TI 2022-025/V
Tinbergen Institute Discussion Paper

Good or Bad? Short- versus Long-Term Effects of Multigrading on Child Achievement

Gian Paolo Barbetta[1]
Patrick Chuard-Keller[2]
Giuseppe Sorrenti[3]
Gilberto Turati[1]

1 Università Cattolica del Sacro Cuore
2 University St. Gallen
3 University of Amsterdam, Tinbergen

Tinbergen Institute is the graduate school and research institute in economics of Erasmus University Rotterdam, the University of Amsterdam and Vrije Universiteit Amsterdam.

Contact: discussionpapers@tinbergen.nl

More TI discussion papers can be downloaded at https://www.tinbergen.nl

Tinbergen Institute has two locations:

Tinbergen Institute Amsterdam
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 598 4580

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900

Good or Bad? Short- versus Long-Term Effects of Multigrading on Child Achievement*

Gian Paolo Barbetta Patrick Chuard-Keller

Giuseppe Sorrenti Gilberto Turati

March 24, 2022

Abstract

This paper studies the effect of multigrading—mixing children of different ages in the same classroom—on students' short- versus long-term academic achievement in Italy. We cope with the endogeneity of multigrading (and class size) through an instrumental variable identification strategy based on a law that disciplines class composition. By relying on longitudinal data that follow a cohort of Italian students over their compulsory school career, we show that multigrading has a positive short-term effect on achievements. This effect fades away over time to become negative in the long run if students spend several years in a multigrade class. The analysis of mechanisms points to the fundamental role of teachers and suggests that no negative long-term effect arises when multigrade classes are taught by more experienced and motivated teachers. These results reconcile contrasting findings in the literature based on cross-sectional data and a short-term focus.

JEL CODES: I26, I28, R53, H52.

*We wish to thank participants at IWAEE 2021 for helpful comments. We also thank Lucia De Fabrizio (MIUR) for kindly providing data about Italian schools and teachers. Barbetta: Università Cattolica del Sacro Cuore, Milan (IT); Chuard-Keller: University of St.Gallen, St.Gallen (CH); Sorrenti: University of Amsterdam and Tinbergen Institute, Amsterdam (NL), g.sorrenti@uva.nl (corresponding author); Turati: Università Cattolica del Sacro Cuore, Rome (IT).

1 Introduction

About one-third of all children worldwide attend mixed-grade or multigrade classes, i.e., classes that mix children of different ages in the same classroom (UNESCO, 2004). Understanding the effect of multigrading on child development is a priority, first to ascertain the impact of this teaching practice in remote areas where multigrade classes are born out of necessity, due to constrained economic and human resources, and second to understand the consequences of this method in areas where it has gained support even if the circumstances would allow for single-grade classes.

Supporters of multigrade classes argue that exposing children to peers of different ages might increase their cognitive and socio-emotional skills. This positive impact might be generated by, e.g. exposure to a richer vocabulary, a more demanding school curriculum, or the need to "take care" of peers of different ages. However, elements like students' relative age in a multi-age class or teachers with limited experience with multigrading might undermine the potential of this teaching practice.

Despite the sizable (and increasing) use of multigrading worldwide, the evidence of its *causal* impact on child development is mixed, mostly based on cross-sectional data, and exclusively focused on short-term outcomes.[1] On the one hand, studies such as Leuven and Rønning (2016) and Barbetta et al. (2021) find evidence of a positive short-term impact of attending a multigrade class on a child's development. On the other hand, studies like Checchi and De Paola (2018) find a negative impact of multigrading for fifth-graders in Italian primary schools. How can these results be reconciled? Should multigrade teaching be considered an effective educational practice?

This paper tries to answer these questions by assessing the short- and long-term impacts of multigrading on children's cognitive achievements.[2] The analysis builds on a new longitudinal data set that contains repeated test scores for the

[1]Some excellent earlier studies—surveyed in Little (2001)—provide interesting analyses on the effect of multigrading. However, they do not properly address endogenous sorting of students into multigrade classes.

[2]We label as long-term outcomes those measured more than five years after attendance of a multigrade class.

universe of Italian students in primary and lower-secondary education. Specifically, the data report the math and language test scores in the standardized INVALSI (National Institute for the Evaluation of the Instruction and Training System) test that Italian students take in Grade 2 and 5 of primary school, and in Grade 8 of lower-secondary (or middle) school.

In the spirit of Barbetta et al. (2021), (i) we develop an algorithm to identify students in multigrade classes, and (ii) we deal with endogenous sorting into multigrade classes through an instrumental variable (IV) identification strategy. Indeed, we exploit the Italian law DPR 81/2009—prescribing cut-offs defined out of the number of students in a specific cohort—that disciplines the formation of single- versus multigrade classes and class size. Such an identification strategy—inspired by Angrist and Lavy (1999)—allows us to estimate the effect of multigrading on child development net of the confounding effect of class size.[3] Specifically, our approach relies on predicted-by-the-law grade composition of classes and predicted-by-the-law class sizes as instruments for the observed grade composition of classes and class sizes.

We start by analyzing the impact of attending a multigrade class in Grade 2 on the INVALSI test taken at the end of the same grade. This analysis sheds light on the short-term effect of multigrading. Subsequently, we consider the effect of having attended a multigrade class in Grade 2 on the performance of the same student in Grade 5 three years later. Finally, to consider a longer time horizon, we analyze the effect of having attended a multigrade class in Grade 2 on the performance in Grade 8, at least six years later. At this last stage, students attend a different cycle of education in which multigrading is rarely used.[4]

Our results display an interesting pattern: multigrading positively affects short-term cognitive achievements, but the effect tends to vanish over time. Indeed, attending a multigrade class in Grade 2 of primary school positively affects the

[3] Indeed, multigrade classes are on average smaller than single-grade classes, and class size is an important potential driver of students' achievement, see e.g. Chingos (2013).

