VIII.
Challenges and opportunities of counterfactual evaluation in a school setting: lessons learned

Marcello Cabria, Renzo Carriero, Alessia Rosa

Introduction

This chapter aims at illustrating and discussing the results of the evaluation activity concerning the field application of the KIDS4ALLL method. To provide tangible data on the effectiveness of diverse tools implemented across heterogeneous educational environments, delivering an evidence-based assessment was a pivotal part of the project. For this purpose, and to strengthen the impact assessment, the project team implemented a counterfactual research design. The evaluation of the actions carried out involved two samples (intervention and control groups), with the aim of observing whether the methodologies applied during the KIDS4ALLL's activities have had an impact on its final targets. The primary goal was to compare the treated and control samples, assessing any differences between the two groups. This involved specifically measuring whether adolescents from the treated sample showed variations in pre-selected dimensions compared to those who didn't participate in the project activities.

Accordingly, the assessment focused on measuring changes in three key dimensions of the project: socio-emotional competences, active citizenship competence, and perception of inclusion. The evaluation took place in two phases – pre-test at the project's outset and post-test upon completion – utilizing the same sample of young people involved in the project. Additionally, this phase aimed to refine an assessment tool based on competences tests previously experimented within existing literature.

The counterfactual analysis was conducted in a few middle and secondary schools in Turin,[1] a choice closely aligned with the evaluation design

1 A similar approach is adopted also in Israel, but such experience is not described in this contribution.

requirements. The formal educational context provided better control over all involved variables and enhanced results comparability. In contrast, informal settings, such as voluntary associations, presented higher heterogeneity, with respect to both the timing of the activities and the targets of the experiment, considering age and personal characteristics of the youth involved in the program.

The chapter begins by examining the objectives and challenges associated with employing a counterfactual analysis design within school contexts, like the ones considered. The subsequent section provides a detailed account of our specific assessment process, encompassing school selection, sample construction, and evaluation design. Following this, we conduct a comprehensive presentation of the results, also discussing a multivariate analysis to highlight causal connections leading to outcomes and to grasp their implications. Notably, the conclusion of our work revealed substantial changes in the measured competence levels among the student sample, prompting a thorough discussion later in this chapter. This discussion entails a detailed analysis of the strengths and weaknesses associated with our work. As a counterfactual-type evaluation imposes rigorous constraints on both the conducted intervention and the working criteria, an in-depth review of empirical limitations influencing observed outcomes will provide a solid foundation for refining further analyses in similar contexts. To wrap up, the chapter provides insights for future interventions, drawing from the experiences gained throughout the project.

1. Monitoring and impact evaluation in school contexts

When we speak of evaluation of an intervention like an educational initiative, we may think of very different activities implying the assessment of the achieved results, the implementation process or both. Achieved results can be measured by means of empirical indicators regarding the beneficiaries of the initiative, using "objective" data such as some kind of measurable performance/characteristic, or beneficiaries' subjective assessments of the initiative in which they were involved (e.g., a satisfaction survey). The implementation process can be evaluated by reviewing and analysing step-by-step all the activities that were envisaged and how they actually unfolded. To perform assessment activities, we can figure out different tools such as questionnaires, interviews, observations, focus groups, data analyses and so on. However, a crucial distinction in this field concerns the overarching

goal of evaluation rather than its tools. On the one hand, the goal of evaluation can be *monitoring*, i.e. identifying the aims and objectives of the intervention, selecting the empirical indicators to be used to observe progress against the objectives, and setting the target to be achieved through the intervention. On the other hand, the goal of evaluation can be conceived of as an assessment of the *effect* of an intervention on its beneficiaries. Underlying this conception is the idea that what we observe *after* the intervention is not necessarily its result or consequence. When the goal of evaluation is to assess the *causal effects* of an intervention, we speak of *impact evaluation* (Khandker et al., 2010).

Both evaluation goals, monitoring and impact assessment, are important, legitimate, and worth pursuing. They are also complementary to some extent, as monitoring activities regarding outcomes and process implementation may help us explain why an intervention actually did or did not produce the expected effects that we measure by means of impact evaluation. Yet, impact evaluation poses specific challenges that make it an extremely risky business, especially in a formal educational context like school. In general terms, if we want to claim that an intervention had an effect on a given outcome, we need to rule out that other concomitant confounding factors may have generated such effect. In other words, our claim requires to make a robust causal inference which implies answering the following question: what would have been the outcome measured on the participants, had they not participated in the intervention? To answer this question, we need to identify what it is called the *counterfactual*: an outcome that we cannot directly observe on the *same individuals* that were exposed to the intervention, as one cannot *at the same time* be exposed and not exposed to the intervention.

Here comes the main challenge of impact (or counterfactual) evaluation, that is how to identify a proper counterfactual.

