratory practice. As an example, Mathematics is learned during the cooking practice: fractions and proportions are easily understood in practical contexts.

Beyond that, the pedagogical approach is individually targeted: each student can in fact rely on a tutor during VET classes, whose function consists of educational support and linguistic mediation. In order to promote social inclusion, peer tutors got involved (i.e., older migrant students, attending the training center, that share their personal experiences and encourage the inclusion process through peer learning). Thanks to these professionals and peer-tutors, students' motivation can improve, producing positive effects on school paths. This is a big

issue, since the project lasts about one year and, after this period, each student ought to be able to come back to traditional school being less exposed to ELET.

An interesting feature of the pilot is represented by the so-called "experiential weeks" during which students and their schoolteachers are involved in joint laboratory activity. The aim is to make students share with their teachers their process of change, reversing the traditional educational path (from teacher to pupils) into a non-conventional one, where teachers can learn from their pupils (e.g. how to make a coffee or a pizza in the laboratories) and students can teach to their professors. Thanks to these activities, teachers can realize their students' efforts and changes while students can appreciate their new skills that may be played for an improved self-esteem, in the relations with peers and teachers at school.

4   THE EVALUATION DESIGN: METHODOLOGY AND MOTIVATION

The OiM pilot is intended to understand the impact on educational attainment and social integration of a training program addressed to non-EU migrant students at high dropout risk. Hence, the evaluation design should answer the following questions:

- *Is the training program effective for migrant students?* (general effectiveness)
- *Are there sub-groups of students for whom the program is more (less) effective?* (conditional effectiveness)
- *How large is this effect?* (impact quantification)

The experimental approach is rather fashionable, since it represents a fair (causal) test[5]. It is aimed at estimating the average impact of the treatment on participants. However, in social field experiments, causal inference can be particularly challenging in terms of four issues: experimental setting-up, intrinsic variability, randomization and substitution biases (Heckman & Smith, 1995). All these points were tackled in the OiM pilot.

Concerning the experimental setup, the VET operators expressed strong resistance to randomization, since their well-established *modus operandi* is a cherry-picking selection of students from a large bulk of applicants, which is not compatible with a correct impact evaluation. However, the intrinsic nature of the project was particularly suitable to field experimentation, which allows the researchers to estimate the impact of pilot interventions on small samples in real-world settings (Shadish *et al.*, 2002; Duflo, 2006). Hence, a convenient randomization design was finally established.

Concerning the intrinsic sampling variability, both small sample bias and sample heterogeneity (school vs. CPIA) can be controlled for by blocked-randomization design, which guarantees unbiasedness and improves precision (Gerber & Green, 2012). According to this design, the N applicants are divided into J blocks, based on the observable characteristics that are supposed to predict the outcome (in the present case, the educational institution of origin[6]). In each block, individuals are randomly assigned to the treated and control groups (respectively m_j treated units and N_j-m_j not treated units) and the block-specific Average Treatment Effect (ATE_j) is unbiasedly estimated by the difference between average outcomes[7] Y, i.e.

$$\widehat{ATE}_j = \left[\frac{\sum_{i=1}^{m_j} Y_i}{m_j}\right] - \left[\frac{\sum_{i=m_j+1}^{N_j} Y_i}{N_j - m_j}\right].$$

Then, the overall ATE is estimated by a weighted average of the block-specific ATEs, i.e.

---

[5] Random assignment implies that both observable and unobservable variables that affect outcomes are equally likely to be present in both the treatment and control groups.

[6] Previous school grades and behavioural indicators, which could be highly predictive, are not available for some students.

[7] Unbiasedness is guaranteed by randomization whenever excludability and non-interference assumptions are fulfilled. The excludability assumption is satisfied as long as non-treated students do not receive any extra-treatment at school. The non-interference assumption is satisfied as long as non-treated subjects are not affected by the treatment of other subjects, which is the case of person-specific training programs.

$$ATE = \sum_{j=1}^{J} \frac{N_j}{N} ATE_j.$$

The same reasoning can be applied to difference-in-differences (DID) estimators, which compare the change between pre-pilot and post-pilot outcomes in the treatment and control groups. Hence, in each block,

$$E\left(\widehat{ATE_j}\right) = E(Y_i - X_i | D_i = 1, J = j) - E(Y_i - X_i | D_i = 0, J = j),$$

where D represents treatment and X the pre-test value. DID estimators are particularly useful whenever pre-tests are highly predictive, since the strategy allows gains in precision that are equivalent to a relevant increase in sample size, compared to the difference-in-mean estimator.

Since the OiM pilot is characterized by a very small sample and different sorts of outcomes (counts, scores, marks), hypothesis testing is assessed by randomization inference (Heβ, 2017; Good, 2005), i.e. it is not based on large-sample theory, rather the p-values are calculated based on the inventory of a large number of possible randomizations.

Concerning other common biases, no substitution was needed, but non-random attrition occurred when a couple of counterfactual students dropped-out from school in the middle of the pilot. It is a result *per se*, since the pilot was mainly aimed at contrasting school dropout, and no case occurred in the treated group. However, missing students' outcomes were no longer collected, seriously undermining ATE inference based on those outcomes (Dinardo *et al.*, 2006; Huber, 2012). A cautious and simple approach to deal with non-random attrition in such small samples consists of placing extreme bounds on the ATE, i.e. estimating the largest and smallest ATEs by filling the missing values by extremely high and extremely low outcomes. This approach imposes few assumptions, but it generates large extreme bounds (Gerber & Green, 2012).