[4] The analysis of Grade 8 is particularly important as upon completion of Grade 8 students are tracked into different secondary schools. The school performance in Grade 8 plays a major role in the choice of the school. Different secondary school tracks are associated with different probabilities to access university and different future labor market returns.

short-term performance—about 16 percent of a standard deviation—in the standardized test in math and language. However, in Grade 5 the effect is zero, while in Grade 8 it turns into a negative and statistically significant 10 percent of a standard deviation. The results are robust across different empirical specifications.

In the second part of the paper, we rationalize the switch in the sign of the effect of multigrading when comparing Grade 2 and 8. First, we stress the importance of a correct interpretation of the effect of multigrading. Indeed, we show that relative age in a multigrade class plays an important role in shaping the effectiveness of this educational practice. Children who share their multi-age class with older peers largely benefit from interactions with more-mature peers. However, given the strong persistence of multigrading in Italian primary schools—multiple years of attendance of a multigrade class—the initial relative advantage for multigrade students is lost over time as students switch from being the younger cohort to the older cohort in the classroom. This evidence—aligned with the findings in Leuven and Rønning (2016) and Barbetta et al. (2021)—points to the importance of correctly defining the effect of multigrading on child development. In the early grades of primary school, e.g. Grade 2, the effect should be considered the "pure" effect of multigrading. In later grades, e.g. Grades 5 or 8, in the majority of cases the effect turns to be the "cumulative" effect of attending several years of multigrading.

As a second step, we consider whether the long-term (or cumulative) detrimental effect of multigrading in the Italian primary school system could be avoided, despite the persistence of multigrading. Our results suggest that the quality of multigrade teaching implementation is the key driver for the success of this practice. For instance, we proxy quality of multigrading implementation with some teachers' characteristics. It is well known that teachers are one of the main actors when it comes to fostering a child's development (Alan et al., 2021b; Jackson et al., 2014; Xu and Ran, 2020). This is also true for multigrade teaching (INDIRE, 2019, 2020). Teaching in a multigrade class implies a considerable effort by teachers who are required to adapt standard forms of teaching to a class made by children of different ages and therefore skills.

INDIRE (2019) argues that teachers' low turnover represents one of the main predictors of successful use of multigrade classes. Teachers exposed to high turnovers experience disruptions in their learning process concerning how to run a multigrade class. Moreover, their motivation can be negatively affected by their temporary role in a certain school. As a result, students attending classes with high levels of teacher turnover risk being exposed to low-quality multigrade teaching and being discouraged by continuous changes of teachers. We test this hypothesis by collecting additional data on teachers' quality as measured by their turnover at the school level. The analysis of teacher turnover displays heterogeneous effects. Specifically, when multigrade classes in Grade 2 are taught by teachers with a high level of motivation and involvement—as proxied by being employed in a school with a large share of teachers under a permanent contract—the long-term detrimental impact of multigrading disappears. This evidence suggests that multigrading can provide no harm to children's cognitive development, including in contexts with strong persistence.[5]

As a final step, we verify whether attending a multigrade class can affect students' probability of enrolling in worse lower-secondary schools. If so, this would potentially explain the long-term negative impact of multigrading on students' achievements. To test this assumption, we collect information on school resources, e.g. computer availability, and investigate possible heterogeneous long-term effects of multigrading by school resources. We do not find any evidence supporting the view that attending a multigrade class in Grade 2 affects school quality in the lower-secondary cycle of education in Grade 8. The negative impact of multigrading in Grade 8 seems independent of lower-secondary school quality. Therefore, this channel is unlikely to explain the long-term effect of multigrading.

Overall, our findings shed lights on why different studies display a heterogeneous impact of multigrading on students' cognitive development. First, given the strong persistence of multigrading, the interpretation of its impact should be mostly considered as a cumulative effect of several years of exposure to multi-age

[5] Admittedly, our analysis might mask some correlation between teachers' characteristics (and type of contract) and other school resources. This limitation does not undermine our analysis as this exercise only aims at understanding whether it is possible to avoid the negative long-term impact of multigrading through the provision of a high-quality standard of multigrade teaching. Whether the effect is only driven by teachers or also other school resources holds second-order importance for the analysis.

teaching. This explains the negative impact of multigrading found by some cross-sectional studies, e.g. by Checchi and De Paola (2018) or in a very specific sample of Californian students by Sims (2008), while other studies, e.g. Leuven and Rønning (2016) and Barbetta et al. (2021), find positive effects. Second, our study reinforces existing evidence on the positive short-term effect of multigrading. Moreover, it also shows that multigrading—if effectively taught by teachers—can prevent harming the long-term cognitive development of children. The latter result is crucial for policymakers in charge of providing learning opportunities to students living in scarcely populated areas where multigrading is the only available (and possible) schooling option. Finally, this work—like the others in the literature—might neglect part of the potential of multigrading. Even if the long-term effect on cognitive achievements is zero, this work is unable to unveil possible effects of exposure to peers of different ages on other important skills such as non-cognitive and socio-emotional ones. Recent works such as Alan et al. (2019), Alan et al. (2021a), Kosse et al. (2020), and Sorrenti et al. (2020) stress the malleability of soft skills and the high returns of investments in such skills. Multigrading might play an important role in shaping these skills, e.g. through exposure of children to more mature peers or by fostering the sense of responsibility toward younger peers. For this reason, future research should focus on multidimensional measures for child development to gain a comprehensive assessment of the impact of multigrading.

The remainder of the paper is structured as follows. In Section 2, we provide the essential background information on the Italian schooling system and the rules governing class formation. Sections 3 and 4 describe the longitudinal data used in the estimations and the empirical strategy to identify the impacts of multigrading. Section 5 reports the results of the empirical analysis. Section 6 investigates the mechanisms underlying the multigrade effect. Finally, Section 7 concludes the paper.

2 Institutional Background

The analysis in this study is focused on Italian students in their first cycle of education, which includes primary and lower-secondary school. The entire first cycle of education is compulsory and free of charge in Italy. Primary education (ISCED 1) starts at the age of six and lasts five years (Grades 1 to 5), while lower-secondary education (ISCED 2) starts at the age of eleven and lasts for three years (Grades 6 to 8). At the end of the first cycle of education, in Grade 8, students take a national exam to gain a lower-secondary education diploma that gives access to upper-secondary education (ISCED 3), the first two years of which are also compulsory.