How to deal with this issue? Under the right circumstances, it is possible to *create*, rather than look for, a counterfactual, by means of randomization. A group of subjects is randomly selected from the target population and assigned to the intervention (i.e., invited to participate or enrolled in a program); another randomly selected group, from the same population, is assigned to the so-called "control" condition and is not exposed to the intervention or enrolled in a program. Outcomes are measured on both groups, before and after the intervention. Given that subjects in the two groups are randomly selected, *if the sample size is large enough*, the differences before the intervention will be very small and random, i.e. non-sys-

tematic. For this reason, the control group can be legitimately considered a proper counterfactual: it represents how the participants, as a group, would look like, had they not participated in the intervention. The average difference in outcomes between the two groups, measured *after* the intervention, represents the net effect (or impact) of the intervention itself.[2]

The kind of impact evaluation just described looks very like a classical social-psychological experiment, but unlike the latter, it is carried out in the field rather than in the lab. This implies a few complications that limit the applicability and challenge the validity of the method, especially in a school setting. Let's briefly review the main challenges.

First of all, in a school setting, children are embedded in classrooms and hence it is not usually possible to select a few of them at random and exclude the others from participating in the program, because that would disrupt the normal functioning of the classroom group. Thus, the intervention or (henceforth) the treatment is administered at classroom level, meaning that classrooms, not individuals, are randomly assigned to treatment or control. This is not a problem *per se*, but makes impact evaluation more burdensome in organizational terms. Indeed, when treatment assignment is made at classroom rather than individual level, this creates clustering of individuals within the treated and control groups. To properly account of clustering and to preserve the statistical power of the analysis (i.e. the capacity of detecting statistically significant differences between groups) a much larger sample size is required compared to the situation where individuals rather than classrooms are randomized. Even if in principle this is not a serious threat, it is clear that with limited resources impact evaluation drains much of the available financial and human resources and make the whole evaluation process much more onerous.

In the second place, random selection of classrooms may be difficult to accept by teachers and principals. Using this device a few pupils will be excluded from a potentially beneficial educational intervention, just for the sake of assessing whether it is really effective. While this may be appealing and worthwhile from a scientist's viewpoint, it is much less so from the teacher's or principal's perspective. Moreover, teachers and principals may be legitimately convinced that only specific classrooms need or can benefit

2 It is worth underlying that this way of assessing the impact of an intervention concerns *group differences*, not individual ones. What we measure is an *average effect* because for any single individual it is not possible to find a proper counterfactual, unless some usually unrealistic assumption is made.

from a certain educational intervention, which would suggest avoiding random selection. Therefore, staff's resistances can only be won by providing control (excluded) classrooms with the same opportunity of benefitting from the initiative, but at a later time, when the impact evaluation task is accomplished (see Ballarino et al., 2022 for an example). Again, this makes impaction evaluation quite onerous.

Another issue concerns one of the assumptions on which the correct identification of an intervention's causal effect relies, that is the independence and (physical) separation between subjects assigned to treatment and subjects assigned to control. This is necessary to avoid reciprocal influences that alter subjects' assigned status and consequent outcome.[3] In a setting like a school, where pupils have opportunities to interact not only within but also between classrooms, this assumption is less likely (though not impossible) to be met (see Barone et al., 2021 for an example).

Finally, all interventions that take place in the field are exposed to vulnerabilities due to unforeseen events (e.g., changes in teaching staff, classroom's participation in other competing initiatives, classroom's drop out from the initiative, etc.) that might alter or even disrupt the correct way of bringing about the experimentation.

2. Impact evaluation in the KIDS4ALLL project

After shortly reviewing the main issues concerning the application of impact evaluation, it should be clear that impact or counterfactual evaluation imposes strong and stricter requirements compared to other types of evaluation because the overarching goal is different and more ambitious. However, while the ambition of assessing whether an intervention actually produced (or induced) a change in its beneficiaries is certainly justified by the necessity of wisely invest our limited resources, such an ambition may be easily frustrated if impact evaluation is applied under circumstances that stretch its feasibility beyond the limits. In the next paragraphs, we describe how we addressed these issues in order to best apply the impact assessment given the actual constraints.

3 This assumption is known as SUTVA (Stable Unit Treatment Values Assumption) in the econometrics literature (Rubin, 1980).

2.1 *Target selection and sample size definition*

The recruitment and enrolment of participating schools and classes followed a few criteria established ex-ante. Firstly, we were interested in schools having a substantial number of students with migratory background[4]. This criterion was crucial for assessing and improving students' perceived levels of inclusion, especially those with this specific characteristic.

A second focal point was the selection of classes within institutes. In lower secondary schools (covering a 3-year time period), priority was given to second-grade classes (12-year-old students), to avoid a sample consisting of students who were too young. At the same time, they had to be of a different age from students in the upper secondary schools involved: this approach facilitated the evaluation of the method impact on clearly defined age groups. For secondary school, our interest was in students from the first two years (14-15 years old), as they were best suited for the proposed teaching materials and activities.

Geographic distribution in urban areas constituted another selection criterion, as the idea was to encompass both central and peripheral (or semi-peripheral) districts. This strategy aimed to mitigate the social stratification inherent to different areas, thus preventing excessive social homogeneity within the sample.