4.1   Random-block assignment in practice

Since the pilot offers just a very small number of VET positions (15), eligible students could be easily divided into treated and control groups by random assignment techniques. Students' heterogeneity between schools and CPIAs was managed by random-block assignment. Practically, teachers from schools and CPIAs recommended 43 non EU students that could enter the pilot. Then, the VET operators selected the eligible ones by individual colloquia, excluding 8 students that did not fulfil their pre-requisites. Then, the eligibles were divided into two blocks (j=school, CPIA), consisting of $N_{school}=22$ and $N_{CPIA}=13$ students. Since CPIA students are generally more subject to home transfers, the majority of VET positions were assigned to school students ($m_{school}=11$, $m_{CPIA}=4$) in order to prevent a failure of the evaluation. Then, each block was randomized by complete assignment, i.e. random numbers were assigned to students in each block, then students were sorted and the first m_j were selected into treatment. In each block, a counterfactual student moved away at the very beginning of the program and was not replaced. In this case, attrition is random and independent from potential outcomes; hence, it does not affect ATEs estimates[8].

---

[8] This is the so-called MIPO assumption, Missing Independent of Potential Outcomes (Gerber & Green, 2012). It implies that missing observations can be simply discarded in the ATE estimation.

Table 1. Eligible students' distribution by block and treatment status. Absolute values

| | | Assigned treatment status | | | |
		Treated	Not treated	Ex ante attrition	Total by block
Blocks	**CPIAs**	4	8	1	13
	Schools	11	10	1	22
	Total by status	15	18	2	35

Table 1 illustrates the final distribution of the treated and control groups. The sample is very small, but randomization guarantees unbiased inference of average treatment effects (ATEs) based on difference-in-mean and DID estimators. Moreover, the blocked random assignment guarantees unbiased estimates by block (Gerber & Green, 2012), allowing us to estimate reliable ATEs on the school students only, when the CPIA students left the pilot before its conclusion, i.e. at the end of their CPIA program. Hence, the rest of the paper focuses on the school students, who completed the program.

4.2   Measuring OiM outcomes: tools and indicators

Since the OiM pilot mainly aims at the *social and educational inclusion* of non-EU adolescents, the evaluation design investigates two classes of indicators, respectively focusing on educational attainment and behavioral issues.

The educational attainment is assessed by both school attendance and basic knowledge scores. In fact, the target population exhibits high dropout risk and the pilot aims at both contrasting early dropout behavior and at strengthening migrants' basic knowledge skills, i.e. their knowledge level in fundamental school subjects as Italian language, English language, and Mathematics. Indicators and tests were collected both at the beginning and at the end of the project (DID estimation approach). In the educational context, the impact evaluation should answer the following questions:

- *Is the pilot effective in contrasting dropout and early school leaving among migrant students?*
- *Does the pilot strengthen students' basic knowledge skills?*
- *Does the pilot hinder students' learning, since they miss school lessons?*

Social inclusion is two-fold investigated: on the one side, schoolteachers fill an hetero-evaluation questionnaire, because they closely observe students in the socio-educative context. The tool was borrowed by a kin project aimed at preventing school dropout: it mainly investigates students' relationships with peers and adults, and their general behavior at school, providing an idea about scholastic inclusion and individual behavior that could eventually cause social exclusion.

On the other side, students' soft skills are assessed by a self-evaluation pictographic questionnaire (Maćkiewicz & Cieciuch, 2016; Barbaranelli *et al*., 2003; Heckman & Kautz, 2012). The questionnaire was originally conceived for children aged 7-10, but it represents a satisfactory solution to overcome migrants' language barriers. In fact, the pictographic form mixes written sentences and images, thus facilitating comprehension. The questionnaire aims to identify five behavioral dimensions that are considered strong predictors of social inclusion: extraversion, agreeableness, conscientiousness, openness, and neuroticism (Ahadi & Rothbart, 1994).

In the social inclusion context, the impact evaluation should answer the following questions:

- *Is the pilot effective in improving students' social integration?*
- *Does the pilot improve students' relationships with peers and adults?*
- *Does the pilot strengthen students' soft skills?*

5 OiM EFFECTS ON EDUCATIONAL ATTAINMENT. SOME EVIDENCE FROM ATTENDANCE RATES AND TEST SCORES

This section analyses educational attainment by means of attendance rates and scores from basic subject tests (Italian language, English language, Mathematics).

The attendance rate is a crude but important indicator, since the target population exhibits high dropout risk. The OiM pilot carefully addresses this issue by offering person-specific intervention based on treated students' needs, as observed by the VET operators. At the end of the experiment, all treated students still attended school, while two out of ten non-treated students dropped-out. In practice, if the treatment effect TE is defined as the difference in attendance rate between the treated and control groups, the OiM effect is $TE = \frac{11}{11} - \frac{8}{10} = 0.2$, i.e. the treatment causes a 20% increase in students' attendance rate. However, the result is poorly significant, showing a p-value=0.2168 in a 100000-permutation Montecarlo test[9].

On the contrary, test scores give more refined outcomes concerning educational attainment. In principle, the OiM pilot aims at producing positive effects on students' skills, since it involves innovative, person-specific, and inductive teaching methodologies. However, the program provides just 30 hours per basic subject out of 290, which is little amount to produce any measurable impact on knowledge. Moreover, schoolteachers were afraid of eventual negative impacts, since treated students miss their lectures when attending the pilot. In this context, test score evidence is particularly important.

Tests were developed by VET teachers and administered twice, at the beginning and at the end of the pilot, aiming at assessing the effects on students' basic knowledge in a DID evaluation framework. This is particularly important in small samples, when pre-tests are highly predictive of outcome values[10] (Gerber & Green, 2012). Both the Italian and the English language tests were designed to assess the students' level of comprehension in a communicative context, based on the Common European Framework of Reference for Languages (Verhelst *et al.*, 2009). Exercises contain a different number of items and weights, in order to evaluate skills on a 0-100 scale. All tests produce similar results[11], hence we just discuss Italian scores.

Entry scores in Figure 1 upper panel confirm the heterogeneity hypothesis between school and CPIA students, which affects outcome variables and justifies the random-block assignment design: school students (square-shaped) systematically perform better than CPIA students (diamond-shaped), showing an higher knowledge level at the beginning of the program. Moreover, the lower panel shows a pre-treatment gap between (lower performing) treated and controls within the school block, which is purely random and statistically not significant[12].

