sciendo
Swiss Journal of Sociology, 48 (1), 2022, 21–46

Experimental Research in Education: An Appraisal of the Italian Experience

Giovanni Abbiati*, Gianluca Argentin**, Davide Azzolini***, Gabriele Ballarino****, and Loris Vergolini*****

Abstract: This work provides an assessment of the Italian experimental literature in the field of education. We review 25 RCTs completed between 2009 and 2020, analysing their alignment to the CONSORT guidelines. Our findings show that the scientific reporting is on average of good quality; however, there are areas where a significant improvement is needed. We suggest viable solutions, aimed at improving the robustness of experimental research in sociology, taking advantage from consolidated rigorous praxis developed in other disciplines.

Keywords: Italy, education, randomized controlled trials, CONSORT, review

Recherche expérimentale en éducation : une évaluation de l'expérience italienne

Résumé : Cet article dresse une revue de la littérature expérimentale italienne dans le champ de l'éducation avec un état des lieux de 25 essais contrôlés randomisés (ECR) menés de 2009 à 2020 en analysant leur conformité aux lignes directrices de CONSORT. La littérature est, en moyenne, de bonne qualité, cependant, une amélioration significative est nécessaire dans certains domaines. Nous proposons des solutions viables pour améliorer la robustesse de la recherche sociologique expérimentale en tirant parti des pratiques rigoureuses d'autres disciplines.

Mots-clés : Italie, éducation, essais contrôlés randomisés, CONSORT, revue

Experimentelle Forschung in Bildungssoziologie: Eine Bewertung der Italienischen Erfahrung

Zusammenfassung: Dieser Beitrag liefert eine Bewertung über experimentelle Studien in der italienischen Bildungssoziologie zwischen 2009 und 2020. Wir betrachten 25 randomisierte kontrollierte Studien und analysieren ihre Ausrichtung an den CONSORT-Richtlinien. Unsere Ergebnisse zeigen, dass die Berichterstattung im Durchschnitt hohen Standards entspricht. Es gibt aber Bereiche, in denen eine Verbesserung erforderlich ist. Wir schlagen praktikable Lösungen vor um die Robustheit von experimenteller Forschung in der Bildungssoziologie zu verbessern.

Schlüsselwörter: Italien, Bildung, Experimente, CONSORT, Bewertung

* University of Brescia, I-25122, Italy, giovanni.abbiati@unibs.it
** University of Milan-Bicocca, I-20126 Milan, gianluca.argentin@unimib.it
*** FBK-IRVAPP, I-38122 Trento, azzolini@fbk.eu
**** University of Milan, I-20122 Milan, gabriele.ballarino@unimi.it
***** University of Bologna and FBK-IRVAPP, I-40126 Bologna and I-38122 Trento, vergolini@irvapp.it

Seismo
DOI 10.2478/sjs-2022-0004

1 Introduction

In the framework of the recently widespread interest in field experiments in the social sciences, in this paper, we provide an assessment of the wave of experimental work that has marked the field of educational research in Italy over the last decade. After the early pioneering studies conducted at the end of the twentieth century, an increasing number of scholars adopted this approach, and the resulting literature constantly grew.

This rapid diffusion, however, has proceeded in a disciplinary field in which experimental designs represent a novelty, and the awareness of their key features and constraints is often not adequate. Random assignment alone is not in fact a sufficient condition for robust causal inference. As Berk (2005) effectively states it, causal inference retrieved from experimental designs represents the "bronze standard" of empirical research, and its potential can be undermined in cases in which the complexity that this method entails becomes disregarded. Researchers in this field, moreover, should also bear in mind that the acknowledgement of the methodological limitations inherent to randomized controlled trials (RCTs) and to counterfactual methods in general has relevant implications for the scope of their application in sociology (Gangl 2010).

Drawing on this consideration, this review is aimed at evaluating the recent Italian experimental literature in education under both methodological and substantial perspectives. Regarding the former point, we evaluated the accuracy of reporting of the Italian RCTs against the CONSORT (Consolidated Standards of Reporting Trials) grid (Schulz et al. 2010; Moher et al. 2012). Building on the results of this analysis, we identify relevant implications for the design and implementation of RCTs in sociology and in the social sciences in general.

We argue that Italy represents an interesting case study for performing such an exercise. To our knowledge, Italy is the only European country – except for the UK – where more experimental research in education has been conducted, and sociologists have made key contribution to it, marking a sharp difference from the international context, in which experimental research has seen limited involvement from sociologists compared with other social science disciplines (Jackson and Cox 2013; Baldassarri and Abascal 2017). As detailed below, field experiments in education conducted in Italy over the last decade have covered a wide range of topics and involved all of the core actors of the school systems, such as stakeholders financing and promoting research (central and local governments, foundations and charities) or experimental subjects (teachers, pupils and families).

This growth was all the more unexpected since it occurred in a context in which demand for scientific, policy-relevant evidence from the political system and decision-makers has been scarce (De Blasio et al. 2021). There is, however, an interesting underside to the Italian situation. Indeed, during the 1990s and the

2000s, a number of policy initiatives in the field of evaluation were undertaken, which proved to be robust to partisan, ideology-driven policies. In particular in the early 2000s international pressure favoured the introduction of two agencies for the evaluation of schools and universities: INVALSI (National Institute for the Evaluation of the School and the Training Systems) and ANVUR (Italian National Agency for the Evaluation of the University and Research Systems). The bipartisan support that marked their creation might explain their retention in the succession of administrations of different colours. Despite neither being committed to experimental research or counterfactual evaluation, both agencies have been performing important work in promoting systematic data collection on schools, teachers, and students and in introducing a modern, evidence-based school culture to Italian public opinion (Ballarino 2015). In particular, INVALSI has been placed in charge by the Ministry of Education to lead regular assessments of pupils' competencies via standardized tests, the results of which are now among the key sources for the study of educational achievement and its differentiation and also serve as an information infrastructure for researchers developing their experiments.[1]

In the next section, we present the methodology of the review, while Section 3 provides a description of the main features of the studies. Section 4 systematically analyses the quality of reporting of these studies with reference to the established CONSORT grid. Section 5 elaborates on these results, underscoring the critical issues emerging from this analysis, which are relevant to the future development of experimental research in the social sciences in general. Section 6 concludes the review with recommendations for future experimental research.

2 The Review: Methodology

This paper reviews experimental research in the field of education conducted in Italy in the 2009–2020 period. Restricting the analysis to a specific subfield allows us to compare a relatively uniform body of studies in terms of social and institutional counterparts, on the one hand, and of challenges and constraints, on the other hand. To further foster homogeneity, we implemented three inclusion criteria. First, we included field experiments only, namely studies based on randomization (the key element) evaluating interventions introduced in actual school contexts, thus excluding lab and survey experiments. Second, we limited our sample to K-13 education. Due to their specificity, we excluded experiments conducted among university students, with many of those at the boundary between field and lab experiments. Third, we

1 The availability of funding for the experiments reviewed in this paper was related to this cultural process, which was also stimulated by a number of private, not-for-profit foundations and charities and – importantly – by the individual initiative of a small number of scholars, who were influential in proposing the counterfactual approach as a key tool for policy research (Martini 2008; Martini and Sisti 2009).

defined a time span: the first experiments that we surveyed dated back to 2009, and we decided to limit our sample to those experiments with experimental results circulating as reports, articles, working papers, conference papers, books, chapters, or talks by December 2020. We excluded studies still ongoing as of that date, i. e. studies for which not even early experimental evidence was available.