Primary school is mainly aimed at providing basic training in writing, reading, and mathematics, and a basic knowledge of several different subjects (science, history, geography, art, music, and English language). This basic training is mostly provided by non-specialized teachers. Training provided in primary school is further developed in lower-secondary school, where the same subjects (and a few additional ones) are taught by teachers specialized in each subject.

Within the first cycle of education (as well as for the following cycles), most schools are state run and parents can freely enroll their children in the school that they prefer. According to data provided by the National Institute of Statistics (ISTAT), 16,948 primary schools operated in the country in 2018, hosting more than 2.7 million students over the five grades. The number of lower-secondary schools is much lower, with a total of 8,064 institutions hosting more than 1.7 million students over the three grades. Private schools represent about eight percent of institutions and host about six percent of students in primary education, while they include about eight percent of institutions and host less than four percent of students in lower-secondary education. No official statistics about multigrading are currently available. However, according to our data, about 23 percent of primary schools and about five percent of lower-secondary schools located in municipalities with no more than one school for each level adopt multigrade teaching.

Class formation in each Italian school is regulated by DPR 81/2009.[6] This law defines thresholds—based on the number of students of the same grade enrolled in a specific primary or lower-secondary school—that influence both the proba-

[6]Two autonomous regions—Valle d'Aosta and Trentino Alto Adige—are not subject to these rules, as they enjoy broad autonomy for educational policies.

bility of being assigned to a multigrade class and class size. Rules for primary and lower-secondary schools are very similar.

For *primary* schools, law DPR 81/2009, article 10 specifically establishes that:

- single-grade classes should enroll a minimum of fifteen and a maximum of 26 students;
- multigrade classes should enroll a minimum of eight and a maximum of eighteen students;
- in isolated villages, small islands, and areas characterized by linguistic minorities, single-grade classes could be created with a minimum of ten students. Moreover, the law allows to reduce the maximum number of students when children with disabilities are enrolled in a class.

For *lower-secondary* schools, article 11 of the same law prescribes that:

- single-grade classes should enroll a minimum of eighteen and a maximum of 27/28 students;
- multigrade classes should enroll a minimum of ten and a maximum of eighteen students;
- in isolated villages, small islands, and areas characterized by linguistic minorities single-grade classes could be created with a minimum of ten students. Moreover, the law allows reducing the maximum number of students when children with disabilities are enrolled in a class.

Within the first cycle of education, admission of students to schools is driven by uniform national criteria. While the school year (SY) starts by mid-September, families need to apply to a specific school well in advance—by January–February—for both primary and lower-secondary schools. Based on the number of applications, school principals decide about admission, the allocation of children to classes and consequently class size. Within a month from their application, families are notified about admission, which is almost always granted if students live nearby the school. In fact, the distance between the student's house and the school is the most relevant criterion determining admission. Usually, students (and also teachers) are assigned to classes shortly before the beginning of the SY.

3 Data

We use longitudinal data following a single cohort of Italian students over time. For each student we observe individual test scores on the second-, fifth-, and eighth-grade national standardized test administered by INVALSI (National Institute for the Evaluation of the Instruction and Training System). The so-called INVALSI test was introduced by Law 176/2007, and it has subsequently been administered yearly to second-, fifth-, eighth-, and tenth-grade students attending public or private schools.[7] The INVALSI written test is aimed at monitoring the skills of students in mathematics and language, but it also includes other subjects, such as science and foreign language. Each test comprises a set of multiple-choice items followed by open-response questions. Students must complete the test in 45 to 90 minutes, depending on the subject and grade.

Besides individual test scores, INVALSI data also contain background information about students' characteristics such as gender, nationality and pre-primary school attendance, parental education level, and profession. Data are fully anonymous. Data only report the numeric "INVALSI code", which identifies each student for her entire school career, plus numeric class and school codes. Neither students, classes, nor schools can be identified based on these codes.[8]

The unit of observation in our data is the single student. In particular, we consider the whole cohort of students who were second-graders in the SY 2012/2013. This is the first cohort for which longitudinal data can be created. We include in the data set the INVALSI test scores of this cohort of students in Grades 2, 5, and 8, i.e. SY 2012/2013, 2015/2016, and 2018/2019, respectively. Subsequently, we match student data with information on schools and classes. Before SY 2018/2019,

[7]Due to the COVID-19 pandemic, the test was not administered in SY 2019/2020.

[8]The INVALSI data set does not cover the two autonomous regions of Valle d'Aosta and Trentino Alto Adige. However, students in these regions account for less than two percent of the population of Italian students.

INVALSI data did not report the grade composition of each class, making it impossible to classify students as attending a single- versus multigrade class. However, the INVALSI data contain geographical and demographic information about schools, including the province as well as the population, size, and altitude of the municipality where the school is located.

Following Barbetta et al. (2021), we merged the longitudinal data set of students with information included in two different administrative archives. First, we use the School Register data provided by the Italian Ministry of Education (MIUR). The data set contains information about each Italian school, individually named, including the number of multigrade classes. Second, we use the Municipality Register data produced by ISTAT, which—for each Italian municipality—include the same demographic and geographical information described in the INVALSI data set.

We use geographic and demographic data about municipalities to bridge information in the INVALSI and in the School Register data sets. Subsequently, for municipalities hosting a single (primary and/or lower-secondary) school, we are able to identify each school's name and characteristics, including the number of multigrade classes. Therefore, in each of these schools, through the information provided by the School Register, we can assess the grade composition of each student's class—single-grade versus multigrade—as well as class size.[9] Moreover, the identification of schools' names together with the information on the municipality in which these schools are located allow us to merge our data set with additional data on school and municipality characteristics that are important to explore the mechanisms underlying the effect of multigrading (see Section 6).