---

[9] The test involves 28.3% of all possible permutations, given the actual experiment size. Results show that, under the sharp null hypothesis of no treatment effect $Y_i(D_i = 0) = Y_i(D_i = 1)$, there is a 21.7% chance of obtaining the same TE estimate simply by chance.

[10] Basic subject scores are supposed to be highly persistent over the course of a year (i.e., the pilot duration). In fact, in the counterfactual group Italian pre-test scores predict 83% of the variance in the final test outcomes, as measured by the R^2 obtained from the regression of $Y_i(0)$ on the corresponding pre-test values.

[11] The Pearson's correlation coefficients between basic subjects' pre-tests are in the range [0.88; 0.93].

[12] The Welch's t-test statistic for independent treated and control samples in the school block (Figure 1, lower panel) allows to accept the null hypothesis that treated and controls have the same pre-treatment average score.

Figure 1. Italian test entry scores: school vs. CPIA blocks (upper panel) and treated vs. non treated in the school block (lower panel)
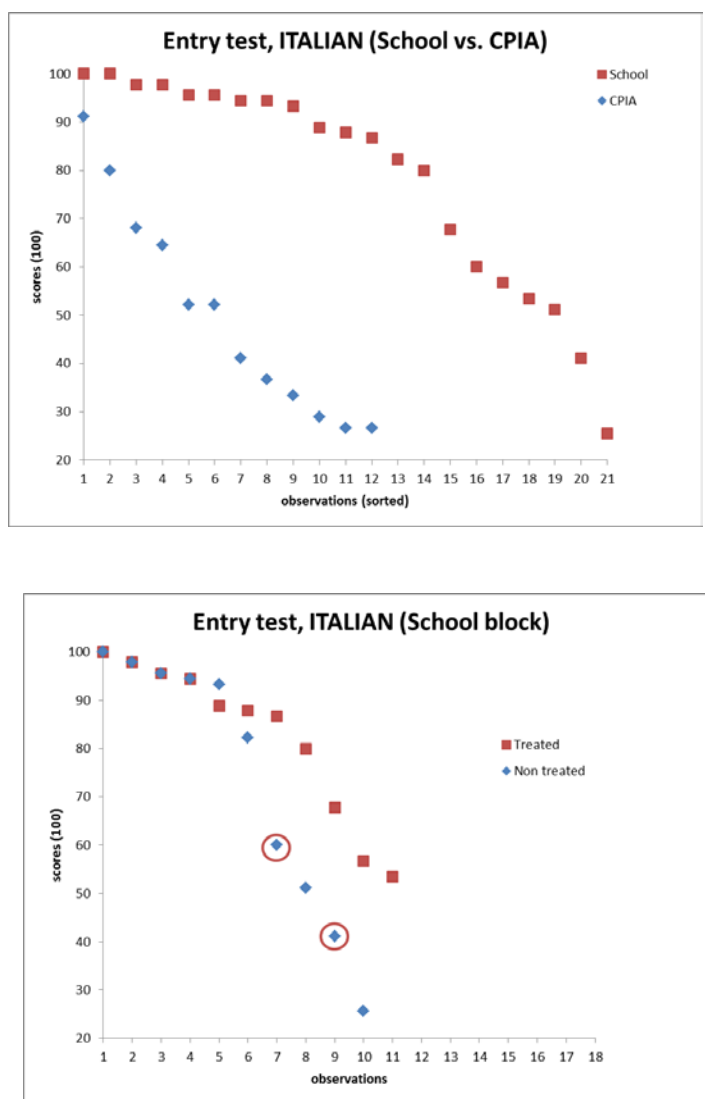
As explained, impact assessment is limited to school records. Hence, the ATE can be estimated by the difference in the average post/pre score gaps between the treated and controls. However, the two red circles in Figure 1 lower panel indicate that non-random attrition occurred, i.e. two students in the counterfactual sample did not report exit scores because they dropped out. This fact can cause biased ATE estimation, which is solved by extreme value bound techniques (Manski, 1999).

The method does not impose strong functional assumptions, but generally, it produces wide and weakly informative bounds. In practice, missing outcomes are filled with the highest (lowest) possible outcome to estimate the upper (lower) ATE bound. In this specific case, we can take advantage of preliminary pre-test information, which is highly predictive and places missing students' scores in the middle of the counterfactual distribution[13]. Hence, missing outcomes are filled at first with the highest and then with the lowest post/pre gap observed in the counterfactual group. Then, the upper (lower) ATE bound is assessed by computing the

---

[13] Italian scores are in fig. 1, lower panel. Other scores are available upon request.

difference between the observed average factual gap and the lower (higher) imputed average counterfactual gap.

Calculations show that the estimated OiM ATE is bound between -3.2 and +6.2, i.e. on average the treated students show score gaps between 6 points higher and 3 points lower than the controls. The same calculations on final scores (rather than on gaps) estimate the ATE between -0.9 and +12.2, i.e. on average the treated students show final scores that are between 12 points higher and 1 point lower than the controls' scores on a 0-100 scale basis. Hence, results cannot confirm a strictly positive effect of the pilot on students' basic knowledge; however, the range of the effect is considerably committed towards positive gap values. Moreover, the ATE based on final scores show that in the worst case the controls averagely perform 1/100 points better than the treated, which is a very negligible difference.

These result are highly compatible with the aims of the pilot, which is mainly devoted at reinforcing migrant adolescents' character skills, rather than their basic knowledge.