The studies to be included in the review were collected using a four-step procedure. The starting point was to draft a list of the experiments conducted by the authors of this review – a list that encompassed a notable proportion of the studies included in the final sample (16 studies). We enriched this list by adding all other experiments fitting our inclusion criteria of which we were aware (8 studies), thanks also to the collaboration of various colleagues. Then, we surveyed the relevant literature indexed by Scopus using the keywords "Italy", "RCT", "experiment", "randomized controlled trial" and "education". No further studies were added at this stage. Finally, we took advantage of a recent initiative developed by a multidisciplinary group of Italian social scientists aimed at screening experimental research in Italy,[2] checking once again the completeness of our list by comparing it to theirs. We finally had the 26 studies listed in Table 1.[3]

We then evaluated the accuracy of the reports – considering articles, chapters, reports, presentations, etc. – by documenting their alignment with the CONSORT checklist (Schulz et al. 2010). CONSORT provides a standard, comprehensive, and authoritative guide for trial reporting. This list, discussed and developed within the biomedical field, was first published in 1994 (Andrew et al. 1994) and was intended as a practical tool to help authors to prepare complete and transparent reports and to allow readers to evaluate them. In this paper, we use the updated 2010 version (Schulz et al. 2010).

As anticipated above, robust and reproducible experimental evidence does not merely require the observation units to be randomized. There is a vast array of issues that should be in place beforehand and that can arise during the running of an experiment that must be properly addressed and documented. Moreover, the scope of applicability of a set of findings should be clearly defined and discussed to inform both social theory and policy-making. Given the influence that might be exerted by structural, cultural, and contextual factors on the interventions evaluated, this

2 The first Italian blog entirely focusing on randomized controlled trials, named *"Studi Randomizzati" (Randomized Studies),* https://studirandomizzati.wordpress.com.

3 Different from systematic reviews, we do not provide the standard PRISMA flow diagram for literature searches. The reason for this choice lies in the peculiar branch of literature under investigation: many studies are not indexed, partly because their indexing has not been properly conducted (e. g., book chapters) and partly because of the formats of the documents (e. g., technical reports, working papers in non-indexed WP series, presentations). Hence, a standard literature search via Sociological Abstracts or Scopus would have resulted in a much smaller number of studies. To be precise, our review, despite covering the entire literature in the specified field and despite using a standardized analytical framework, lacks the elements allowing it to be defined as "systematic" (first of all, a pre-specified protocol).

Table 1 List of the Studies included in the Review

Number	Name of the program	School year (or years) of the intervention	Reference
1	PON M@t.abel+ (wave 1)	from 2009/10 to 2011/12	Argentin et al. (2014)
2	Comunicazione sul rischio di radiazioni ionizzanti (Comunicating the Risk of Ionizing Radiations)	2009/10	Fasanella and Maggi (2011)
3	SAM - Scacchi e Apprendimento in Matematica (Chess and learning in Math)	2010/11	Argentin et al. (2012)
4	PON M@t.abel+ (wave 2)	from 2010/11 to 2012/13	Abbiati et al. (2021)
5	MOS-4	2012	De Poli et al. (2018)
6	Family Background, Beliefs about Education and Participation in Higher Education	2013/14	Abbiati et al. (2018)
7	Riunioni di Famiglia (Family Meetings)	2013/14	Argentin et al. (2015)
8	EOP (Equality of Opportunity for Immigrant Students)	2013/14	Carlana et al. (2018)
9	Comunicare il rischio chimico (Comunicating Chemical Risk)	2014/15	De Cataldo et al. (2016)
10	Dispersione scolastica, equità sociale, orientamento (School Dropout, Social Equality, Orientation)	2014/15	Barone et al. (2017)
11	Affording College with the Help of Asset Building (ACHAB)	from 2014/15 to 2016/17	Martini et al. (2021)
12	SCUOLINSIEME (School-Together)	from 2014/15 to 2016/17	Abbiati et al. (2019)
13	Relazioni a scuola (relationships at school) – National Efficacy Trial	2016/17	Argentin et al. (2020)
14	Relazioni a scuola (relationships at school) – National Effectiveness Trial	2016/17	Argentin et al. (2020)
15	Non solo a scuola (Not Just in School)	2016/17	Argentin et al. (2018)
16	Twitteratura	2016/17	Barbetta et al. (2019)
17	Il bias implicito degli insegnanti (Teachers' Implicit Bias)	2016/17	Alesina et al. (2018)
18	MENTEP	2016/17	Abbiati et al. (2020)
19	La torta dell'economia (Economy's Pie)	2016/17	Rinaldi and Argentin (2020)
20	Family star	from 2016/17 to 2017/18	Argentin et al. (2019)
21	Digital Well-being – Schools	2017/18	Gui et al. (2018)
22	Relazioni a scuola (relationships at school) – Trentino Local Trial	2018/19	Argentin and Gerosa (2020)
23	TEACHUP	2018/19	Azzolini et al. (2020)
24	Mathesis Mathematics Camp	2018/19	Aparicio Fenoll et al. (2020)
25	FA.C.E. Farsi Comunità Educanti (Creating an Educating Community)	2019/20	Del Boca et al. (2020)
26	Mathematics Active Learning Teaching Practices	2019/20	Di Tommaso et al. (2021)

element is particularly relevant for sociology. In this particular discipline, as in the social sciences in general, awareness of the relevance of some methodological issues remains limited, partly because the adoption of experimental research protocols has been relatively recent. In other disciplines, such as medicine, in which the adoption of RCTs has been common practice, these concerns have called over time for the institution of standards.

The CONSORT scheme is structured in five main areas, 25 subareas and 37 indicators. We considered the differences between the scientific field in which the checklist originated and our use in the domain of the social sciences by making marginal changes[4] and deleting a few items that we regarded as not pertinent[5]. Our final checklist comprises 26 indicators organized into 20 subareas, as reported in Table 2.

For each study included in the review, we applied the following procedure: i) we emailed the authors asking for the complete set of documents pertaining to the experiment (published or unpublished papers/reports/talks); ii) we coded each study using the adapted version of the CONSORT checklist[6]; iii) we again emailed at least one of the authors, asking them to validate our codification or to bring evidence to correct eventual inaccuracies in the coding. Thanks to this last action, we were able to collect unpublished data about the experimental process and to verify methodological details not fully available to the public.

3 The Studies

Table 1 provides the full list of the 26 studies considered for this paper, while the full information is provided in Appendix Table A1. First, as already stated, all of the studies were conducted between the 2009/2010 school year and the 2019/2020 school year, proving how experimental studies are a relatively new venture for Italian education researchers.

4 We applied slight changes to the items that we used, in particular: the setting description could be limited to the geographical area or to specific and relevant aspects of the population of interest (4b); the treatment could be described emphasizing its general organizational and content features (5); given the frequent lack of outcome data beyond the experimental samples, some flexibility was allowed in the outcome pre-specification, e. g., lack of pre-specification of coding details or data reduction techniques (6a); we did not set the presence of details' block sizes as a condition (8b); and we accepted p values as a measure of statistical uncertainty (17a); for the full original protocol see Moher et al. (2012).

5 In particular, we deleted the following items, of limited/no applicability in our case: 1a, 1b (referring to the presentation of the RCT in the title and the abstract of the paper/report); items connected to changes after the trial commenced or to interim and stopping guidelines (4b, 5b, 7b, 14b); items connected to the procedure of double blinding (9, 10, 11a, 11b); and harms (19).

6 The need for an adaptation of the CONSORT checklist, the results of which have been described before, emerged during the coding process. The authors coded the studies themselves and met frequently to share doubts about the interpretation of single items or coding choices.