Furthermore, again to increase social heterogeneity, the intervention aimed to incorporate different school tracks (vocational, technical, lyceum), covering the whole spectrum of educational paths available in Italy.

Given these required features, the actual selection had to adapt ultimately to the availability of teachers and school principals willing to participate in the experiment.

The number of selected schools was also functional to enrol enough students to ensure a reasonable level of statistical power. As mentioned above, in a school setting, children are embedded in classrooms and it is not usually possible to select a few of them at random and exclude the others from participating in the program, because that would disrupt the normal functioning of the classroom group. To properly account of clustering and to preserve the statistical power of the analysis, a much larger sample size is required compared to the situation where individuals rather than classrooms are randomized. This circumstance increases the minimum number of ob-

4 Under this category we include students born in another country or those born in Italy with at least one parent born abroad.

servations required to detect a given effect size. In our case, we fixed a minimum effect size of 0.2 standard deviations of the outcome measures, i.e. a small effect[5]. Using PowerUp tool[6], we calculated that 30 classrooms and about 600 pupils were necessary. To concentrate organizational efforts on a smaller number of treated classes, we decided to randomly allocate one class to the treated group and two classes to the control group in each school. However, we could not reach exactly the desired sample size because we were able to enrol only eight schools rather than ten, equally split between lower and upper secondary levels (see Table 1). Nevertheless, the final sample can be considered still fairly satisfactory.

Lower secondary schools
– 5 institutes (including 2 "comprehensives" and 1 "parificata"[7])
– 11 second grade classes
– 3 first grade classes

Upper secondary schools
– 5 institutes (2 lyceums; 2 vocational; 1 technical)
– 14 classes (8 second grade classes; 3 first grade; 3 third grade)

Table 1. The sample in brief

2.2 *The Socio-emotional and citizenship competences questionnaire*

Overall the KIDS4ALLL Project has used many different tools to detect the outcome of the intervention itself. The tools were both qualitative and quantitative, to gather as complete a scenario as possible. In fact, as happens in any humanistic investigation it is difficult to isolate other variables which may influence the intended construct. For this reason, quantitative survey tools have been combined with qualitative tools aiming to investigate the reading and interpretation of teachers and students with respect to the educational path carried out.

5 According to Cohen (1988), effect size is the difference between means of the treated and control groups, divided by the standard deviation of the data. An effect size of 0.2 is small, while 0.5 is medium, and 0.8 or bigger is large.
6 https://www.causalevaluation.org/power-analysis.html
7 "Comprehensive" schools include primary and lower secondary classes. "Parificata" is a private school that provides officially recognized education.

In the following paragraph we illustrate the Socio-emotional and citizenship competences questionnaire in order to clarify the whole administration process and the main results.[8]

The questionnaire has been administered at the beginning and at the end of the activities with students and it has been developed through the integration of several survey instruments, that will be shortly described in their theoretical underpinnings.

The chosen measurement tool has been developed according to the following criteria:

– simplicity and clarity of items;
– the overall length of the tool;
– the appropriateness of items to the students' age;
– the ease and wide administration of the questionnaire by different professionals;
– the time, i.e. the possibility of administering the test quickly;
– the adequacy of the questionnaire to be run in different locations and contexts;
– the possibility of administering the test both in its online version and paper version.

In the first part of the questionnaire, profiling questions were included in order to get to know the respondent group.

Several evaluation tools were used in the definition of the questionnaire. The following rating scales have been identified and briefly described.

General Self-Efficacy (GSE)

The General Self-Efficacy Scale (GSE; Schwarzer & Jerusalem, 1995) was created to assess the general sense of perceived self-efficacy, with the aim of predicting the ability to cope with everyday difficulties and adaptation after experiencing stressful life events of all kinds. The German version of this scale was originally developed by Jerusalem and Schwarzer in 1981, first as a 20-item version and later as a reduced 10-item version (Jerusalem & Schwarzer, 1985; 1989;1992; De Caroli & Sagone, 2014).

8 Besides the Socio-emotional and citizenship competences questionnaire the main tools of the evaluation toolkit consist of an Ethnographic observation, whose structure and results are treated in the following chapter; a Final self-assessment questionnaire for students; Teachers and educators check-list; Social Network Analysis; Teachers, educators and stakeholders interviews.

Sense of Community in the School (SOC)

The school, where pupils spend most part of their day, can be regarded as a primary physical and social context for young people and refers to the sense of belonging to the school as a community, the perception of an emotional bond with other students and the feeling that personal needs are met through such belonging.

Regarding belonging, McMillan and Chavis (1986) point out that "belonging has boundaries; this means that there are people who belong and people who do not belong. Boundaries provide members with the emotional security needed to expose needs and feelings and to develop intimacy".

The School Sense of Community (SOC) Scale was developed to measure students' sense of school as a community. This is our empirical measure for what we called "perception of inclusion" above. Such measure covers the dimensions of belonging, emotional connection and opportunity.

Many researchers have used the concept of sense of community to describe the psychological aspects of physical and social contexts that satisfy the need for belonging (Fisher et al., 2002). Sense of community was defined by McMillan and Chavis (1986) as "the feeling of belonging by members, the feeling that members are important to each other and to the group, and a sense of sharing".