6   OiM EFFECTS ON SOFT SKILLS. (VERY PRELIMINARY, BASED ON INITIAL DATA)

6.1   Hetero-evaluation questionnaire

It was possible to collect the hetero-evaluation data for State school pupils only, since CPIA teachers have no full familiarity with their students. The questionnaire is administered to teachers and investigates three main aspects: relationship with peers, relationship with adults, and behavior at school of the students they recommend for the project (see sec. 4.2). For each student, teachers express an ordinal judgment (-9, 0, 1, 2) about a set of behavior of interest: e.g., score 2 signals that the issue deeply affects the student, while score -9 signals that it is rather a strong point.

Table 2. Hetero-evaluation of selected students from State school, average scores by subset

Type	# items	Average scores
Relations with peers	14	0,62
Relations with adults	5	0,71
Learning	22	0,81

Table 2 shows the average score by subset for students selected into the project. Statistical tests reveal no significant difference between the factual and counterfactual groups[14], hence scores are considered globally. Analyzing the whole indicator set, the most problematic aspects emerge in learning. In fact, participant students do distinguish (median greater than 1) for difficulties in reading, writing, speaking.

Spearman's rank correlation analysis reveals several significant relations (95% confidence level). First of all, it emerges that negative attitudes towards both peers (schoolmates) and adults (teachers) are truly connected and related in turn to negative attitude towards school activity in general (e.g., scarce rule respect, no participation, no class-work, no homework). In particular, a clear correlation structure emerges between the set of indicators measuring relations with adults and a subset of learning indicators concerning general attitude towards school and school performance itself.

In teachers' opinion, students having logic-mathematical problems generally show lack of communication abilities (reading/writing/speaking), as well as lack of interest and low rule respect. Students who do not easily understand school tasks show difficult peer and adult relations as well, self-excluding behavior, poor school performance, and lack of attention. Timid

---

[14] Statistically significant differences at the 95% confidence level were detected in one indicator out of 41 by rank sum test for ordinal variables.

pupils manifest anxiety and difficulties in handling school failures, while low self-esteem pupils do not generally ask for schoolmates' help, and are frequently distracted. Finally, low family involvement in sons' scholastic activity is not significantly related to any negative attitude, apart from inadequate care for scholastic materials.

Polychoric principal component analysis (Kolenikov & Angeles, 2009) reveals four components[15], explaining 61.0% total variance. Analyzing loadings, an interesting interpretation emerges involving psychologists' attachment theory (Bowlby, 1988), which correlates secure attachment relations to teachers and significant adults with higher academic attainment, better self-regulation, and social competences (Geddes, 2006; Commodari, 2013). This is particularly important in the case of young migrants, who are generally subject to attachment issues due to stress and trauma that derive from home and school moving.
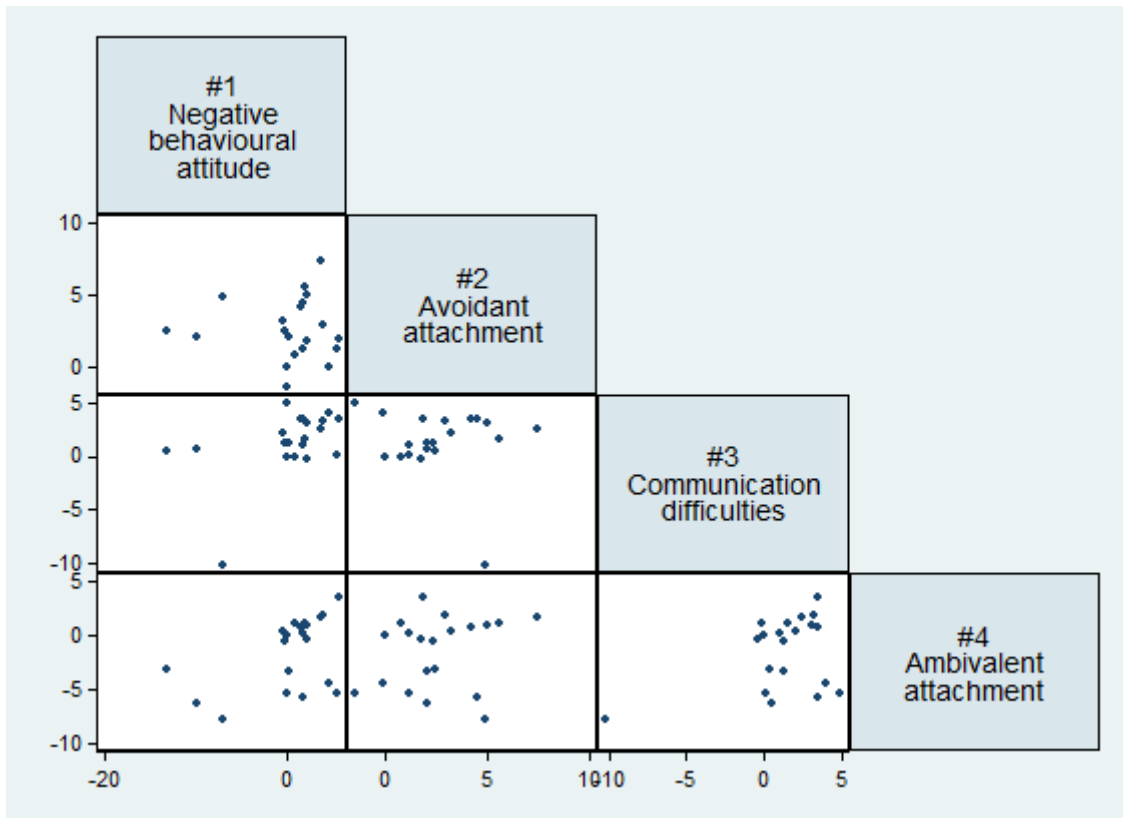
Table 3. Polychoric PCA (oblimin oblique rotation): Pearson's correlation between PCs and entrance test scores for non-CPIA students

Polychoric rotated component and interpretation	Total explained variance (%)	Correlations			
		#1	#2	#3	#4
# 1: Negative behavioural attitude	22,8	1,000			
# 2: Avoidant attachment	16,6	-0,012	1,000		
# 3: Communication difficulties	11,2	0,467**	-0,161	1,000	
# 4: Ambivalent attachment	10,4	0,476**	0,220	0,381*	1,000

*90% confidence level; **95% c.l.; ***99% c.l.