Our data construction process generates a final sample including the entire cohort of Italian second graders who attended a primary school located in a municipality hosting only one primary school in SY 2012/2013. This sample of students is followed over time until SY 2018/2019, when they become eighth-graders and finish lower-secondary school. Overall, we end up with 3,999 primary schools

[9] Appendix A describes the algorithm used to classify whether a school used single- or multigrade teaching in each grade. It is important to mention that it is never observed that a school has both single- and multigrade classes for the same grade.

out of 15,248 covered in the School Register data and about 70,688 students out of the about 500,000 in each year-cohort all over the country. It is important to highlight that municipalities hosting more than one primary school are relatively rare in Italy—no more than 35 percent of those hosting a primary school—and quite common only in urban areas. Furthermore, 53 percent of Italian municipalities are classified as rural or inner areas. In most instances, these are small municipalities, where multigrade classes are quite common.

Admittedly, focusing exclusively on municipalities that host no more than one primary school represents a potential limitation of our data. In these areas, multigrading is often a necessary choice due to the small population. At the same time, focusing on municipalities with a unique primary school helps to keep the problem of non-random assignment of students in single- or multigrade classes under control. In fact, families living in municipalities hosting only one school find it more difficult to choose the preferred school for their children, unless they are willing to bear the costs of driving to a different municipality. These commuting costs increase directly with the distance from the closest alternative school.[10]

Table 1 shows summary statistics for our sample. To ease the interpretation of the results, the test scores in math and language are combined to obtain a single performance measure with a mean of zero and standard deviation of one. The combined measure takes the average of the normalized reading and math scores and then by normalizing the combined score. About six percent of students have attended a multigrade class in Grade 2, and the share is similar in Grade 5, whereby this similarity confirms the strong persistence of multigrading in the Italian education system. Persistence of multigrade teaching in primary schools is supported by the fact that students in a multigrade class in Grade 2 spend on average 3.6 years in such type of class. On the contrary, multigrade teaching in lower-secondary school is rare, with about one percent of the sample in Grade 8. The average class size in the sample is nineteen students per class. The sample is equally split by gender and about eight percent of students have

[10] Barbetta et al. (2021) show that distance to the closest alternative school plays a limited role in shaping the point estimates for the effect of multigrading. Therefore, in the following empirical analysis, we will abstract from considering this aspect, e.g. by including an extra control variable for distance to the closest alternative primary school.

parents who are migrants.[11] With reference to family characteristics, about ten percent of parents have a university degree.

Table 2 tests differences (in means) between the sample of students attending a multigrade versus a single-grade class in Grade 2. As the last two columns show, the majority of variables are significantly different. While multigrade students overperform single-grade students in Grade 2, the opposite is true for Grade 8. As expected, class size differs by sub-groups, with multigrade classes that are usually smaller—on average with six fewer students—than single-grade classes. Isolating the effect of multigrading from the effect of class size will be discussed and addressed in the empirical model section. The two subsamples are equally split by gender and despite some statistically significant differences, the migration background is also comparable across groups. On average, students in single-grade classes have slightly more educated parents than their counterpart in a multigrade class. Geographically, multigrade classes are less common in the north of Italy and more widespread in the south.

The difference in observables between the sample of students in multigrade versus single-grade classes will be addressed by (i) controlling for all of the observable characteristics displayed in the table in our empirical model; and (ii) dealing with possible unobservable factors shaping the attendance of a multigrade class in an IV setting.

4 Empirical Strategy

This section describes the empirical strategy of the study. First, we introduce the empirical model underlying the analysis. Second, we discuss the identifying assumptions required to interpret the effect of multigrading on students' performance causally.

4.1 Empirical Model

We aim to estimate the causal effect of attending a multigrade class in Grade 2 on a child's standardized test score in Grades 2, 5, and 8. Our empirical model of

[11]Given the longitudinal nature of the data with each student observed in Grades 2, 5, and 8, we do not report age in the table as it is scarcely informative.

interest takes the following form:

$$Score_{i,g\in[2,5,8]} = \beta_0 + \delta MG_{i,g=2} + \beta_1 CS_{i,g=2} + \boldsymbol{X}'_{i,g=2}\boldsymbol{\beta_2} + \epsilon_{i,g=2} \quad , \tag{1}$$

where i is an index for each student and g stands for the grade. $Score$ represents the mean of student i's standardized performance—with a mean of zero and standard deviation of one—in the math and language sections of the INVALSI test taken in grade g.[12] MG is an indicator for student i's attendance of a multigrade class in Grade 2. CS represents the class size in Grade 2. $\boldsymbol{X}$ is a vector containing a set of control variables for child characteristics (age, gender, nationality, first- and second-generation immigrant) and parental background (mother's and father's education and profession). It also includes controls for the five macro-regions of Italy: the northwest, the northeast, the center, the south, and the islands. All control variables refer to Grade 2 of primary school. ϵ is the error term of the model.

Identifying the causal effect of multigrading on school performance is a challenging task, as attendance of a multigrade class might correlate with unobserved determinants of a student's achievement. Thus, ordinary least squares (OLS) estimates of the effect of multigrading on children's school performance might suffer from bias due, e.g. to selection on unobservables. This threat is reinforced by the descriptive analysis, which highlighted that multigrade classes usually include fewer students than single-grade classes. Therefore, a credible empirical analysis needs to isolate the effect of multigrading from the effect of class size.

In the spirit of Barbetta et al. (2021), our identification strategy exploits some features of the Italian framework to address the identification challenges in an IV setting. Specifically, as previously described, the Italian law DPR 81/2009 disciplines how classes should be formed in primary (and lower-secondary) schools and constitutes an exogenous source of variation to be used to overcome endogeneity concerns relative to MG, the indicator for attendance of a multigrade class in Grade 2.