Multidimensional test of self-esteem (TMA)

The questionnaire investigates the self-esteem level of the observed subjects.

The responses are grouped and coded to provide scores and standard deviations, which describe the level of self-esteem, compared to the peer average, in several areas.

In particular, the theoretical model on which this instrument is based defines six dimensions of self-esteem that identify the six rating scales: interpersonal relationships, environmental control competence, emotionality, school success, family life, and bodily experience.

The questionnaire consists of 150 questions divided into 6 scales, however it is important to notice that individual survey scales can be used.

In coherence with the starting criteria 2 scales (50 items in total) were chosen for the TMA: interpersonal and emotional.

The Multidimensional test of self-esteem marks out those elements:

– Global self-esteem assessment.
– Evaluation of individual areas explored.
– Administration to individuals or groups.
– Measure easily comparable with other tests.
– Interpersonal and intrapersonal interpretation.

TMA scoring is very simple but must follow a specific procedure.

The scale has both positive and negative items, so different scoring procedures are required for the two types of items (Bracken, 2003).

Career Adaptation Scale (CAAS)

The Career Adapt-Abilities Scale – Italian form consists of four scales made by 6-items, which measure concern, control, curiosity and confidence as psychosocial resources for managing occupational transitions, developmental tasks and work trauma. The Italian form of the 24-item CAAS is identical to the international 2.0 form. The estimation of the internal consistency of the four subscales and the total scores is good. Concurrent validity tests were collected on perceived internal and external barriers, breadth of interests and quality of life. Correlations were as expected and showed that adaptability was negatively related to perceived barriers and positively related to breadth of interests and quality of life. As expected, the analysis of variance showed that adolescents with greater adaptability perceived fewer barriers, expressed a wider range of interests and reported a higher quality of life. Throughout the work we used only the confidence subscale; in the text we refer to such dimension by the term *adaptability*.

PISA 2018

PISA 2018 defines and assesses Global Competence in a multidimensional way, recognizing students' socio-emotional skills and attitudes as key indicators of global competence, in addition to their cognitive reasoning on global and intercultural issues. As the test focuses only on the cognitive knowledge and skills students need to address global and intercultural issues, the student questionnaire collects information on students' skills (both cognitive and socio-emotional) and their attitudes towards global and intercultural issues. The PISA 2018 assessment uses the following definition of global competence: "Global competence is the capacity to examine local, global and intercultural issues, to understand and appreciate the perspectives and world views of others, to engage in open, appropriate and effective

interactions with people from different cultures, and to act for collective well-being". (PISA 2018 Global Competence Framework, 2018).

This definition outlines four target dimensions of global competence that people need to apply successfully in their everyday life: 1. the capacity to examine issues and situations of local, global and cultural significance; 2. the capacity to understand and appreciate different perspectives and world views; 3. the ability to establish positive interactions with people of different national, ethnic, religious, social or cultural backgrounds or gender; and 4. the capacity and disposition to take constructive action toward sustainable development and collective well-being In the dimensions 2, 3 and 4 a number of functional items were chosen for the survey of interest in the project, in order to specifically detect and measure the level of active citizenship competence.

2.3 *Survey tools and teachers' role in the intervention*

The data collection for the impact evaluation was based on a questionnaire administered before and after that the intervention was carried out. The questionnaire was built in the form of an online survey hosted by the LimeSurvey platform. The students participated either by using the digital equipment provided by the school or their personal devices (notebook, tablet or smartphones). Supplementary devices were made available, if necessary, in order to avoid digital divide as well as any technical problem during the survey administration.

The questionnaire was composed by 3 main sections.

- The first one aimed at collecting socio-demographic data on respondents and their families: migratory background, household composition, economic and social status. The purpose of this part was to collect information on the students' learning environment when they are not at school, as well as on the cultural resources they can rely on.
- The second part constituted the core section of the survey: several item sets focused on socio-emotional skills have been administered to the interviewees. The selected tools are mainly of psycho-pedagogical inspiration and already validated by the literature, oriented to measure the level of some key dimensions: general self-efficacy, multidimensional self-esteem (two sub-scales), sense of community in school, and adapt-ability (see previous section).

– The third and last section was devoted to measure students' self-evaluation of one of the lifelong learning key competences: active citizenship.

The questionnaires were slightly different between the two rounds: the socio-demographic section was included only in the first round, while in the second one a specific part was devoted to collect students' feedback on the performed activity, i.e. their experience with the KIDS4ALLL online platform and the buddy learning method (see Table 2).

In each round, we asked teachers to highlight relevant issues about the questionnaire administration. Moreover, overall feedback was collected from the researchers who were attending each questionnaire administration.