Practically, two components are consistent with behavioral traits, while the remaining two can be associated to attachment styles. The first component can be interpreted as "Negative behavioral attitude", towards both peers and teachers: in the sample, it correlates to low pupils' scholastic commitment. The second component is consistent with an "Insecure avoidant attachment" profile and it correlates to low academic achievement (Geddes, 2006). The third component is mainly related to "Communication difficulties" (reading/writing, speaking, logical-mathematical), mostly in a linguistic sense. Finally, the last component is consistent with an "Insecure ambivalent attachment" style, characterized by both positive peer relations and teachers' attention-seeking behavior. Table 4 shows that the chosen PCA decomposition implies significant correlations between individual scores from ambivalent attachment (comp. #4) and both communication difficulties (comp. #3) and negative behavioral attitudes (comp. #1). On the behavioral side, Figure 2 shows that individuals that behave negatively (high comp. #1 scores) show strong communication difficulties too (high comp. #3 scores): both aspects are particularly evident in individuals that show high ambivalent attachment scores. On the contrary, no clear relation emerges in the case of avoidant attachment. In line with psychological literature, the relation between avoidant and ambivalent attachment reveals that pupils can show mixes of both styles, rather than a clear polarization (Bergin & Bergin, 2009).

---

[15] Components are selected by Horn's test (Dinno, 2009) and rotated by the oblique oblimin criterion. The analysis was performed on a reduced set of variables, in ordered to reduce the noise due to low correlations.

Figure 2. Pairwise plot of polychoric Principal Component scores

Finally, it is worth noting that ambivalent attachment significantly correlates to (Italian/English) entrance test scores (Table 4 and Figure 3), while communication difficulties have no effect at all.

Table 4. Polychoric PCA (oblimin oblique rotation): Pearson's correlation between PCs and entrance test scores for non-CPIA students

Polychoric rotated component and interpretation	Entrance tests correlations		
	Italian	English	Mathematics
# 1: Negative behavioural attitude	0,171	-0,121	0,046
# 2: Avoidant attachment	0,250	-0,191	0,069
# 3: Communication difficulties	0,003	0,093	-0,232
# 4: Ambivalent attachment	0,735***	0,404*	0,296

*90% confidence level; **95% c.l.; ***99% c.l.

Figure 3. Scatterplot of entrance test and ambivalent attachment scores
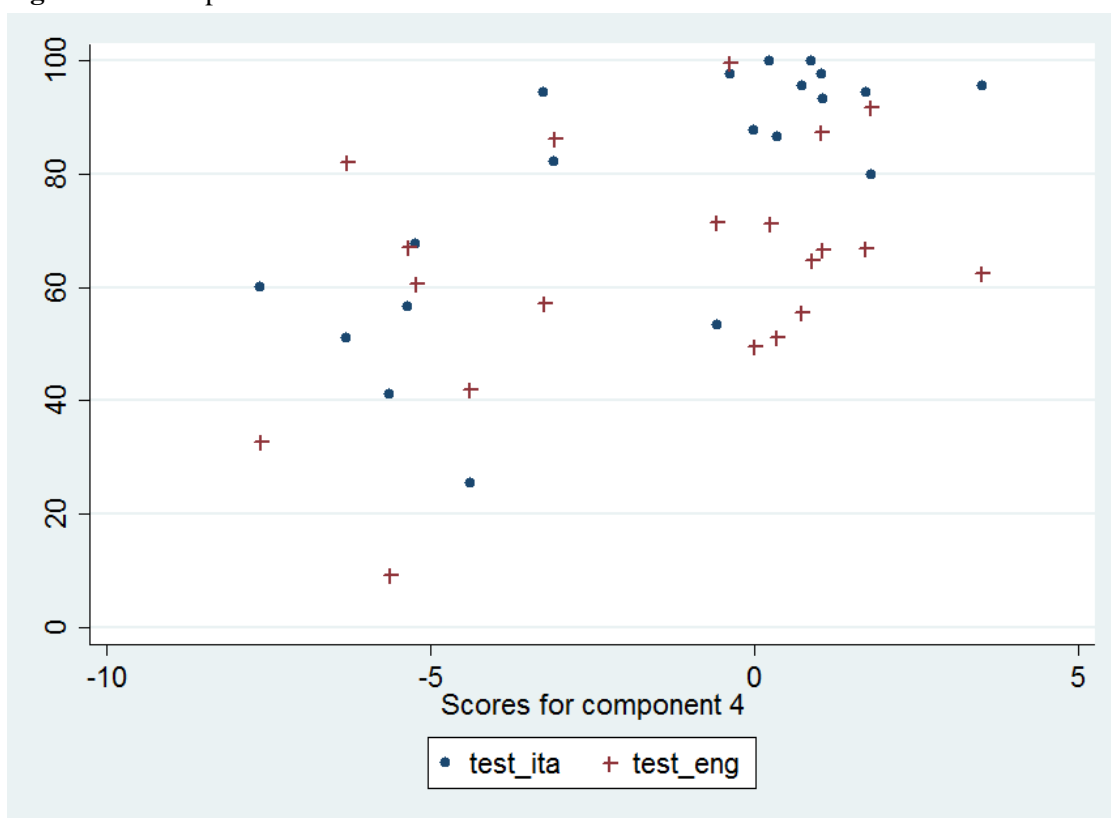
6.2 Self-evaluation questionnaire

This section shows results from pictographic questionnaires for State school pupils, that investigate soft skills by 15 items associated to the "Big 5" personality trait notion. Factors (Openness, Consciousness, Extraversion, Agreeableness, and Neuroticism) are described by three Likert-scale ordinal (1-5) items each. Questionnaires were collected at the beginning of the program, hence they describe the *ex ante* standpoint. Analyzing overall median values, it emerges that most students have friendly attitude when helping people and noticing friends' needs (median value 5 for each item). Moreover, no student declares to be worried when going to school or to avoid sharing their gifts.