Table 2 Adapted CONSORT Checklist used in the Analysis

Area	Sub-area	Item	Label
Intro	Background and objectives	2a	Scientific background and explanation of rationale
		2b	Specific objectives or hypotheses
Methods	Trial design	3a	Description of trial design (such as parallel, factorial) including allocation ratio
	Participants	4a	Eligibility criteria for participants
		4b	Settings and locations where the data were collected
	Interventions	5	The interventions for each group with sufficient details to allow replication, including how and when they were actually administered
	Outcomes	6a	Completely defined pre-specified primary and secondary outcome measures, including how and when they were assessed
	Sample size	7a	How sample size was determined
	Sequence generation	8a	Method used to generate the random allocation sequence
		8b	Type of randomization; details of any restriction (such as blocking and block size)
	Statistical methods	12a	Statistical methods used to compare groups for primary and secondary outcomes
		12b	Methods for additional analyses, such as subgroup analyses and adjusted analyses
Results	Participant flow (a diagram is strongly recommended)	13a	For each group, the numbers of participants who were randomly assigned, received intended treatment, and were analysed for the primary outcome
		13b	For each group, losses and exclusions after randomization, together with reasons
	Recruitment	14a	Dates defining the periods of recruitment and follow-up
	Baseline data	15	A table showing baseline demographic and clinical characteristics for each group
	Numbers analysed	16	For each group, number of participants (denominator) included in each analysis and whether the analysis was by original assigned groups
	Outcomes and estimation	17a	For each primary and secondary outcome, results for each group, and the estimated effect size and its precision (such as 95% confidence interval)
		17b	For binary outcomes, presentation of both absolute and relative effect sizes is recommended
	Ancillary analyses	18	Results of any other analyses performed, including subgroup analyses and adjusted analyses, distinguishing pre-specified from exploratory
Limitations	Limitations	20	Trial limitations, addressing sources of potential bias, imprecision, and, if relevant, multiplicity of analyses
	Generalizability	21	Generalizability (external validity, applicability) of the trial findings
	Interpretation	22	Interpretation consistent with results, balancing benefits and harms, and considering other relevant evidence
Other info	Registration	23	Registration number and name of trial registry
	Protocol	24	Where the full trial protocol can be accessed, if available
	Funding	25	Sources of funding and other support, role of funders

3.1 Social Experimentation and Policy Interventions: RCTs in the Social Sciences

RCTs in the social sciences respond to two distinct knowledge purposes that can be considered the two extremes of a continuum: impact evaluation studies, on the one hand, or theory-driven experiments, on the other hand. Typically, in the first case, sociologists play the role of external evaluators, asked by policy-makers to design an impact evaluation through an appropriate RCT. The evaluators should not participate in the intervention design or in its implementation. At the opposite pole of the continuum, the distinction between the evaluators and the managers of the intervention becomes blurred or nonexistent: interventions are designed, implemented, and assessed by the researchers themselves to test specific theoretical hypotheses.

Among the Italian RCTs reviewed examples of RCTs close to the theory-driven pole are Studies 6 and 13. In the first case, a team of researchers designed an intervention based on the application of rational choice theory to educational choices, with the aim of understanding whether informing students about the costs, risks, and occupational returns of tertiary degrees would lead to a reduction of social inequalities associated with university enrolment. In the second case, the researchers designed an intervention aimed at assessing whether it was possible to increase student achievement by increasing a specific component of teachers' effectiveness (their relational skills) via light-touch professional training. On the opposite side, we find, for example, Studies 1 and 3, in which the evaluators had no voice in determining the content of an already existing treatment or its implementation. The former study evaluated a professional development program for math teachers, while the latter evaluated the impact of a chess course for students. A common situation lying between the poles is given when sociologists and practitioners/policy-makers join forces, typically in the early stages of program development. In this case, exemplified by Study 7, the evaluators contribute to the design of the intervention with the aim of driving practitioners to identify and strengthen the core elements of their policy idea (family group conferences applied to the school setting), the success of which can then be reliably (as much as possible) assessed using a given set of predetermined outcomes.

3.2 Programs' Content and School Settings

Most of the studies refer to interventions in the fields of informational guidance, information campaigns addressed to students or professional development addressed to teachers, thus falling into three consolidated research traditions. Other studies investigated the potential of less standard leverages, especially when the target group was composed of teachers, such as support for online courses via SMSs, self-awareness of teachers' own prejudice, and self-assessment of their teaching practices.

In approximately half of the cases (14 of 26), interventions were targeted to students. In 5 of these cases, they included the provision of information via outreach interventions, particularly concerning future school choices, while in 2 more cases,

information was provided in a personalized manner. In the remaining 7 cases, the interventions included non-curricular activities, such as courses to be held in a typical classroom framework (summer courses, chess courses, math labs, introductory classes on finance), extra-school activities (volunteering, project funding), or a combination of both (the provision of financial aid in the form of an incentivized savings account combined with financial education classes). In 9 cases, the intervention was centred on teachers. In 6 of these cases, it included professional development courses; in two cases, it was based on forms of self-assessment (in one case concerning their prejudice, in another their performance in technologically enhanced teaching). In the remaining 2 cases, the intervention included the provision of information to parents.

Most of the studies were addressed to secondary school students and/or teachers: 13 studies concerned lower secondary schools, while 10 studies concerned upper secondary schools. Two studies referred to primary schools and one to a kindergarten, but the latter did actually concern families. In only one case, all school levels were involved in a local-level study (Study 6) concerning new permanently hired teachers.

Concerning the geographical scope, approximately half of the studies (13) were multisite, that is, implemented in more than one geographical location across the country; 9 were local-level studies, among which those at the regional level were also included; 2 were national-level studies; and 2 were the Italian section of multicountry studies. The scarcity of national-level studies is likely to be related both to the complexity entailed by national-level designs and to a scarce institutionalization of counterfactual evaluation in the central government. This fact implies that support for this type of research is mainly guaranteed by local administrations, EU institutions and programs, or private organizations (such as philanthropic foundations) (see also Section 3.4). In such a context, a multisite study might turn out to be the best compromise between the difficulty in raising the funding required by national-level studies and the issues of generalization plaguing local-level studies in a country such as Italy, where deep geographic socioeconomic cleavages extend to the school system (Bratti et al. 2007; Argentin et al. 2017).

3.3 Research Designs

Most designs took the form of clustered, randomized, controlled trials with blocking, taking advantage (or being constrained to) the nested nature of the school system, which is organized around school institutions. In such cases, randomization might occur either at the school level or, within schools, at the class level (the latter being a typical case of blocked randomization, in which the schools constitute the randomization urns). Consistent with the heterogeneity of geographical scope, the size of the samples investigated also varied widely. If we consider the student population, sample sizes varied from a minimum of 261 individuals to a maximum of more

than 27 000[7]. Both the average and the standard deviation were pushed upwards by an outlier (Study 13, counting 27 000 students) so that the values excluding the outliers, at approximately 2480 for the average and 3000 for the SD, are more informative. Overall, they tell us that the average size of the studies was not negligible, with 15 studies involving more than 1000 students, and that there was substantial heterogeneity. Similar considerations can be made for studies involving teachers.

In 6 cases, the schools to be involved were randomly sampled to be invited and included in the projects to lend external validity to the findings. This practice, unsurprisingly, is a function of the scale of the project: among the 8 studies having a number of students greater than average, 3 performed random sampling of the units to be included in the experiment.[8] The same can be said for the only studies that can be considered genuinely implemented at the national level (Studies 14, 18 and 23), which targeted teachers.

Regarding data collection, student outcome variables were collected by means of standardized assessments (4 cases), ad hoc surveys (9 cases), or both (5 cases). In other cases, administrative records for students were used, in combination with a standardized assessment (1 case) or a survey (5 cases). In 3 of these cases, outcome data collected from students were supplemented with corresponding data collected from teachers, while in Study 6, results were collected only concerning teachers (typically in the form of survey data).

3.4 Funding and Institutional Drivers

We now come to an RCT feature apparently less relevant from a methodological perspective, namely the institutional bases of the experiments and their funding sources. Funding sources were mostly national or EU public agencies, or not-for-profit institutions. In 13 cases, slightly more than half of the total resources came from different national ministries, agencies or departments. In 3 cases, funding was directly provided by EU programs, in 5 cases by the Italian Ministry of Education and Research, in 3 cases by regional governments, and in 3 cases by public research institutes. A significant role was also played by philanthropic foundations of banking

7 We include in this consideration only efficacy trials, which aim at assessing whether an intervention produces the expected impact but are usually realized using small samples and under circumstances facilitating the intervention implementation delivery. Conversely, effectiveness trials assess the intervention's impact under real world circumstances, hence facing challenges due to the intervention's implementation occurring at scale on large samples of beneficiaries and with intervention's staff being less motivated, less strictly trained and supervised, etc. In our case, we do not include Study 14 here, the estimates of which were calculated on 400 000 individuals, the whole Italian grade 8 population, as the intervention group, was randomly selected among the whole country's population. Schools not selected served as a control group.