In order to maintain maximum comparability between different schools and classes, during the experimental phase teachers were asked to meet two basic requirements: working with students on two competences, active citizenship and digital competence, for at least 15 hours for each competence. The decision to expand the number of treated competences, as well as the hours devoted to the work, was left to the discretion of the teachers on a voluntary basis. Teachers could extend the suggested standard intervention, but only after its completion within the classes. In addition, teachers had the liberty to adapt the schedule of the activities to align with their school calendar. Indeed, a degree of flexibility was necessary to tailor the activities to the wide heterogeneity of settings (two school levels; students spanning different ages engaged in distinct curricular paths, etc.). These elements led to a margin of adaptation of the work carried out, within a common framework.

This short summary about the teachers' role shows that an intervention like the one experimented in the KIDS4ALLL project heavily relies on teachers' commitment, as they are in charge of delivering the proposed educational program. This poses the issue the intervention's *uniformity* across classes and schools. Therefore, in addition to the general problems of impact evaluation in school contexts (see section 2), the KIDS4ALLL intervention had to face this specific challenge threatening its overall effectiveness.

Timing of the evaluation activities
– Beginning of November 2022:
  • Start of the activities
  • First round of data collection (pre).

– April-May 2023:
  • End of the first pilot phase
  • Second round of data collection (post)
  • Final self-assessment questionnaire.

– Both rounds: administration notes.

Table 2. Activities' timetable

3. Evaluation findings

In an evaluation setting with outcome measurements taken both pre- and post-intervention, data lend themselves to a double reading. On the one hand, it is possible to look at what happened in the treated and control groups separately. This reading allows to gauge variations in outcomes across time, by examining differences between the two measurements on the same group. On the other hand, we can compare such differences between treated and control group. In this way, by computing the difference-in-differences, we assess the effect of the intervention net of confounding factors that may have affected both groups during the observation period. In addition, the random composition of the treated and control groups ensures that differences in outcomes cannot be attributed to prior differences between groups. We apply this double reading to the data presented in Table 3.

	Sense of community (scale 1-5)	Active citizenship (scale 1-5)	Emotional self-esteem (scale 25-100)	Interpersonal self-esteem (scale 25-100)	General self-efficacy (scale 1-4)	Adapt-ability (scale 1-5)
Treated group: difference post-pre	-0.14*	-0.06 ns	-1.00 ns	1.05 ns	-0.05 ns	-0.06 ns
Control group: difference post-pre	-0.19 ***	0.04 ns	-2.13 ***	-1.30*	0.01 ns	0.02 ns
Treated-control group difference	0.05 ns	-0.10 ns	1.13 ns	2.35 **	-0.06 ns	-0.08 ns

*: p-value<0.05. **: p-value<0.01. ***: p-value<0.001. ns: not significant

Table 3. Pre-post mean differences between outcome measures among control and treated groups

Looking at the treated group, we see that there was a significant drop in the level of sense of community from pre- to post-intervention, while for the other outcomes the variations were not statistically significant. Thus, if we judged simply from these observations, we should (wrongly) conclude that the intervention negatively affected sense of community while leaving the other socio-emotional skills unchanged. But if we also take the variations in the control group into account, we come to a different conclusion. The sense of community indeed decreased also in the control group, about to the same extent. Thus, the mean difference between treated and control subjects becomes almost null and not significant. Conversely, interpersonal self-esteem significantly declined in the control group while it increased in the treatment group (albeit not significantly).

The result is that interpersonal self-esteem is the only outcome that was positively and significantly impacted by the treatment. The magnitude of the effect is 2.35 points, which amounts to 0.23 standard deviations (SD) of the outcome, i.e. a small but non-negligible effect size. The other outcomes were not significantly affected.

It is worth emphasizing that if we limited ourselves to observe change in only one (treated) group, any variation would have been assumed as a direct result of the experimental activities, while we know that many other confounding factors might play a role. To stress the advantage of the counterfactual evaluation, it is important to consider that if we had not employed a control group, a simple pre-post comparison would have been

highly misleading. Conversely, with two randomly sampled groups, we had the possibility to control for confounding factors unrelated to the intervention itself.

However, regarding the positive effect of the intervention on interpersonal self-esteem, we must acknowledge that we also found a *pre-treatment* statistical difference between control and treated groups. This imbalance is not due to a failed randomization, but it stemmed from sample attrition between the pre- and post-intervention measurements. Thus, restricting the analysis to usable cases (i.e., those who filled in *both* questionnaires), the two groups are not strictly comparable any longer. To account for this imbalance, it is necessary to turn to regression analysis where treated and control groups can be compared net of their pre-intervention outcome levels.

For each outcome, we estimated two regression models. The first model includes the dummy variable indicating treatment assignment and the pre-intervention (t0) level of the dependent variable. This allows to control for possible imbalances in pre-intervention levels across treated and control groups and, at the same time, allows to estimate the treatment effect more precisely (i.e. with a lower standard error), as pre- and post-intervention levels are highly correlated. In the second model, we added socio-demographic variables (gender, age, migratory background) as additional (pre-treatment) control variables.