Comparing factual and counterfactual distributions, the Mann-Whitney (1947) test detects statistically significant differences at the 95% confidence level in both calm/worried attitudes and amusement behavior. When aggregating item scores by personality trait (Maćkiewicz & Cieciuch, 2016), significant differences emerge in both Neuroticism and Extraversion (Table 5): counterfactuals appear to be a little more extroverted and less neurotic than the factual sample. Overall, agreeableness is the more represented trait. Nevertheless, differences are random and due to small sample.

Table 5. Personality trait statistics by participants' group and rank-sum test. Aggregated score range: 3-15

Personality trait	Factual		Counterfactual		M-W test
	Mean	Min/Max	Mean	Min/Max	p value
O: Openness	9,5	6/14	10,2	5/15	0,702
C: Consciousness	10,1	5/13	11,2	7/15	0,283
E: Extraversion	9,5	6/12	12,7	4/15	0,011**
A: Agreeableness	12,7	10/15	14	12/15	0,061*
N: Neuroticism	8,4	5/13	5,6	3/9	0,011**

*90% confidence level; **95% c.l.; ***99% c.l.

Analyzing personality trait correlations (Table 6), it emerges that Neuroticism is negatively correlated with all other traits, while Extraversion significantly correlates with Openness, and Consciousness with Agreeableness. Concerning entrance tests, Mathematics scores are highly positively correlated with Neuroticism, and negatively correlated with Extraversion (Figure 4).

Table 6. Pearson's correlation coefficients between personality traits and entrance test scores

Personality trait	O	C	E	A	N	IT	ENG	MAT
O: Openness	1,00					-0,04	0,31	-0,33
C: Consciousness	0,30	1,00				-0,24	0,13	-0,38*
E: Extraversion	0,47**	0,18	1,00			0,03	0,16	-0,48**
A: Agreeableness	0,25	0,43*	0,36	1,00		-0,24	-0,25	-0,27
N: Neuroticism	-0,20	-0,38*	-0,45**	-0,24	1,00	0,42*	0,10	0,70***

*90% confidence level; **95% c.l.; ***99% c.l.

Figure 4. Scatterplot of Mathematics entrance test scores and personality traits (Extraversion and Neuroticism)
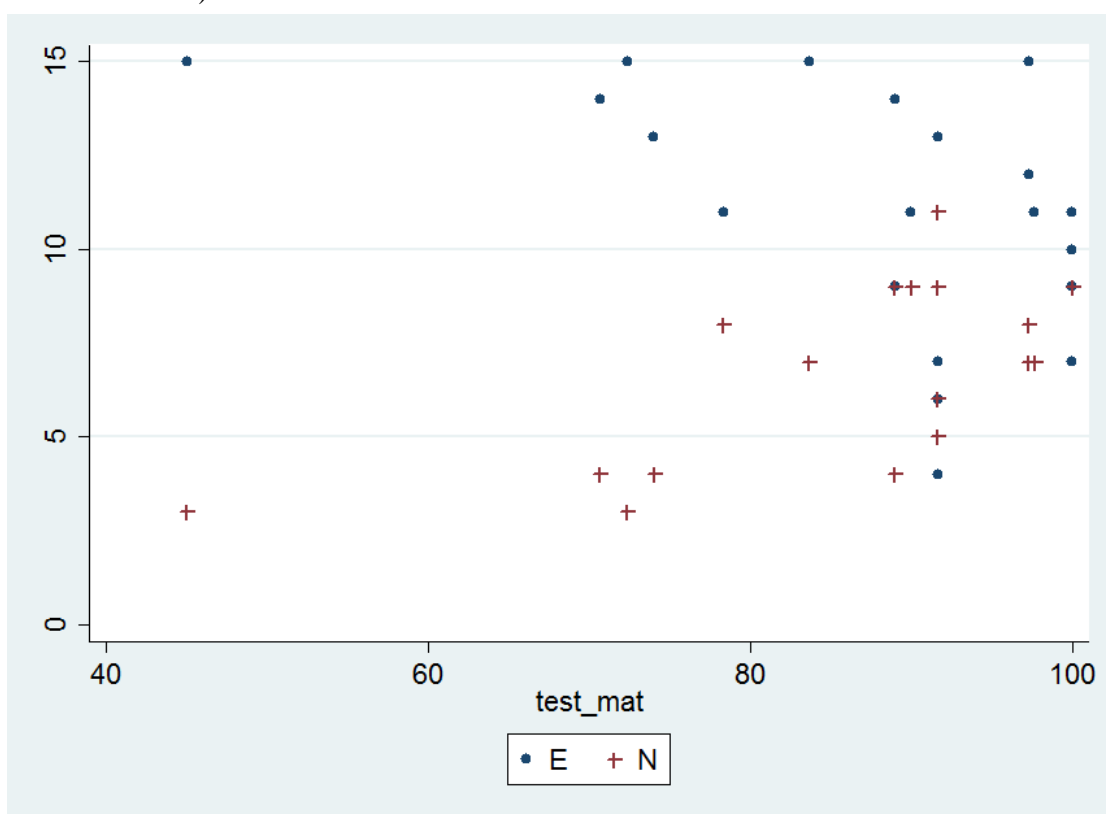
Finally, analyzing the polychoric principal components derived from the hetero-evaluation questionnaires, it emerges that Extraversion is positively and significantly correlated with "Communication difficulties" (comp. #3), while Consciousness is negatively correlated with "Negative behavioral attitude" (comp. #1)[16].