8 Study 10 had a very small number of experimental units and seems to escape this rule; however, the students included in the experiment were subject to a very selective procedure of inclusion out of a much larger sample for targeting reasons.

origins.[9] In 4 cases, the interventions were fully funded by philanthropic foundations, while in 3 further cases, they joined forces with other charities (in one of the latter cases, EU funding was also available). Two interventions were funded by charities and a further intervention by a charity and a business association of small local banks. Two interventions were funded by universities: in one case fully, by a private university, while in the other case, a public university joined forces with a large corporation. Finally, one intervention was funded by a private association.

Substantial heterogeneity of funding sources is apparent. This heterogeneity depends, again, on the low level of institutionalization of experimental research in Italy, as well as on the way in which most of these projects were created. Indeed, in our correspondence with the authors (see Section 2), we collected information about the latter point and coded it in Table A1. An explicit request from public institutions for an experimental research design intended to gather robust empirical evidence to inform policy-making was a procedure present in a minority of cases, 8 of 26 studies. In only 3 of these cases was the experimental evaluation assigned on the basis of an open public competition, while in the remaining 6 cases, policy-makers asked researchers to provide an experimental evaluation of a program already designed. In the remaining 17 cases, more than ⅔ of our sample, experimental designs were autonomously established by researchers. In 10 of these studies, researchers designed both the program to be tested and its evaluation, while in the remaining 7, researchers asked policy-makers to fund the evaluation of an already existing program. Such a prominent role of the researchers in the creation of the interventions underscores some key points already raised above, particularly weak institutionalization and the role of key individuals in the promotion of experimental research in Italy. In many cases, indeed, researchers managed to gather funders for their own projects by promoting their ideas to key officers in governments, public agencies, bank foundations, and charities. While this phenomenon might be considered a laudable example of professional voluntary effort, it might also have a number of downsides. First, voluntarism alone can hardly address the systemic lack of evidence-based policy-making in our institutional context. Second, the frequent identification of program evaluators with program designers might be problematic in light of the well-known perverse incentives connected to publication market (Ravallion, 2008), which favours academic novelty over null effects and leaves even the most solid research designs vulnerable to cherry-picking activities.

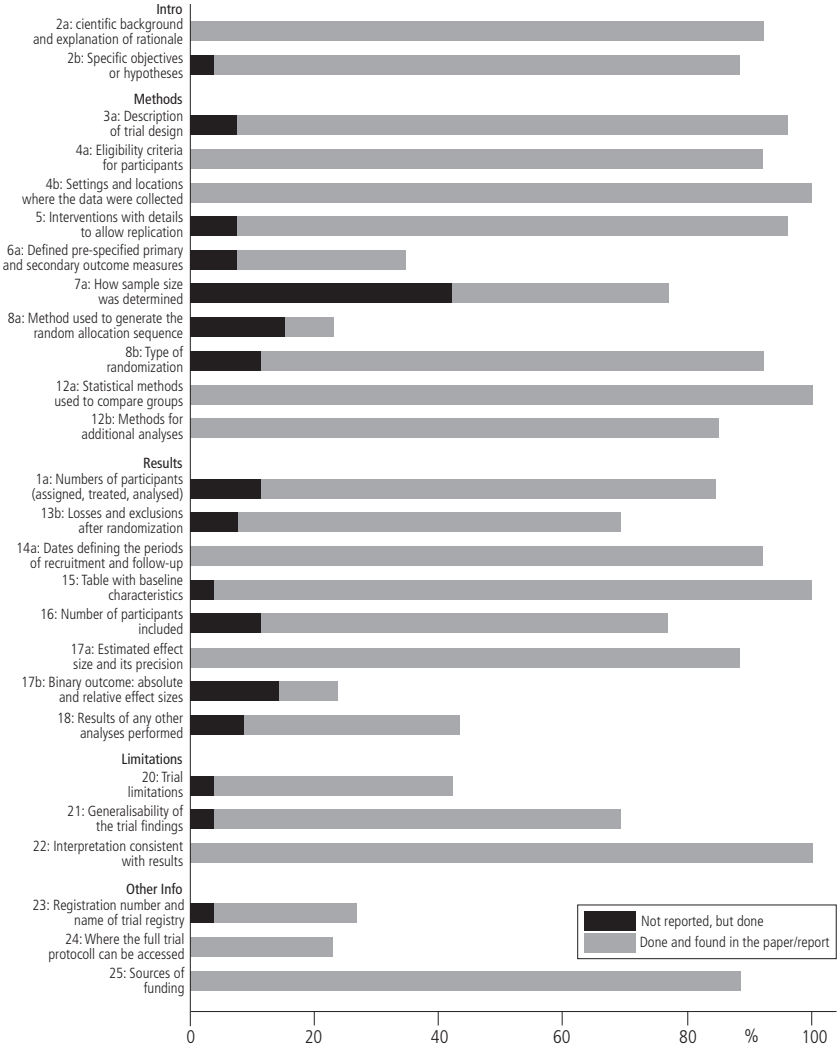
9 These institutions are typically Italian institutions, created during the 1990s when the largest Italian banks, who were national property since the 1930s, were privatized, and a substantial proportion of their assets was transferred to the foundations – nominally private bodies with a not-for-profit charitable mission.

4 Assessing RCTS' Reporting Accuracy

We come now to the reporting accuracy of our studies, assessed against the CONSORT criteria. The results are presented in Figure 1, in which each vertical bar expresses the percentage of studies fully complying with the requirements indicated by the corresponding item. For each item, we distinguish between studies for which all relevant elements related to an item were found in published material (grey area within the bars) and studies for which relevant elements were found in working documents that were not publicly released (black area). The weight of the black area over the grey area can hence be interpreted as an indicator of transparency of reporting. In the case of multirequirement items (for instance, Item 20), we considered a study to be compliant only if all of the points indicated by CONSORT were covered.
Three main elements can be identified at first glance from the graph. The first is that, on average, the scientific reporting of the studies can be considered of good quality, with a significant proportion of items bordering on or exceeding 80 % compliance. The second is the variability among items and areas, immediately appreciable by the presence of peaks of different heights, showing that the aforementioned quality is unevenly distributed across items and areas, and it is possible to identify areas in which improvement is needed. The third refers to the level of completeness of the reports, which is, with some minor exceptions, rather high. This result is visually illustrated by the smaller extension, within each bar, of the blue area compared to the red area, indicating that the vast majority of studies explicitly address the issues under inspection.

Let us now take a closer look at each dimension. Introductory sections are, on average, accurate, with more than 85 % of the studies describing the main methodological features of the study and declaring its objectives. If we exclude three relevant exceptions (discussed further below), methodological section items show high scores, on average. Studies provide a clear description of the trial design and of the randomization strategy adopted (Items 3a and 8b). Regarding the interventions tested, the details contained in the reports are sufficient to allow for replication (at least intended in a broad sense) since eligibility criteria are explicitly stated, locations and settings are declared, and the interventions are almost always carefully described (4a, 4b, and 5). Finally, the effects of the treatments are estimated using properly described and specified models (12a, 12b). Weaker points concern the complete prespecification of primary and secondary outcomes (6a), the determination of the sample size (7a), and the disclosure of random allocation sequences (8a). The first two instances are interconnected, at least in part. Many studies lack prespecification of the outcomes, and this omission happens both for the indicator of interest and its coding. Some studies report effects on multiple outcomes, often without clarifying the hierarchy among them, or they split the sample into multiple, not predefined, subgroups, making it difficult to disentangle the main outcomes and

Figure 1 Accuracy of the Studies Included in the Review



analyses from exploratory ones. This lack of clarity is also reflected by the low level of compliance with Item 18 in the area of "results", in which approximately half of the studies fail to distinguish between main and additional analyses. Moreover, the data treatment of prespecified outcomes (e. g., data reduction technique; choice of specific thresholds in ordinal or continuous variables) is almost never provided in advance. Perhaps more than on a lack of awareness, this absence could depend on

both data scarcity in the Italian context and the lack of internationally validated measures in our discipline. More precisely, there are often no available data on the outcomes of interest outside the experimental samples (with the notable exception of achievement test scores). These weaknesses are particularly relevant, considering the already mentioned frequent identification of the programs' evaluators with the programs' designers.