DPR 81/2009 prescribes cut-offs based on the number of students of the same

[12]The math and language test scores are the only test scores available for all grades.

cohort enrolled in the same school. It states that in principle, a cohort made by fewer than ten students should not determine the creation of a single-grade class. This ten-student cut-off represents a possible instrument to correct the endogeneity underlying the attendance of a multigrade class. The law also introduces other cut-offs. If the size of the cohort of students enrolled in the same school is between ten and fifteen, both multigrade and single-grade classes are possible. Finally, the law establishes that for cohorts of 15 to 26 students or cohorts with at least 27 students, multigrade classes should never be formed. Moreover, for the latter group more than a single-grade class should be created, therefore implying a direct effect on class size. Our IV analysis exploits these cut-offs to deal with the endogeneity of MG and—in one of the specifications—class size.

As in our previous work, we construct four mutually-exclusive indicator variables for cohort size to be used as instruments for the (actual) attendance of a multigrade class and class size. The first indicator ($\mathbb{1}[CohortSize_s < 10]$) takes the value of one if the cohort enrolled in a certain school s contains fewer than ten students. This variable should be a strong predictor of the probability of attending a multigrade class for students in school s. The second indicator ($\mathbb{1}[10 \leq CohortSize_s < 15]$) is for cohorts made by ten to fourteen students and should show a positive correlation with attendance of a multigrade class. The third indicator takes value of one for a cohort with 15 to 26 students ($\mathbb{1}[15 \leq CohortSize_s < 27]$). It should not affect the attendance of a multigrade class, although it should shape class size. Finally, the last indicator ($\mathbb{1}[CohortSize_s \geq 27]$) takes value of one if the cohort size in school s exceeds 26 students and should only explain class size as the cohort should be too large for multigrade classes to be formed.[13]

Starting from the definition of the four indicator variables for cohort size, our IV strategy leads to the following first-stage equation:

$$\begin{aligned} MG_{i,g=2} = \lambda_0 + \lambda_1 \mathbb{1}[CohortSize_s < 10] + \lambda_2 \mathbb{1}[10 \leq CohortSize_s < 15] \\ + \lambda_3 \mathbb{1}[15 \leq CohortSize_s < 27] + \lambda_4 \mathbb{1}[CohortSize_s \geq 27] \\ + \lambda_5 CS_{i,g=2} + \boldsymbol{X}'_{i,g=2} \boldsymbol{\lambda_5} + \delta_{i,g=2} \quad , \end{aligned} \tag{2}$$

[13] It is important to recall that in Italy it is almost impossible to observe a school with both a multigrade and a single-grade class for the same grade.

where s is the school attended by student i.

Before discussing the identifying assumptions underlying the IV strategy, it is crucial to discuss how we deal with the role of class size in Equation 1. We consider class size in a dual way: first, we consider class size as a standard control variable; and second, given that class size might suffer from the same sources of endogeneity as attendance of a multigrade class, we treat class size as an additional endogenous variable in our model. Therefore, we instrument class size with the set of instruments defined above under the testable assumptions that two of the cut-offs identified by DPR 81/2009 affect class size while they do not play any role in shaping the probability of observing a multigrade class. The specification that considers class size as endogenous—our preferred specification—yields an additional first stage of the following form:

$$\begin{aligned} CS_{i,g=2} = \tau_0 + \tau_1 \mathbb{1}[CohortSize_s < 10] + \tau_2 \mathbb{1}[10 \leq CohortSize_s < 15] \\ + \tau_3 \mathbb{1}[15 \leq CohortSize_s < 27] + \tau_4 \mathbb{1}[CohortSize_s \geq 27] \\ + \boldsymbol{X}'_{i,g=2} \boldsymbol{\tau_5} + \mu_{i,g=2} \end{aligned} \tag{3}$$

4.2 Identifying Assumptions

Two main assumptions need to be fulfilled to interpret the IV estimates for the multigrade effect on child development causally: first, we need the exclusion restriction to hold; and second, we need our instrumental variables to be relevant.

Exogeneity. The exclusion restriction implies that our instruments—the four indicator variables for cohort size—only affect students' achievement through the endogenous variables of the model, i.e. attendance of a multigrade class and class size. Although the exclusion restriction is empirically untestable, the process and timing of enrollment in Italian primary schools should reassure about its validity.

First, as parents are free to enroll their children in any primary school nationwide, it is difficult (if not impossible) to have sufficiently precise expectations of cohort sizes at time of enrollment. In other words, parents who intend to avoid a multigrade class for their children would need to forecast if the cohort size in a certain school in the next SY will be slightly above versus slightly below ten students.

Second, parents have to be willing to bear commuting costs in order to consider other schools than the closest one. These commuting costs might be sizable especially for parents living in remote or inner areas of the country. Moreover, due to the enrollment timing and process, the choice of a school in a different municipality would not guarantee a specific teaching practice (single- versus multigrade teaching) or class size.

Finally, it is important to highlight that parents apply to primary school by January–February each year. However, the SY starts in mid-September and families are notified about acceptance within a month after their application. On the contrary, students' (and teachers') assignment to classes only occurs during the summer. Therefore, parents, who are not involved in the assignment process only learn of their children's class composition shortly before the beginning of the SY.

Despite the impossibility to formally test the exclusion restriction, the analysis of discontinuities around the cut-offs defined by DPR 81/2009 is an interesting exercise to validate our approach. Indeed, in the presence of discontinuities around the cut-offs we would not be able to exclude the notion that principals or other subjects manipulate the enrollment process to favor the formation of certain types of classes. We investigate this aspect in Table 3 by comparing average characteristics of the students included in our sample around (+/- 2 students) the critical cut-offs identified by the law. As the table shows, with a few exceptions for the 27-student cut-off, there are no significant differences around the cut-offs. This evidence suggests that manipulation or other strategical behavior around the cut-off is unlikely to occur.

Relevance. We test instruments' relevance by analyzing first-stage estimates. Table 4 reports the first-stage estimates for two different specifications. The first specification (column 1) includes the endogenous variable MG and considers class size within the set of control variables. The second specification (columns 2 and 3) tackles the possible endogeneity of class size and considers both attendance of a multigrade class and class size as endogenous variables. All specifications in this and the following analyses include the control variables listed in Section 4.1.