7   PRELIMINARY CONCLUSIONS

7.1   Lessons for a good implementation of the random trial evaluation

This paper discusses the evaluation experience of a project aimed at a better school integration of young migrants, which is still in progress. For this reason, it is currently not possible to show the results of impact evaluation, and final considerations are limited to the lessons deriving from the evaluation design implementation.

Project effectiveness will be evaluated by means of an experimental approach. We did not rely on quasi experimental techniques for many reasons, including the small size of the trial, the lack of suitable variables for matching, the important role played by unobservables in policies that address disadvantaged people.

On the other hand, it is well known that random trial implementation may be difficult, because it has to be planned from the beginning of the project, it may face declared opposition or unexpressed resistance by the operators implementing the policy, and it may still suffer of selection bias in the case of non-random substitution of treated individuals.

We may acknowledge that all these situations occurred and they had to be properly faced. There only case of non-random substitution was excluded from the evaluation sample.

---

[16] The correlation coefficient between Extraversion and "Communication difficulties" is 0.40, while between Consciousness and "Negative behavioural attitude" it is -0.,39, both at the 90% confidence level.

However, the most interesting lessons derive from the other issues, i.e. planning/timing and resistance. Starting discussing the initial implementation, the counterfactual evaluation has been included in the project because its promoters showed a high commitment in understanding the effectiveness of a very experimental approach to ELET. In its initial design, the project should have lasted one school year, including a fair time to implement the random trial before the beginning of the teaching activities. Due to problems in the program management by the Ministry, there was a 5-month delay in the kick-off.

	Sept	Oct	Nov	Dec	Jan	Feb	Mar	April	May	June July Aug	Sept	Oct	Nov	Dec	...
School year	██	██	██	██	██	██	██	██	██	Vacation time	██	██	██	██	██
Scheduled timesheet	K.O.	Trial design				Laboratories					Evaluation				
Actual timesheet						K.O.	Trial design		Laboratories		Laboratories		Evaluation		

This fact caused a number of practical implications. For example, it was impossible to include in the trial last-year students, who would not have been able to attend laboratories in the next school year, so sharply reducing the number of candidates. Moreover, the project was designed so as to follow the school calendar, with a lot of advantages for both the project itself, and the evaluation task. For example, the possibility to rely on school notes as an outcome indicator is sharply reduced, because in the actual timesheet it is hard observe an impact at the end of the first cycle of laboratories (June 2017), while at end of the project final school reports are not available. Activities for a medium term evaluation of the project effects on school performance in June 2018 are possible, but the cost will not be covered.

Finally, the most impacting consequences were suffered by the random trial design. In fact, preliminary implementation activities suffered a hard rush-up, since teaching activities (laboratories) may not be compressed. Consequently, it was difficult to select in a short time a fair number of candidates, thus reducing the homogeneity of the selected target (e.g., by the inclusion of CPIA students). Moreover, both selection and enrolment procedures had to heavily rely on the promoter, so the evaluator could not supervise all steps. Finally, some activities that should have occurred sequentially, actually overlapped (e.g., entrance tests administered during laboratory classes). Unfortunately, we observed some effects on the results of entrance tests (see section 5). Since they were administered after selection procedures, different students' commitment induced a small positive bias in favor of the treated individuals. This inconvenience will be mitigated by the diff-in-diff method, which estimates changes in the objective variables, measured both at the beginning and at the end of the treatment.

To conclude, timing is fundamental. This is often true in policy evaluation (e.g., policies addressed to firms that are subject to business cycle), but it is crucial in educational policies, which have to coordinate with the school calendar. If this fails, both the project effectiveness and the evaluation feasibility are threatened.

As far as the relationship with implementing actors, we received full support by the top managers, who agreed on the initial design, i.e. the random assignment to main and counterfactual groups of all students recommended by the schools. On the contrary, VET operators had the unexpressed idea to assign non-suitable students to the control group, causing a serious selection bias. In general, they refused to include whatever candidate without a preliminary selection of the most suitable ones and insisted on creating a well-balanced treatment group. The final solution, that was suitable for both the operators and the evaluators, was to define at first a list of eligible students, then to assign them randomly to either treatment or control groups. This procedure reduced the sample size, but ensured both the evaluation reliability, and project feasibility.

All above considerations prove that a serious preliminary work aiming at creating a common agreement on objectives and procedures is fundamental in random trial evaluation. This common wisdom is particularly true when the client of the evaluation service is a non-

government organization, as in this case, which cannot be forced to adapt its empirical strategy. However, the preliminary brainstorming step is in the evaluator interest as well: in fact, social innovators generally need non-conventional evaluation strategies, because they generally produce very innovative and advanced experiences.

Since the target population of pilot projects is often fairly unknown (few available statistics), a great preliminary work is also needed to specially design the measurement tools. Our experience, that merged researchers', teachers' and psychological counsellors' competencies, shows that it remains very difficult to correctly calibrate the instruments for the measurement of the output/outcome variables. Both time and budget have to be invested in these activities.

In the specific case of young migrants with scarce school performance and social inclusion, we learnt that they are a very heterogeneous group. Some are fairly skilled, but it is difficult to detect their skills for both the linguistic barrier and the low social inclusion. The entrance tests had to include a much wider range of levels than firstly expected regarding teachers' files. Their judgement in this case failed, because of the scarce integration of their pupils.

8   REFERENCES

Andersson, C. (2007). Teacher Density and Student Achievement in Swedish Compulsory Schools, *Working Paper* (4). Uppsala: Institute for Labour Market Policy Evaluation (IFAU).

Barbaranelli, C., Caprara, G. V., Rabasca, A., & Pastorelli, C. (2003). A questionnaire for measuring the Big Five in late childhood. *Personality and Individual Differences*, *34*(4), 645–664.

Bergin, C., & Bergin, D. (2009). Attachment in the Classroom. *Educational Psychology Review, 21,* 141–170.