The determination of the sample size was reported in 77 % of cases, although most of the time, it was not publicly available. In many cases, power calculations were lacking, possibly due to the lack of data on the selected outcomes and because, in many cases, the sample size was determined by the amount of volunteer participation from schools/teachers or by budget constraints. Whatever the reason, we recorded that such elements rarely find their place in the published material. Finally, the random allocation sequence was almost never disclosed, and researchers seem to be unaware of its importance since they kept almost no track of it, even in unpublished material.

The "results" section is also characterized by high variability in item compliance. Items 13a, 13b, and 16 refer to the ways in which treatment compliance, response rates and attrition are addressed. The compliance level for the three items was decent to good (81 %, 65 %, and 77 % overall, respectively). In our view, however, efforts should be made to improve the handling of these issues. Moreover, these pieces of information are sometimes difficult to collect since they are often reported in various reports or papers. Furthermore, participant diagrams or flowcharts were mostly missing, but they would be definitely helpful. A very high level of completeness in reporting characterizes Items 14a (dates defining the periods of recruitment and follow-up), 15 (baseline equivalence), and 17a (effect sizes and their precision). Compliance with Item 17b (presentation of both absolute and relative effects for binary outcomes) was rather low and should definitely be improved, even if the issue is, all things considered, quite marginal.

Items related to study limitations showed lower scores. Item 20 refers to the trial limitations properly defined (e. g., biases, imprecision, multiplicity of analyses). While imprecision and sources of bias were generally transparently stated, the low compliance marking this item was mostly due to the lack of discussion of the multiplicity of analyses. Awareness of this problem is making its way through the Italian experimental community, but it has been vastly ignored until recently. This result does not come as a surprise, given its close link with the frequent lack of prespecification of the outcomes and the often-missing distinction between types of outcomes and the hierarchy of analyses. The average compliance characterizing Item 21 concerns the quality of the discussion of the external validity of the findings. This element is always mentioned as a caveat but rarely profitably discussed or assessed with empirical data. As we argue in the next section, the discussion about the external validity of experimental findings in the social sciences should not be limited

to the observation of the geographical boundaries of research, but it should address the very specificity of the units involved (and often self-selected) in the experiments. Doing so would help us to gain a deeper understanding of the context in which the mechanisms elicited by the interventions are at play and to solicit reflections on the transferability of the results to other contexts. Interpretation of the results (22) is, instead, well embedded in the relevant literature in the majority of cases.

The sections providing information about registration of the trial (23), availability of experimental protocol (24), and funding (25) are quite deficient, especially if compared with the methods and results areas. Italian trials in education are normally transparent in defining the role of the funders, although in the Italian educational context – given the lack of patents and of a market for educational programs – the issue itself is not very relevant. However, the same cannot be said for the preregistration of trials. Preregistration is a relatively recent practice in the social sciences, and Italian studies comply only in a minority of cases (19 %). Preregistration requires researchers to draft a logic model of their research and explicit outcome measures and the hierarchy between them, impeding the creation of breeding grounds for conscious or unconscious "fishing for effects" or similar practices. The scarcity of preregistered experiments is, in our view, the unifying element underlying the negative aspects of the item compliance observed in Figure 1 and is particularly risky considering that the programs' evaluators and designers are frequently the same people. In fact, the lack of logic models feeds back on the problematic elements discussed before, particularly the prespecification of the outcomes and the multiplicity of analyses. Finally, lack of experimental protocols is the last element that sees vast room for improvement, even if, normally, interventions are well described (see Item 5).

5 Discussion

In reviewing the 26 RCTs illustrated above, we identified three critical points to which future experimental studies should pay more attention: preregistration of the trials, description of the randomization procedure, and assessment of external validity. These weaknesses seem particularly relevant in the Italian context, in which as we saw, the role of policy designers frequently overlaps with that of policy evaluators.

Regarding common weaknesses in the assessed RCTs, the first one has to do with the scarce diffusion of the practices of preregistering the experimental studies and, relatedly, of prespecifying primary and secondary outcomes of the research. This lack of clarity in the experimental design is connected – in some cases – with the inclusion of multiple outcomes in the analyses. In some circumstances, the formulation of precise hypotheses and the elaboration of a precise preanalysis plan are made difficult by contextual conditions (e. g., codesign of the intervention and the research with a policy-maker; uncertainty about the actual rollout of the tested

intervention; complex, multifaceted programs) or by the lack of well-established outcome measures. This point stated, there is surely room for improvement. The practices of prespecifying the research hypotheses and of preregistering the studies' protocols would significantly improve the transparency of the experimental studies, reduce the risk of researchers' "fishing for effects", and avoid post hoc interpretations of the findings. In doing so, it would be helpful to ground the research design more explicitly in sociological theory, thus improving the connections between the intervention to be tested and its hypothesized outcomes.

A second weakness is that the reviewed studies do not always provide detailed descriptions of the adopted randomization procedures, nor do they always prove the integrity of the design. In fact, the studies would benefit from a richer and more detailed description of the randomization process in a broader descriptive framework, encompassing both the participants' enrolment and their retention in the study. The adoption of the CONSORT-recommended study flow charts could be an easy and effective solution to this problem.

Third, the assessment of the external validity is frequently limited to the consideration that the sample on which the study is based is not randomly chosen from the population and hence not fully representative. While this aspect is surely one that must be explicitly acknowledged, it is clearly not sufficient, especially when assessing interventions or the relevance of theoretically defined causal mechanisms. The process by which subjects are enrolled in the study and the peculiar characteristics of the analytical sample should be described in detail and fully considered when discussing the results of the experiments to assess the relevance of contextual factors in determining the detected impacts.

On the basis of what we detected, we now complement the analysis by discussing some proposals for the adaptation of the CONSORT criteria to the social sciences, in light of the presence of peculiar features of experiments in social research not covered by the present standards since they were formalized for a different scientific field.

The need to assess more broadly the external validity of the experimental sample, going beyond the mere issue of representativeness by fully considering the recruitment process and contextual factors, leads us to the second set of considerations emerging from this review. We refer to some methodological aspects of RCTs in social research being not entirely considered by the CONSORT checklist. More precisely, we argue that the peculiar features of experiments in social research – compared to clinical trials – call for the creation of an autonomous set of recommendations based on the CONSORT criteria but tailored to the needs of our field of inquiry. Beyond the already identified issue of broadly intended external validity, we refer here to (at least) two other crucial features of social experiments.

The first is that the underlying logic model and/or theory of change should be explicitly discussed when assessing interventions through experiments in social research. Indeed, the link between the features of the intervention and its outcomes

is not always self-evident. Otherwise, there is a serious risk of dealing with interventions that, due to their complexity, appear as "black boxes", not allowing researchers to learn which of their components were effective and which were not. In addition, a detailed theoretical model would also be useful to promote the recommended practice of prespecifying the outcomes, forcing researchers to identify and register them. Systematic integration of experimental studies with a detailed theory of change (Weiss 1995) and a logic framework (Kellogg Foundation 2001) would be extremely beneficial (Martini and Sisti 2009).

A related, second feature of social experiments that should draw more attention is implementation analysis. Monitoring compliance with random assignment does not seem sufficient, especially when addressing complex interventions implemented by multiactor networks, which are not always homogeneous in terms of implementation standards and quality. Developing in-depth knowledge about what happened during the delivery of an intervention is crucial to interpreting the experimental estimates and contributes to the policy relevance of experimental evidence. Hence, more common use of qualitative evidence and mixed methods approaches is likely a promising way to enhance the quality and usability of experimental findings. Moreover, in this case, a prespecified theory of change and logic framework would also be extremely useful.