Regression results (Table 4) confirm that the intervention has affected positively and significantly (at 10% level) interpersonal self-esteem, proving the positive response of this socio-emotional dimension to the experimental activity. In contrast, it is noteworthy that general self-efficacy and perceived adapt-ability were negatively and significantly (at 5% and 10% level respectively) impacted by the intervention. Although this finding may seem counterintuitive and even opposite to the goals of intervention, we believe there can be a more positive interpretation. Indeed, the negative effect could suggest that the mutual confrontation and the work in pairs required by the buddy method led students to put themselves into question. Even if this may have triggered feelings of inadequacy, leading to decreased levels of self-efficacy and perceived adapt-ability, nonetheless it witnesses the existence of an inner process arisen from the intervention, which would need a longer time to be properly assimilated by the pupils. Anyway, such negative impacts are very small (about one tenth of SD) and thus practically negligible. In this regard it is important to note that, in line with international guidelines, the processes of acquisition and development of appropriate

skills and competences could be favoured by an holistic and co-ordinated approach, also equipping education institutions and their staff with the adequate knowledge and tools, to facilitate young people's learning and socio-emotional growth (OECD, 2018; OECD, 2022).

In Models 2, reported in Table 4, we add socio-demographic variables to the regressions: the estimates of the treatment effects did not change. Gender turns to be associated with emotional and interpersonal self-esteem, as well as with general self-efficacy and adapt-ability, with females showing lower levels on all these outcomes. Age is positively correlated with active citizenship competence, indicating that older students are more responsive in this domain. Conversely, younger students perceive more sense of community compared to older ones. Students with migratory background do not significantly differ from the others on any outcome. The empirical finding is inherently significant as it highlights how the sense of community, as perceived by students within their classrooms, transcends cultural differences and family's geographical origins. These findings serve as concrete evidence that, in students' perceptions, shared growth trajectories can pave the way for broader paths of inclusion.

	Sense of community							*Active citizenship*						
	Model 1			Model 2				Model 1			Model 2			
	B	SE		B	SE			B	SE		B	SE		
Treated	0.02	0.10		0.02	0.10		Treated	-0.07	0.05		-0.06	0.05		
Sense of C._t0	0.53	0.07	***	0.52	0.08	***	Citizen_t0	0.51	0.05	***	0.51	0.05	***	
Female				0.00	0.05		Female				-0.03	0.08		
Age				-0.03	0.01	*	Age				0.07	0.02	*	
Migratory background				-0.06	0.05		Migratory background				-0.10	0.05		
Constant	1.76	0.33	**	2.33	0.43	**	Constant	1.49	0.13	***	0.52	0.35		
Rsq	0.26			0.27			Rsq	0.24			0.27			
N	372			368			N	356			353			

	Emotional self-esteem							*Interpersonal self-esteem*						
	Model 1			Model 2				Model 1			Model 2			
	B	SE		B	SE			B	SE		B	SE		
Treated	0.90	0.88		0.83	1.06		Treated	1.49	0.73	+	1.41	0.69	+	
Emotion_t0	0.82	0.04	***	0.78	0.04	***	Interpersonal_t0	0.73	0.05	***	0.72	0.04	***	
Female				-3.17	1.01	*	Female				-2.78	0.48	***	
Age				0.21	0.26		Age				0.14	0.33		
Migratory background				0.31	1.08		Migratory background				0.49	0.82		
Constant	10.12	2.68	**	11.41	7.07		Constant	19.36	3.60	**	18.91	5.26	**	
Rsq	0.65			0.66			Rsq	0.52			0.54			
N	312			309			N	308			307			

	General self-efficacy							*Adapt-ability*						
	Model 1			Model 2				Model 1			Model 2			
	B	SE		B	SE			B	SE		B	SE		
Treated	-0.07	0.03	*	-0.07	0.03	*	Treated	-0.09	0.04	+	-0.09	0.04	+	
Self-efficacy_t0	0.61	0.07	***	0.58	0.07	***	Adapt-ability_t0	0.63	0.04	***	0.61	0.04	***	
Female				-0.11	0.05	*	Female				-0.10	0.04	*	
Age				0.00	0.01		Age				-0.01	0.02		
Migratory background				-0.07	0.07		Migratory background				-0.05	0.07		
Constant	1.14	0.19	***	1.23	0.20	***	Constant	1.31	0.15	***	1.55	0.31	**	
Rsq	0.36			0.37			Rsq	0.38			0.39			
N	369			365			N	368			365			

+: p-value<0.10. *: p-value<0.05. **: p-value<0.01. ***: p-value<0.001.

+: p-value<0.10. *: p-value<0.05. **: p-value<0.01. ***: p-value<0.001.

Table 4. Regression results

4. Why counterfactual evaluation did not detect positive effects?

Counterfactual evaluation is primarily conceived of to find out whether an intervention had an effect, not to discover why. For the latter goal, other means are necessary, such as interviews with participants and stakeholders and ethnographical observations in the sites where the intervention takes place. Nonetheless, it is possible to advance a few educated guesses, also based on the experience gained during fieldwork, about the reasons why we were not able to detect statistically significant and meaningful impacts of our intervention.