The first column of the table confirms the highly predictive power of the instruments based on cohort size. With respect to the omitted category of schools with more than 26 second graders, students in schools with fewer than ten second graders are 85 percentage points more likely to end up in a multigrade class. As predicted by the law, the probability drops to a statistically significant positive effect of fifteen percentage points for those students in schools with 10–14 second graders. Students from schools with 15–26 or more than 26 second graders are equally likely to end up in a multigrade class.

The results do not substantially differ for the specification also considering class size as endogenous. Column (2) replicates the first-stage estimates for multigrading and displays very similar results as in column (1). The first stage for class size in column (3) shows that schools within the first two cut-offs—namely with fewer than fifteen second graders—are characterized by an average class size that is five students lower than those schools with fifteen or more students.

First-stage estimates highlight (at least) three important findings. First, the tests provided in the bottom part of the table confirm the instruments' relevance. Second, the two different specifications are very similar in terms of point estimates and precision. This similarity reassures the marginal impact on estimates when it comes to considering class size as an endogenous variable of the model versus as a simple control variable. Third, the four instrumental variables differently affect the two endogenous variables of the model. On the one hand, the first two cut-offs only play a role in determining a student's probability of ending up in a multigrade class. On the other hand, the two other variables only shape class size. These different effects of the instrumental variables enable estimating a specification with both multigrade and class size treated as endogenous variables and identified out of two different subsets of instruments.

5 Multigrading and Child Achievement Over Time

This section provides the estimates of the effect of multigrading on child achievement. Table 5 displays the ordinary least squares (OLS, columns 1 to 3) and the IV second-stage estimates (columns 4 to 9) of our model. In columns (4) to (6), the model includes class size as a control variable, while in columns (7) to (9) we also consider class size as endogenous. Each analysis is replicated for students' test scores in Grades 2, 5, and 8.[14]

The OLS estimates in column (1) suggest that attendance of a multigrade class in Grade 2 is associated with a significant improvement in the test score at the end of the same grade. Students in multigrade classes experience an increase by eight percent of a standard deviation in their Grade 2 test score with respect to second graders attending a single-grade class. This relative advantage disappears in Grade 5. Indeed, as shown in column (2), the test score in Grade 5 seems unaffected by attendance of a multigrade class three years earlier. Conversely, the multigrade effect is reverted when the test score in Grade 8 is considered as the outcome of interest (column 3). In this case, attendance of a multigrade class in Grade 2 is associated with a lower performance on the test by about seven percent of a standard deviation.

Columns (4) to (6) display IV results with attendance of a multigrade class considered as the endogenous variable of the model. In column (4), we analyze the short-term impact of attending a multigrade class in Grade 2 on the standardized test score in the same grade. The effect of multigrading is positive, statistically significant, and sizable. Attendance of a multigrade class causes an average increase in the test score by about thirteen percent of a standard deviation. The result is similar to that found by Barbetta et al. (2021). As for the OLS, the positive effect of multigrading fades away three years after students' attendance of a multigrade class. Column (5) displays a zero impact of multigrading on the test score at Grade 5. The point estimate for the multigrade effect turns negative in Grade 8, namely at least six years after the first attendance of a multigrade class and once students are in the lower-secondary cycle of education. At this stage of education, the effect of multigrade is statistically significant and accounts for a ten percent of a standard deviation decrease in the standardized test (column 6).

Our findings are robust to considering both attendance of a multigrade class and class size as endogenous variables. Indeed, the analysis performed in columns

[14] All the analyses consider the combined math-language test score as the outcome of interest. All results remain remarkably similar if single test scores, i.e. math and language, are considered in isolation. Math and language test scores are the only test scores available for all grades.

(7) to (9) is remarkably similar to that with a single (MG) endogenous variable. Again, the effect of multigrading is strongly positive in the short term and tends to vanish over time, becoming negative in Grade 8.

The analysis presented in this section calls for two important remarks. First, the choice among different empirical specifications plays a limited role in shaping the interpretation of the effect of multigrading on students' academic performance. From a quantitative viewpoint, IV estimates tend to be slightly larger than OLS estimates, therefore suggesting the importance of taking into proper account possible endogeneity concerns related with attendance of a multigrade class. At the same time, the two different IV specifications mimic each other, limiting possible concerns on the best way to isolate the effect of multigrading from that of class size. Second, the analysis sheds important light on the evolution over time of the effect of multigrading on students' academic performance. Students assigned to a multigrade class tend to benefit from this experience in the short term although the effect turns negative in the medium/long term. In the next section, we investigate the potential mechanisms underlying these results to show (i) the determinants of the effect of multigrading over time, and (ii) how to avoid the negative long-term effect of multigrading.

6 Mechanisms

Our main analysis suggests that the positive short-term impact of multigrading fades away over time. Six years after the first attendance of a multigrade class, the effect on children's cognitive development is on average negative. There are three potential mechanisms that can help to rationalize this finding: (i) the relative age of students in multigrade classes; (ii) effect heterogeneity by quality of multigrade teaching; and (iii) the selection of students into different lower-secondary schools after attendance of a multigrade class in primary school. This section investigates each of these mechanisms.

6.1 Relative Age and the Effect of Multigrading

Attending a multigrade class in all grades of the primary school cycle—defined as *persistence* of multigrading—might explain the pattern of the effect of multigrading over time displayed in the empirical analysis. If children ending up in a multigrade class at the beginning of primary school are highly likely to complete the whole primary school cycle in such a class, then the overall effect of multigrading might differ when estimated at different points in time. Indeed, children would switch from being part of the younger cohort in the earlier grades—i.e. Grade 1 and 2—to belonging to the older one in the classroom in Grade 5. This pattern makes the effect of multigrading observed in later stages of the educational career the average *cumulative* effect of multigrading, rather than its *pure* effect as measured in Grade 2.

The persistence of multigrading characterizes the Italian primary school system. In our data, about 72 percent of students who are in a multigrade class in Grade 2 are in the same type of class in Grade 5. As shown in Table 1, multigrading is very rare after primary school. Studies such as Barbetta et al. (2021) and Leuven and Rønning (2016) show that being among the younger cohort(s) in a multigrade class is beneficial to a child as she is exposed to more mature peers, a richer vocabulary, and a more demanding school curriculum.[15] The effect is reverted for the older cohorts in the classroom.