Finally, there are two aspects not considered at all in the CONSORT statement, but that – we argue – should be considered when designing RCTs in the social sciences and when reporting results. First, we refer to the issues of contamination and spillover effects, two features embedded in the social processes occurring in experimental settings. Documenting whether and how experimental groups interacted and to what extent the interventions affected them seems valuable, not only to assess the internal validity of experiments in social research but also to inform participants about unintended social consequences of interventions. Second, the last element not considered in the CONSORT checklist but once again relevant to developing considerations about the effectiveness of the interventions is their scalability. As stated above (see footnote 6), efficacy trials conducted on a small scale and in highly supervised settings (leading to "superrealizations") risk overestimating the impact of interventions that, once scaled up, fail to generate impacts of the same magnitude as those detected during the evaluation phase. This feature seems to be often neglected in the comments on the results. Inducing researchers to reflect upon interventions' scalability would also be useful for several other aspects discussed above, such as external validity, the logic model, and implementation issues.

6 Conclusions

This article provides a review of the experimental literature in education recently produced in Italy. We used the checklist developed by CONSORT as a tool to assess the quality of experimental reporting. While acknowledging the positive aspects of this flourishing context, our work emphasizes the existence of significant areas of improvement in both the quality of reporting and the research protocols adopted thus far. In addition, we identify the need for the integration of the existing scientific standards for trial reporting and suggest specific dimensions to be covered to render the CONSORT checklist a research tool that is fully operational in the domain of the social sciences.

An obvious limitation of this study is the geographical, disciplinary, and temporal scope, which is confined to a specific set of studies conducted in education in the 2010s in Italy. However, the trends highlighted in this paper, we believe, could also be illustrative of the processes at play in other contexts in which RCTs are gaining ground over more traditional research designs. On the basis of the review presented in this paper and the discussion that followed, we propose three main issues to which social researchers interested in experimental methods should pay much more attention.

The first issue is linked to a key recurring element that emerged from our inquiry, which is the tension between the rigidity required by a sound experimental design and the need to adapt the experimental method to the substantive complexity of educational processes. This issue seems related to the weak guidelines provided by social theory and/or the lack of solid descriptive evidence for the issues investigated. In contrast to the medical sciences, in which experimental trials have long been implemented, experiments in education and in social science in general are often a way to "establish the phenomenon", in the sense of Merton (1987), that is, to highlight empirical regularities previously not fully known as such.

In other disciplines, predefined (and preregistered) outcomes and hypotheses have emerged as an internal regulatory procedure to induce researchers to: a) specify the theory of change underlying the intervention; b) underscore the mechanisms that it aims to activate, thanks to a definition of the key elements of the program; and c) emphasize the assumptions not supported by the data. We believe that our discipline would benefit from a broader adoption of this procedure, and this benefit would also be larger considering the blurred boundaries between the roles of policy designers and those of policy evaluators. We again discuss this point in the conclusion of this article. However, it is useful to bear in mind that predefined procedures and guidelines should aim primarily at fostering theory-relevant and transparent research but, at the same time, allow researchers to swiftly adapt research protocols to the changing contexts in which the research takes place. We believe that this tension between flexibility and rigor will endure given that, in this field, the theoretical and

methodological guidance, in terms of definition of outcomes and of measurement scales, are weaker than elsewhere.

We believe that the problems and uncertainties entailed in the process of preregistration would be alleviated by the adoption of a series of preparatory steps: interventions to be assessed should be implemented only after careful preparation, consisting of reviewing the existing evidence, specifying the theoretical framework, and establishing the specification (as precisely as possible) of the theory of change around which the intervention is built (Weiss 1995; Kellogg Foundation 2001). Pretesting the interventions in pilot studies and integrating qualitative researchers in this phase could play crucial roles.

Both issues might be effectively addressed by spreading the CONSORT standard and guidelines among social scientists or, perhaps better, by promoting some type of CONSORT-derived standards adapted for the social sciences, reducing the requirements for some items and better specifying others. Such standards should be circulated among researchers and among journal editors, with the medium-term goal of creating a new benchmark for experimental publications in social science journals sustained by both top-down requirements implemented by the latter and a bottom-up movement pushed forward by the former.

We conclude with a final remark about the use of experimental evidence to inform education policy. As our review of Italian studies has shown, in social science experimental research, the relationship between researchers and stakeholders might be more diversified and less structured than it is in clinical trials. The Italian case might well be representative of the majority of European countries, where the institutionalization of experimental research is relatively weak compared to the US and the English-speaking world in general.

On the basis of the experience described in this article, we outline here two scenarios that are typical of such contexts, emphasizing the potential threats to research validity that are embedded in them. The first is given when policy-relevant questions are addressed by researchers, rather than by policy-makers or program managers. In this case, we must be aware that the structure of incentives (based on what is publishable, rather than what is policy relevant) might introduce distortions, i. e., in the ex post and ad hoc choices of the outcomes to analyse. While experimental research allows for a new role for social scientists in policy-making, its weak institutional embeddedness might, to some extent paradoxically, endanger its very scientific merit. A second scenario arises when researchers must negotiate experimental interventions with project stakeholders – be they public decision-makers, private not-for-profit bodies, or corporations. In a context marked by a limited use of evidence-based policy and scarce contacts between academia and civil society, researchers must invest in clear and transparent communication with stakeholders in the early stage of research to ward off misunderstandings about the role of the evaluation. It is possible that policy-makers and program managers

conceive experimental research as a means to confirm their a priori beliefs about the effectiveness of an intervention. Beyond the specific interests that might be at stake, this type of misunderstanding arises from those who lack research training (and are not used to confronting researchers) disregarding the distinction between a good implementation of an intervention and its efficacy. Selective reporting of the results might then become a matter of negotiation insofar as they do not conform to stakeholders' expectations.

We argue that, once again, the downsides of both scenarios might be alleviated as experimental scientists in the social sciences improve the robustness of their research, adopting the practice of preregistering their trials while the institutionalization of this type of research proceeds.

7 References

Andrew, Erik, Aslam Anis, Tom M. D. Chalmers, Mildred Cho, Mike Clarke, David Felson et al. 1994. A Proposal for Structured Reporting of Randomized Controlled Trials. *Jama* 272(24): 1926–1931.

Argentin, Gianluca, Barbieri, Gianna, Falzetti, Patrizia, Pavolini, Emmanuele, and Roberto Ricci. 2017. I divari territoriali nelle competenze degli studenti italiani: tra fattori di contesto e ruolo delle istituzioni scolastiche. *Politiche Sociali, Social Policies* 1: 7–28.

Baldassarri, Delia, and Maria Abascal. 2017. Field Experiments Across the Social Sciences. *Annual Review of Sociology* 43: 41–73.

Ballarino, Gabriele. 2015. Higher Education, Between Conservatism and Permanent Reform. Pp. 209–236 in *The Italian Welfare State in a European Perspective: A Comparative Analysis*, edited by Ugo Ascoli, and Emmanuele Pavolini, Cambridge: Polity Press.

Berk, Richard A. 2005. Randomized Experiments as the Bronze Standard. *Journal of Experimental Criminology* 1(4): 417–433.

Bratti, Massimiliano, Daniele Checchi, and Filippin, Antonio. 2007. Geographical Differences in Italian Students' Mathematical Competencies: Evidence from Pisa 2003. *Giornale degli Economisti e Annali di Economia* 66(3): 299–333.

De Blasio, Guido, Antonio Niscita, and Pammolli, Guido (eds.). 2021. *Evidence-Based Policy! Ovvero perché politiche pubbliche basate sull'evidenza empirica rendono migliore l'Italia*. Bologna: Il Mulino.

Gangl, Markus. 2010. Causal Inference in Sociological Research. *Annual Review of Sociology* 36: 21–47.

Jackson, Michelle, and David R. Cox. 2013. The Principles of Experimental Design and Their Application to Sociology. *Annual Review of Sociology* 39: 27–49.

Kellogg Foundation. 2001. *Logic Model Development Guide. Using Logic Models to Bring Together Planning, Evaluation, & Action*. Kellogg Foundation

Martini, Alberto. 2008. How Counterfactuals Got Lost on the Way to Brussels. Paper presented at the *Policy and programme evaluation in Europe: cultures and prospects*. Strasbourg, July 3–4, 2008.

Martini, Alberto, and Marco Sisti. 2009. *Valutare il successo delle politiche pubbliche. Metodi e casi.* Bologna: Il Mulino.