In the first place, we cannot exclude that the low sample size played a role. The theoretical sample size established ex-ante, based on statistical power calculations, should allow to detect an effect of at least 0.2 standard deviations, while our actual effect sizes were all below 0.11, except for interpersonal self-esteem (0.23), the only one found to be significant. However, we were not able to reach the theoretical sample size of about 600 pupils due to an unexpected low schools' availability which was a consequence of the pandemic. It should be borne in mind that the intervention was implemented during the post-covid period when all schools were overburdened by numerous projects and initiatives. This circumstance induced a few schools to withdraw their previous availability; in other schools the teaching staff in charge of attending at the practical implementation of our project was reduced.

In the second place, one kind of reasons for ineffectiveness concerns the specificities of the proposed learning method. The buddy method is thought to foster sense of community, socio-emotional and citizenship skills because of its collaborative nature, regardless of the specific learning content to which it is applied. In other words, it should influence expected outcomes by strengthening the social relationships it creates and by developing the social-emotional skills needed to maintain those relationships. Given this hypothetical mechanism, it is quite clear that the effectiveness depends on the availability of sufficient time for the social relationships established by the buddy method to develop and thus produce the expected benefits. As a result, it is possible that our trial lasted too short to produce tangible effects. Moreover, the buddy method was applied in the teaching of citizenship and digital competences because these competences were sufficiently transversal to all disciplines and could be taught by all teachers, regardless their disciplinary background. This was functional to maximize the chance to recruit teachers willing to participate and to experiment with

the buddy method with two different teaching contents. However, if the content and the time devoted to teaching with the buddy method had been entirely focused on citizenship skills, perhaps the measurable impact of the intervention would have been bigger, at least on the specific outcome relating to citizenship skills. In other words, applying the buddy method on two different topics may have softened an already soft intervention. Finally, the buddy method is a form of collaborative learning and collaborative learning was not completely new to the students at our target schools. Teachers told us, before the intervention took place, that at times they use group work strategies. Given that we could not prevent control group students from using collaborative learning strategies, it is possible that another reason for non-effectiveness was the fact that the control group was not really "untreated" in that sense.

In the third place, another kind of reasons for ineffectiveness regards the ways the intervention was implemented. The effectiveness of an intervention based on a teaching method implies a certain degree of commitment both on the part of the recipients (the students) and of those who implement the method (the teachers). Regarding the former, all we could measure was the so-called "intention to treat", that is the impact of the intervention on all subjects, also those who passively attended the classes, since we had no tool to measure students' degree of commitment in the proposed activities. Even if we had it, eliminating uncooperative students from the analyses would have biased the measurement of the impact. Regarding the teachers, notwithstanding the guidelines we provided, we could not guarantee that the way they implemented the buddy method was the same in each treated class. This may have generated heterogeneity in the use of the buddy method which decrease its effectiveness. On the other hand, a uniform implementation of the educational intervention would have required a top-down approach, i.e. an intervention carried out by staff external to the educational institution, but this did not correspond to the spirit of the project. The intention was in fact to leverage teachers' professionalism and to stimulate them to experiment actively and creatively with the material we had made available to them. As a counterpart, it was inevitable to allow a certain degree of heterogeneity in the application of the intervention between schools.

5. Advice for future applications of the KIDS4ALLL methodology

On the basis of the collected data and the pedagogical literature on the subject, we shall try to identify some suggestions for future applications. First, the KIDS4ALLL project as underlined in the previous paragraph was carried out mainly within the citizenship lessons. Over the last few years there has been a renewed interest in citizenship issues and in particular its relation to young student (Lawy & Biesta, 2006). This has been allied to an educational discourse where the emphasis has been upon questions concerning education to citizenship as a school subject rather a cross-curricular disciplinary perspective (Brett, 2022). If education to citizenship were the object of attention of all the disciplines perhaps it would be possible to strengthen the proposed contents and to maximize the relapse. Moreover, to build educational proposals able to better integrate themselves within the different educational contexts, one possible option is to have co-planning sessions of the didactic contents. Co-design in education is 'a highly-facilitated, team-based process in which teachers, researchers, and developers work together in defined roles to design an educational innovation, realise the design in one or more prototypes, and evaluate each prototype's significance for addressing a concrete educational need' (Roschelle et al., 2006, p. 606).

For what concern the "buddy method" an interdisciplinary employment could be functional. The possibility of using the buddy method in different disciplines, and not only in the KIDS4ALLL project, would perhaps be functional to strengthen the bound between students and be more effective. Students would perhaps perceive the methodological proposal not as impromptu. In addition, they could have experimented the buddy method even within paths with traditional methodologies. A further aspect that deserves to be considered regarding the impact of the project is the technological equipment of the schools. Technology can assist learning institutions in facilitating both, personalisation and institutional flexibility (Redecker et al., 2010).

In the future, the process of school involvement might include a phase of technological equipment test. In fact, beyond knowing the number of devices it becomes important to understand the real operational level and functionality. Often schools' technological equipment is slow and obsolete. For this reason, it would be appropriate to consider also the presence of IT technicians able to update technological systems. In the case of KIDS4ALLL, no initial monitoring of the technological equipment in the schools was carried out and it was therefore not possible to understand its role in the success of the project.