The analysis in Table 5 shows that students experience a positive multigrade impact on their academic performance in Grade 2 (contemporaneous effect). The beneficial impact of multigrading likely derives from sharing the classroom with older and more mature peers. This effect is counterbalanced by attendance of a multigrade class in the following SYs when students are exposed to younger peers and probably a less-demanding school curriculum adapted for a multi-age classroom. Therefore, the overall effect of multigrading in Grade 5 should be considered as the average cumulative effect accrued over the whole primary school cycle of education. Due to dynamic complementarities described in, e.g. Cunha et al. (2010) and Heckman and Mosso (2014), this average cumulative effect carries over in Grade 8, where students experiencing multigrade in all grades at primary school are likely to compare with students of the same age who did not attend a multigrade class in lower-secondary school.

[15]Despite sample size limitations, Barbetta et al. (2021) also show that the few students who were only attending a multigrade class in Grade 2 report a positive multigrade effect at the end of primary school (Grade 5).

Two important concluding remarks are due. First, our analysis on longitudinal data confirms some puzzling results in the existing literature—exclusively based on cross-sectional data—pointing to the importance of relative age in shaping the effect of multigrading. In further detail, the pattern that we find on the impact of multigrading rationalizes some opposite-in-sign effects in the literature, e.g. positive impacts for younger cohorts (Leuven and Rønning, 2016; Barbetta et al., 2021) versus negative impacts for older cohorts in Grade 5 (Checchi and De Paola, 2018). This evidence highlights the importance of a correct interpretation of the effects—pure versus cumulative—of multigrading in different circumstances. Second, our analysis leaves an important question unanswered: is it possible—independently of the persistence of multigrading—to adopt forms of multigrade teaching that do not harm long-term child development? The following analyses will try to answer this question.

6.2 Teachers and Teaching Practices

Given the evidence of a negative average long-term impact of multigrading on students' performance, it becomes crucial to understand whether it would be feasible to adopt such a practice without harming the long-term cognitive development of students. To find an answer, the focus moves to school inputs, which include both human resources—teachers and teaching practices—and capital, namely financial resources, equipment, and other school facilities. We start in this section by discussing the role of human resources.

Teachers are among the most important inputs in a child's development process (Alan et al., 2021b; Jackson et al., 2014; Chetty et al., 2014; Xu and Ran, 2020). The role of teachers becomes even more important in contexts of non-standard educational practices such as multigrading. In order to handle classes comprising students of different ages, teachers need to adopt flexible forms of teaching and devote considerable effort to playing their role as educators (INDIRE, 2019, 2020). This section aims to answer a very simple question: Does the effect of multigrading depend on the way in which teachers implement this education practice, or is it almost impossible to adopt multigrading practices without harming long-term child development?

The Italian Institute for Innovation in Schooling (INDIRE) provides qualitative evidence that teachers' turnover represents one of the main predictors of successful adoption of multigrade classes. INDIRE (2019) claims that: *"Teacher turnover has a huge impact: the school requires detailed and operational knowledge of the curriculum in order to operate well in a multigrade environment. If the teacher changes, the process starts from scratch every time."*

As argued by the INDIRE report, a high level of turnover has multiple possible disadvantages for child development. Indeed, on the teachers' side, a high turnover may imply that less experienced teachers face the demanding challenge of teaching a multi-age class. Moreover, a high turnover interrupts and jeopardizes the teachers' learning process on how to run a multigrade class effectively. Finally, teachers might feel less motivated and exert a lower effort if they perceive their role as temporary. On the students' side, children—especially those most in need of support by teachers—might become discouraged by a continuous changes of their instructors. All of these circumstances might reduce the potential of multigrade teaching in fostering child development.

We map the qualitative evidence by INDIRE (2019) into quantitative results by investigating the role of teachers' turnover in shaping the effectiveness of multigrading. for this purpose, we match our original data with an additional data set provided by the Ministry of Education and including information about teachers' characteristics at the school level. This new data set enables analyzing the possible heterogeneity of the effect of multigrading by the level of teachers' turnover in Grade 2.

Table 6 shows the effect of multigrading on child achievement by level of teachers' turnover.[16] To investigate effect heterogeneity, we have performed our baseline IV analysis—with multigrade and class size treated as endogenous variables—by subsamples of second-grade students. Subsamples are determined by the distribution of the share of temporary teachers at the school level. The *low-turnover* subsample comprises students attending schools whose share of teachers with temporary contracts is below the median of the distribution. The *high-turnover* subsample comprises students attending schools whose share of teachers with

[16] The average teacher turnover in the sample is about 11 percent.

temporary contracts is above the median. Columns (1), (3), and (5) report the estimates for the effect of multigrading on children's test scores in Grades 2, 5, and 8, respectively, in schools with low levels of teachers' turnover. Columns (2), (4), and (6) replicate the analysis by focusing on children in schools with high teachers' turnover.

The table conveys some important findings. First, teachers' turnover plays a role in shaping the effectiveness of multigrading. Independently of the grade, multigrading always shows a better impact on school performance when teachers' turnover is low. For instance, the positive effect of multigrading in Grade 2 is considerably larger (21 versus 14 percent of a standard deviation) for students in schools with more permanent teachers compared to students exposed to higher levels of teachers' turnovers.[17] Second, if we consider Grade 8, students who attended a multigrade class six years earlier in a school with more permanent teachers perform similarly to their peers attending a single-grade class in a low-turnover school. The same does not apply to students of schools with high levels of teachers' turnover, whereby in this case attending a multigrade class harms a student's long-term academic achievement (-16 percent of a standard deviation).

In Table 7, we test the sensitivity of our results to the criteria used to classify schools given their share of temporary teachers. Specifically, we now classify schools in tertiles of the distribution of the share of temporary teachers. The new results are similar to the previous ones. The long-term impact of multigrading is zero (column 1) if students attended a multigrade class in a school with a sizable share of permanent teachers. The effect becomes negative for students who attended Grade 2 in schools with higher teachers' turnover (columns 2 and 3).