Merton, Robert K. 1987. Three Fragments From a Sociologist's Notebooks: Establishing the Phenomenon, Specified Ignorance, and Strategic Research Materials. *Annual Review of Sociology* 13(1): 1–29.

Moher, David, Sally Hopewell, Kenneth F. Schulz, Victor Montori, Peter C. Gøtzsche, Philip J. Devereaux, Diana Elbourne, Matthias Egger, and Douglas G. Altman. "CONSORT 2010 explanation and

elaboration: updated guidelines for reporting parallel group randomised trials." *International journal of surgery* 10, no. 1 (2012): 28–55.

Ravallion, Martin. 2008. Evaluation in the Practice of Development. *The World Bank Research Observer* 24(1): 29–53.

Schulz, Kenneth F., Douglas G. Altman, David Moher, and the CONSORT Group. 2010. CONSORT 2010 Statement: Updated Guidelines for Reporting Parallel Group Randomised Trials. *Annals of Internal Medicine* 152(11): 726–732.

Weiss, Carol H. 1998. *Evaluation. Methods for Studying Programs and Policies*. Upper Saddle River, NJ: Prentice Hall.

8 Studies included in the Review

Abbiati, Giovanni, Gianluca Argentin, Andrea Caputo, and Aline Pennisi. 2021. Repetita Iuvant? A repeated RCT on the effectiveness of an at-scale teacher professional development program. *Evaluation review* (forthcoming).

Abbiati, Giovanni, Davide Azzolini, Martina Bazzoli, and Antonio Schizzerotto. 2019. I risultati della valutazione d'impatto. Pp. 223-277 in *È possibile una scuola diversa? Una ricerca sperimentale per migliorare la qualità scolastica*, edited by Daniele Checchi, and Giorgio Chiosso. Bologna: Il Mulino.

Abbiati, Giovanni, Davide Azzolini, Anja Balanskat, Katja Engelhart, Daniela Piazzalunga, Enrico Rettore, and Patricia Wastiau. 2020. Raising Technology-Enhanced Teaching Competences Through Self-Assessment. Experimental Impacts from a European Study on Lower-Secondary Education Teachers. Paper presented at the *ESPAnet Conference Il welfare state di fronte alle sfide globali*. Venezia, September 17-19, 2020.

Abbiati, Giovanni, Gianluca Argentin, Carlo Barone, and Antonio Schizzerotto. 2018. Information Barriers and Social Stratification in Higher Education: Evidence from a Field Experiment. *The British Journal of Sociology* 69(4): 1248–1270.

Alesina, Alberto, Michela Carlana, Eliana La Ferrara, and Paolo Pinotti. 2018. Revealing Stereotypes: Evidence from Immigrants in Schools. *(No. w25333). National Bureau of Economic Research*. https://www.nber.org/papers/w25333.

Aparicio Fenoll, Ainoa, Flavia Coda Moscarola, and Sarah Zaccagni. 2020. Mathematics Camps: A Gift for Gifted Students? *(198/20)* CeRP, Center for Research on Pensions and Welfare Policies. https://www.cerp.carloalberto.org/wp-content/uploads/2020/04/WP_198.pdf.

Argentin, Gianluca, Barbara Romano, and Alberto Martini. 2012. Giocare a scacchi aiuta ad imparare la matematica? Evidenze da una sperimentazione controllata. Pp. 87–98 in *Gli scacchi, un gioco per crescere. Sei anni di sperimentazione nella scuola primaria*, edited by Roberto Trinchero. Milano: Franco Angeli.

Argentin, Gianluca, Aline Pennisi, Daniele Vidoni, Giovanni Abbiati, and Andrea Caputo. 2014. Trying to Raise (Low) Math Achievement and to Promote (Rigorous) Policy Evaluation in Italy: Evidence from a Large-Scale Randomized Trial. *Evaluation review* 38(2): 99–132.

Argentin, Gianluca, Gianpaolo Barbetta, and Francesca Maci. 2015. Cercare soluzioni altrove. Una sperimentazione sull'uso delle Family Group Conferences come strumento di prevenzione del disagio scolastico. Pp. 185–203 in *Politiche sociali innovative e diritti di cittadinanza*, edited by Andrea Bassi, and Giuseppe Moro. Milano: Franco Angeli.

Argentin, Gianluca, Gianpaolo Barbetta, A. Hammad, Mario A. Maggioni, and Domenico Rossignoli. 2018. Non solo a scuola. Relazione finale sulla valutazione degli effetti del progetto. *Research report Milano, Università del Sacro Cuore di Milano*.

Argentin, Gianluca, Gianpaolo Barbetta, and Francesca Maci. 2019. Il progetto Family St. A. R.: analisi degli effetti a breve termine. Research report, Milano, Università Cattolica del Sacro Cuore.

Argentin, Gianluca, and Tiziano Gerosa. 2020. *Migliorare le relazioni a scuola. I risultati della valutazione di impatto*. Trento: IPRASE.

Argentin, Gianluca, Giulia Assirelli, Tiziano Gerosa, and Matteo Moscatelli. 2020. Are Teachers' Relational Skills a Key Leverage for Their Effectiveness? Results from a Large Scale RCT. Paper presented at the *Durham School of Education.*

Azzolini, Davide, Katja Engelhardt, Benjamin Hertz, Sonia Marzadro, Enrico Rettore, and Patricia Wastiau. 2020. Raising Teachers' Retention in Online Courses through Personalized Support. A Randomized Controlled Trial in 10 Countries. Paper presente at the *2020 APPAM Fall Research Conference*. November 11-13, 2020.

Barbetta, Gianpaolo, Paolo Canino, and Stefano Cima. 2019. Let's Tweet Again? The Impact of Social Networks on Literature Achievement in High School Students: Evidence from a Randomized Controlled Trial. *Working Paper n. 81, Università Cattolica del Sacro Cuore*, https://ideas.repec.org/p/ctc/serie1/def081.html.

Barone, Carlo, Giulia Assirelli, Giovanni Abbiati, Gianluca Argentin, and Deborah De Luca. 2018. Social Origins, Relative Risk Aversion and Track Choice: A Field Experiment on the Role of Information Biases. *Acta sociologica* 61(4): 441–459.

De Cataldo, Alessandra, Antonio Fasanella, and Manlio Maggi. 2016. *La comunicazione del rischio chimico. Sperimentazione e valutazione nelle scuole di Roma*. Milano: Franco Angeli.

Del Boca, Daniela, Chiara D. Pronzato, and Lucia Schiavon. 2020. How Parents' Skills Affect Their Time-Use with Children? Evidence from an RCT Experiment in Italy. *CESifo Working Paper No. 8795*, https://ideas.repec.org/p/cca/wpaper/628.html.

Carlana, Michela, Eliana La Ferrara, and Paolo Pinotti. 2018. Goals and Gaps: Educational Careers of Immigrant Children, *Centre for Economic Policy Research*, https://cepr.org/active/publications/discussion_papers/dp.php?dpno=12538.

De Poli, Silvia, Loris Vergolini, and Nadir Zanini. 2018. The Impact of a Study Abroad Programme on Learning Abilities and Personality Traits: Evidence from a Randomization. *Applied Economics Letters* 25(8): 562–566.

Di Tommaso, Maria Laura, Dalit Contini, Dalila De Rosa, Francesca Ferrara, Daniela Piazzalunga, and Ornella Robutti. 2021. Tackling the Gender Gap in Mathematics with Active Learning Methodologies. *Working paper series 16/20, Dipartimento di Economia e Statistica Cognetti Martiis*, https://ideas.repec.org/p/uto/dipeco/202016.html.

Fasanella, Antonio, and Manlio Maggi (eds.). 2011. *Le conoscenze giovanili sulle radiazioni ionizzanti. Intervento e valutazione nelle scuole superiori del Lazio*. Roma: ISPRA.

Gui, Marco, Tiziano Gerosa, Gianluca Argentin, and Lucilla Losi. 2021. Mobile Connectivity and Adolescent Well-being. Evidence from a Randomised Control Trial on a Media Education Training Programme in High Schools. *SocArXiv. July 29*, https://osf.io/preprints/socarxiv/8bd43.