Finally, an age-old training question: How long does it really take to learn a new skill? Some experts have attempted to answer the question. The scientific literature on the topic does not have a univocal answer. It is a common opinion that any repeated educational practice for longer can be better acquired both in terms of knowledge and skills.

Although we were aware of the need for longer periods for acquiring such complex skills, we nevertheless considered it important to test the method through both quantitative and qualitative tools (described in the following chapter).

Since there is no consensus on the timing of the development of skills, we believe that any opportunity to increase knowledge on the subject should be exploited.

References

Ballarino G., Filippin A., Abbiati G., Argentin G., Barone C., & Schizzerotto A. (2022). The effects of an information campaign beyond university enrolment: A large-scale field experiment on the choices of high school students. *Economics of Education Review*, *91*, 102308. https://doi.org/10.1016/j.econedurev.2022.-102308

Barone C., Fougère D., & Pin C. (2021). Social Origins, Shared Book Reading, and Language Skills in Early Childhood: Evidence from an Information Experiment. *European Sociological Review*, 37, 1, 18-31. https://doi.org/10.10-93/esr/jcaa036

Bracken A. B. (2003). *Test TMA - Valutazione multidimensionale dell'autostima*. Erikson.

Brett P. (2022). Twenty Reasons Why Cross-Curricular Citizenship Education Might Struggle to Take Flight in Secondary Schools: An Autoethnographic Review. *Curriculum and Teaching*, *37*(1), 5-29. https://doi.org/10.7459/-ct/37.1.02

Cohen J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. New York, NY: Routledge Academic

De Caroli M.E., & Sagone E. (2014). Generalized Self-efficacy and Well-being in Adolescents with High vs. *Low Scholastic Self-efficacy, Procedia - Social and Behavioral Sciences*, 141, 867-874, ISSN 1877-0428, https://doi.org/10.10-16/j.sbspro.2014.05.152

Ford A. (2015). Oxford University Peer Support Programme: Addressing the wellbeing of students. In M. Henning, C. Krageloh, & G. Wong-Toi (Eds.), *Student motivation and quality of life in higher education* (pp. 167-174). Routledge.

Frischmann R.M. (2013). *Skills-based approach to developing a career.* Frischmann, Trafford Publishing.

Jerusalem M., & Schwarzer R. (1992). Self-efficacy as a resource factor in stress appraisal processes. In R. Schwarzer (Ed.), *Self-efficacy: Thought control of action* (pp. 195-213). Hemisphere Publishing Corp

Jerusalem M., & Schwarzer R. (1989). Anxiety and self-concept as antecedents of stress and coping: A longitudinal study with German and Turkish adolescents. *Personality and Individual Differences*, *10*(7), 785-792.

Khandker S.R., Koolwal G.B., & Samad H.A. (2010). *Handbook on impact evaluation: quantitative methods and practices*. Washington D.C., The International Bank for Reconstruction and Development / The World Bank.

Lawy R., & Biesta G. (2006). Citizenship-as-Practice: The Educational Implications of an Inclusive and Relational Understanding of Citizenship. *British Journal of Educational Studies*, *54*(1), 34-50. http://www.jstor.org/stable/3699294

McMillan D. W., & Chavis D. M. (1986). Sense of community: A definition and theory. *Journal of Community Psychology, 14*(1), 6–23. https://doi.org/10.1002/1520-6629(198601)14:1<6::AID-JCOP2290140103>3.0.CO;2-I

OECD (2022). *Recommendation of the council on creating better opportunities for young people*. https://www.oecd.org/mcm/Recommendation-on-Creating-Better-Opportunities-for-Young-People.pdf

OECD (2018). *Education 2030: The Future of Education and Skills.* Position paper. https://www.oecd.org/education/2030-project/contact/E2030%20Position%20Paper%20(05.04.2018).pdf

Redecker C., Leis M.J., Leendertse M., Gijsbers G.W., Punie Y., Kirschner P.A., Stoyanov S., & Hoogveld B. (2010). *The future of learning: New ways to learn new skills for future jobs - results from an online expert consultation*.

Roschelle J., Penuel W. R., & Shechtman N. (2006). Co-design of Innovations with Teachers: Definition and Dynamics. In *Proceedings of the 7th International Conference on Learning Sciences* (pp. 606-612). International Society of the Learning Sciences.

Rubin D.B. (1980). Comment on: "Randomization analysis of experimental data in the fisher randomization test" by D. Basu. *Journal of the American Statistical Association*, 75, 591–593.

Schwarzer R. & Jerusalem M. (1995). Generalised self-efficacy scale. In J. Weinman, S. Wright, & M. Johnston, *Measures in health psychology: A user portfolio. Causal and control beliefs* (pp. 35-37). NFER-NELSON.

Wardak D., Wilson S. & Zeivots S. (2023). Co-design as a Networked Approach to Designing Educational Futures. *Postdigit Sci Educ*. https://doi.org/10.1007/s42438-023-00425-5.