Björklund, A. M., Clark, M. A. Edin, P. A. Fredriksson P., & Kreuger, A. B. (2005). *The Market Comes to Education in Sweden: An Evaluation of Sweden's Surprising School Reforms*. New York: Russell Sage Foundation.

Bowlby, J. (1988). *A Secure Base: Parent-Child Attachment and Healthy Human Development*. London: Routledge.

Bonini E., & Santagati M. (eds,) (2016). Primo Rapporto. *Osservatorio sulla dispersione scolastica delle scuole secondarie di secondo grado del Comune di Milano*. Milano: Fondazione ISMU, Comune di Milano.

Carrington, B., & C. Skelton (2003). Re-thinking 'Role Models': Equal Opportunities in Teacher Recruitment in England and Wales. *Journal of Education Policy*, *18*(3), pp. 253–265.

Christensen, G., & Stanat, P. (2007). Language policies and practices for helping immigrants and second-generation students succeed. The Transatlantic Taskforce on Immigration and Integration. *Migration Policy Institute (MPI) and Bertelsmann Stiftung*, 1–15.

Commodari, E. (2013). Preschool teacher attachment, school readiness and risk of learning difficulties. *Early Childhood Research Quarterly, 28*, 123– 133.

Devaux-Spatarakis, A. (2014). L'expérimentation « telle qu'elle se fait »: leçons de trois expérimentations par assignation aléatoire. *Formation emploi*, 126, Avril-Juin, 17–38. http://formationemploi.revues.org/4180.

Di Nardo, J., McCrary, J., & Sanbonmatsu, L. (2006). Constructive proposals for dealing with attrition: An empirical example. *NBER Working Paper*, July 21, 1–46.

Dinno, A. (2009). Exploring the sensitivity of Horn's parallel analysis to the distributional form of random data. *Multivariate behavioral research*, *44*(3), 362–388.

Duflo, E. (2006). Field experiments in development economics. *Econometric Society Monographs*, 42, 322.

Dunn, O. J. (1964). Multiple comparisons using rank sums. *Technometrics, 6*, 241–252.

Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. B. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child development*, 82(1), 405–432.

European Commission/EACEA/Eurydice/Cedefop, 2014. *Tackling Early Leaving from Education and Training in Europe: Strategies, Policies and Mesures.* Eurydice and Cedefop Report. Luxembourg: Pubblication Office of the European Union.

European Commission, 2013. *Study on Educational Support to newly arrived migrant children.* Luxembourg: Pubblication Office of the European Union.

Geddes, H. (2006). *Attachment in the Classroom: the links between children's early experience, emotional wellbeing and performance in school.* London: Worth Publishing.

Gerber, A. S., & Green, D. P. (2012). *Field experiments: Design, analysis, and interpretation.* New York: WW Norton.

Good, P. I. (2005). Permutation, parametric and bootstrap tests of hypotheses. *Springer Series in Statistics.*

Heckman, J. J., & Kautz, T. (2012). Hard evidence on soft skills. *Labour economics*, *19*(4), 451–464.

Heckman, J. J., & Smith, J. A. (1995). Assessing the case for social experiments. *Journal of economic perspectives*, 9(2), 85–110.

Heβ, S. (2017). Randomization inference with Stata: A guide and software. *The Stata Journal*, 17(3), 630–51.

Huber, M. (2012). Identification of average treatment effects in social experiments under alternative forms of attrition. *Journal of Educational and Behavioral Statistics*, 37(3), 443–474.

Kautz, T., Heckman, J. J., Diris, R., TerWeel, B., & Borghans, L. (2014). *Fostering and measuring skills: Improving cognitive and non-cognitive skills to promote lifetime success* (No. w20749). National Bureau of Economic Research.

Kolenikov, S., & Angeles, G. (2009). Socioeconomic status measurement with discrete proxy variables: Is principal component analysis a reliable answer? *Review of Income and Wealth*, *55*(1), 128–165.

Lodigiani R., & Santagati M. (2017). NEET e Garanzia Giovani. Una politica di empowerment per i più svantaggiati? In Alfieri S., & Sironi E. (eds.). *Una generazione in panchina. Da NEET a risorsa per il paese* (pp. 267-276). Milano: Vita e Pensiero.

Maćkiewicz, M., & Cieciuch, J. (2016). Pictorial Personality Traits Questionnaire for Children (PPTQ-C)—A New Measure of Children's Personality Traits. *Frontiers in psychology*, 7.

Mann, H. B., & Whitney, D. R. (1947). On a test whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics, 18*, 50–60.

Manski, C. F. (1999). *Identification problems in the social sciences.* Harvard University Press.

Nusche, D. (2009). What Works in Migrant Education? A Review of Evidence and Policy Options. *OECD Education Working Papers*, 22. Paris: OECD Publishing.

Orr, L. L., Bloom, H., Bell, S., Lin, W., Cave, G., & Doolittle, F. (1994). *The national JTPA study: Impacts, benefits, and costs of Title II-A*. Bethesda: Abt Associates.

Rodriguez-Planas, N. (2012). Longer-term impacts of mentoring, educational services, and learning incentives: Evidence from a randomized trial in the United States. *American Economic Journal: Applied Economics*, *4*(4), 121–139.

Rothbart, M. K., Ahadi, S. A., & Hershey, K. L. (1994). Temperament and social behavior in childhood. *Journal of Abnormal Psychology, 103*(1), 55-66.

Shadish, W., Cook, T. D., & Campbell, D.T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Schanzenbach, D.W. (2014). Does Class Size Matter? *National Education Policy Center*, February (1).

Thibert, R., 2013. *Le décrochage scolaire: diversité des approches, diversité des dispositive. Dossier d'actualité*. Lyon: ENS de Lyon, Veille & Analyses IFÉ, *84*(5).

Verhelst, N., Van Avermaet, P., Takala, S., Figueras, N., & North, B. (2009). *Common European Framework of Reference for Languages: learning, teaching, assessment.* Cambridge University Press.