Martini, Alberto, Davide Azzolini, Barbara Romano, and Loris Vergolini. 2021. Increasing College Going by Incentivizing Savings: Evidence from a Randomized Controlled Trial in Italy. *Journal of Policy Analysis and Management*, *40:3, https://doi.org/10.1002/pam.22260*.

Rinaldi, Emanuela E, and Gianluca Argentin. 2020. La torta dell'economia. Valutazione di un progetto di educazione finanziaria scuole primarie. Pp. 223–250 in *Scenari ed esperienze di educazione finanziaria. Risultati dell'indagine nazionale ONEEF e* riflessioni *multidisciplinari*, edited by Luca, Refrigeri, Emanuela E. Rinaldi, and Valentina Moiso. Lecce: Pensa multimedia.

9 Appendix

Table A1 Description of the Studies

Intervention and target				Design						
Study number (see table 1)	Type of intervention	Target population	Geographical area	# of schools/ teachers/classes	# of students	Were schools randomly sampled?	Randomization	Outcome data	Funding	Driver for the research[a]
1	Professional development for teachers	ISCED 2 math teachers	4 regions	174 schools/ 581 teachers	11 000	No	Cluster RCT with blocking	Students: Standardized assessment and survey; teachers: survey	Italian Ministry of Education and Research	2
2	Information campaign on radioactivity and its risks	ISCED 3 students	Local	24 schools	1527	No	Cluster RCT with blocking	Survey and standardized assessment	EU Joint Research Centre	1
3	Chess course	3rd graders	Multi-site	30 schools/ 113 classes	2000	No	Cluster RCT with blocking	Standardized assessment	Private (FSI – Federazione Scacchistica Italiana)	1
4	Professional development for teachers	ISCED 2 math teachers	4 regions	44 schools/ 146 teachers	2800	No	Cluster RCT with blocking	Standardized assessment	Italian Ministry of Education and Research	2
5	Summer English courses	12th graders	Local	not school based	281	No	RCT (individual randomization)	Standardized assessment and survey	Trento Province	1
6	Informational guidance	13th graders	4 provinces	62 schools/ 475 classes	9000	Yes	Cluster RCT with blocking	Surveys	Italian Ministry of Education and Research	3
7	Outreach intervention in schools	6–8th graders	Local	14 schools	261	No	RCT (individual randomization) with blocking	Survey and administrative data	EU, Private Foundations (Cariplo, Fondazione con il Sud, Fondazione Peppino Vismara)	3

Continuation of table A1 on the next page.

Continuation of table A1.

Intervention and target				Design						
Study number (see table 1)	Type of intervention	Target population	Geographical area	# of schools/ teachers/classes	# of students	Were schools randomly sampled?	Randomization	Outcome data	Funding	Driver for the research[a]
8	Personalized informational and academic guidance	ISCED-2 high-achieving «immigrant» students	Multi-site	145	1217	No – population data	Cluster RCT with blocking	Survey, standardized assessment and administrative data	Private (Foundation, Cariplo)	2
9	Information campaign on chemical risks	ISCED 3 students	Local	24 schools	1051	No	Cluster RCT with blocking	Survey and standardized assessment	EU Joint Research Centre	1
10	Nudged informational guidance	Mothers of 8th graders	Regional	44 schools/ 147 classes	408	Yes	Cluster RCT with blocking	Survey	Apulia region	3
11	Provision of a savings account and financial instruction classes	12th and 13th graders	Local	not school based	716	No	RCT (individual randomization) with blocking	Survey	EU Programme for Social Policy Experimentations Supporting Social Investments – Progress 2013	2
12	Funding of school- and class-level projects involving principals, teachers and students	ISCED 2 schools	2 regions	50 schools/ 1400 teachers	4800	No	Cluster RCT with blocking	Students: standardized assessment and survey; teachers: survey	Private (Foundation, Compagnia di San Paolo)	2
13	Professional development for teachers	ISCED 2 teachers	11 provinces widespread in Italy	198 schools/ 2400 teachers	27 000	Yes	Cluster RCT with blocking	Students: standardized assessment and administrative data; teachers: survey and admin. Data	Italian Ministry of Education and Research	3
14	Professional development for teachers	ISCED 2 teachers	National	4178 schools (228 treated)	402 306	Yes	Cluster RCT with blocking	Students: standardized assessment	Italian Ministry of Education and Research	3

Continuation of table A1 on the next page.

Continuation of table A1.

Intervention and target				Design						
Study number (see table 1)	Type of intervention	Target population	Geographical area	# of schools/ teachers/classes	# of students	Were schools randomly sampled?	Randomization	Outcome data	Funding	Driver for the research[a]
15	Participation to volunteering activities	9th–10th graders at risk of drop-out	Local	6 schools	169	No	Cluster RCT (randomization unit: student; cluster: schools and application period)	Survey and administrative data	Centro di Servizio per il volontariato Monza Lecco Sondrio; Private (Foundation, Fondazione Peppino Vismara)	1
16	Use of social networks to teach literature	ISCED 3 students	National	70 schools	1465	No	Cluster RCT with blocking	Survey	Private (Foundation, Cariplo)	3
17	Teachers' exposure to their prejudice	ISCED 2 teachers	Multi-site (5 provinces in the North)	65 schools/ 533 teachers	6031	No	Cluster RCT	Survey, standardized assessment and administrative data	Bocconi University	3
18	Teachers' self-assessment of technology-enhanced teaching	ISCED-2 teachers	Multi-country	469 schools/ 7391 teachers	–	Yes	Randomized Encouragement Design; Cluster RCT (randomization level: school) with blocking + random peer effects design	survey	EU ERASMUS Programme Key Action 3	4
19	Financial literacy & pro-social attitudes courses	3rd-5th graders	Multi-site	10 schools/ 60 classes	1200	No	Cluster RCT with blocking	Survey	Fondosviluppo (private fund for cooperative economy), Federazione delle Banche di Credito Cooperativo del Lazio, Umbria, Sardegna	2

Continuation of table A1 on the next page.

Continuation of table A1.

Intervention and target				Design						
Study number (see table 1)	Type of intervention	Target population	Geographical area	# of schools/ teachers/classes	# of students	Were schools randomly sampled?	Randomization	Outcome data	Funding	Driver for the research[a]
20	Outreach intervention in schools	6–8th graders	Multi-site (4 locations in 4 provinces)	34 schools	540	No	RCT (individual randomization) with blocking	Survey	Private (Foundations – Cariplo, Fondazione Peppino Vismara)	3
21	Professional development for teachers	ISCED 3 teachers	Local	18 schools/ 171 classes	3659	No	Cluster RCT with blocking	Survey	University of Milano-Bicocca and Fastweb	3
22	Professional development for teachers	Recently permanently hired teachers (all grades)	1 province	204 teachers	–	No	RCT (individual randomization) with blocking	Survey	Public research Institute (IPRASE)	2
23	Online personalized support through SMSs to online courses participants	ISCED 2 – Senior and junior teachers	Multi-country	511 schools/ 4090 teachers	–	Yes	Cluster RCT with blocking	course platform data	EU ERASMUS Programme Key Action 3	4
24	Math camp for gifted students	9th-12th grade gifted students	regional	–	1346	No – population data	Cluster RCT with blocking	Survey and standardized assessment	Private (Foundation – Compagnia di San Paolo)	1
25	Information campaign for parents on improving the learning environment at home	Families of 0–6 years old children	Multi-site	20 schools	534	No	RCT (individual randomization)	Survey	Con i Bambini (public foundation co-managed by government and bank foundations)	4
26	Math laboratory sessions	8th graders	Local	25 schools	1044	No	Cluster RCT with blocking	Standardized assessment	University of Torino and the private foundation Compagnia di San Paolo	3

[a] 1: policy makers asked (PM) for evidence on a program; 2: researchers proposed to PM to evaluate an intervention; 3: intervention designed by researchers; 4: evaluation requested in a public call for tenders.