Overall, the analysis in this section suggests that the quality of implementation of multigrade teaching matters. In particular, multigrading could be practiced without harming children's long-term development if teachers are sufficiently skilled and motivated. Schools with more experienced or motivated teachers are able to use multigrading effectively without harming students' cognitive development. On the contrary, the use of multigrading by teachers with a lower level

[17] For the latter group, despite sizable, the point estimate is imprecise and statistically non-significant.

of experience and motivation—proxied here by those more subject to frequent turnover—can cause negative long-term impacts on child development. In other words, our results support the view in INDIRE (2019): *"If the expertise remains sporadic and linked to teachers who do not transfer it and are sensitive to turnover, the school is unable to sustain the processes of innovation and change. Only 6 out of 19 institutes display an adequate technological experience and a group of teachers are able to implement innovative forms of teaching with the lack of a stable teaching staff that makes it difficult to carry out multigrade teaching."*

Admittedly, our analysis cannot rule out the notion that other measures for school quality—e.g. low turnover in high-quality schools—are the driver of the heterogeneous effects. This threat does not undermine the main conclusion of this analysis, namely the fact that it is possible to use multigrade teaching without producing long-term negative effects on child cognitive development. If anything, it only highlights one of the aspects of school quality with the potential to make multigrade teaching an effective educational practice.

6.3 Multigrading and (Future) School Resources

The Italian Constitution states that education is an exclusive responsibility of the central government, which defines the allocation and the remuneration of school teachers at the national level and covers their costs. However, for both primary and lower-secondary schools, municipalities have the responsibility to find an appropriate building for hosting the school, bear the costs of maintenance, as well as the cost of some connected services, e.g. a dining hall or school buses. Despite the redistribution pursued at the national level, municipalities differ in terms of available resources, and their choices relative to school buildings and services are also different. Moreover, municipalities in rural and remote areas with few inhabitants are generally poorer than less peripheral municipalities. If students attend a multigrade class at the beginning of their educational career and then continue in "poor" schools—namely schools with low educational, financial, and infrastructural resources—this would confound the interpretation of the long-term effect of multigrading. In this section, we study whether lower-secondary schools resources for students previously attending a multigrade class can explain the effect of multigrading and its pattern over time. In other words, we test whether the effect of multigrading in Grade 8 is heterogeneous based on the quality of the school attended after multigrading.

As a first step, we merge our data with a new data set elaborated by SOSE, a public company owned by the Italian Ministry of the Economy and Finance and the Bank of Italy that collects an extensive set of information on Italian municipalities. This information is used by SOSE to compute standard needs across municipalities. SOSE also elaborates a benchmarking tool called OpenCivitas, which classifies each Italian municipality according to the level and quality of services provided to the population and the level of spending.[18]

OpenCivitas assigns each municipality to a *high-* versus a *low-level* group of municipalities according to the relative position of the municipality with respect to the median of the distribution of services offered in the whole country. The same is replicated for the case of spending. We use this classification to test whether the effect of multigrading changes with the resources of the municipality hosting the school proxied by both the level of services and the level of expenditure.

Table 8 reports the effect of attending a multigrade in Grade 2 on the standardized test score in Grade 8 by municipality resources as measured in the SY 2015/2016. Column (1) reports the analysis for municipalities with lower (or below-the-median) levels of services. Column (2) analyzes the sample with higher levels of services. Column (3) considers the case of lower-than-median expenditures, while column (4) only includes those municipalities with expenditure above the median of Italian municipalities. As for previous analyses, we estimate IV specifications treating both attendance of a multigrade class and class size as endogenous variables.

The table depicts a consistent picture. The restriction to different subsamples based on municipalities' resources does not shape any clear heterogeneity in the multigrade effect. All point estimates are remarkably similar and mimic the point

[18]Since we are considering inputs related with financial resources—which vary slowly over time—we use data relative to the SY 2015/2016 in which students in our sample attended Grade 5. This choice allows us to proxy school resources available when students started attending lower-secondary education.

estimates obtained in the whole sample. If anything, the use of subsamples implies an important loss in precision that makes point estimates statistically non-significant (or weakly significant).

As a further step, we investigate the heterogeneous effect of multigrading by considering more specific definitions of lower-secondary schools' resources and facilities. To perform this analysis, we use an additional source of data by the Ministry of Education that includes information on school resources—such as the availability of rooms with personal computers, a pool/gym or a dining hall—for the whole population of Italian schools. We refer to SY 2018/19, when students were in Grade 8.

We start by considering the share of rooms with a personal computer, which can be considered as a proxy for the digital resources available to schools. We classify schools in three levels, namely *Low*, *Medium*, and *High*, based on the tertiles of the distribution of the share of rooms with a computer in Italian schools. Table 9 shows that the effect of multigrading is independent of computers' availability. In fact, the effect of multigrading on students' achievement is similar—regardless of computers and other information technology (IT) resources available in Grade 8—and aligned with the baseline effects found for the whole sample. Again, the use of subsamples lowers the precision of our estimates.

We then consider the availability of an auditorium, a dining hall, and a sports facility (a gym and/or swimming pool accessible to students) in each school. These three facilities can be considered as proxies of the school environment, and their availability should make the environment livelier and improve students' experience while at school. We consider the auditorium and the sports facility to account for the availability of extra-curricular activities, while the dining hall is a proxy for a friendlier school environment and the possibility that students also remain at school for meals or in the afternoon.[19]

Table 10 reports the analysis of the effect of multigrading on test scores in Grade 8 by school environment. Columns (1) and (2) shows results based on availability of an auditorium, in columns (3) and (4) the case of a dining hall is considered,

[19] In Italy, most of lower-secondary school students only attend classes in the morning and return to their families before lunch.

while the availability of a sports facility is the source of heterogeneity in columns (5) and (6). For each facility, the odd-numbered column refers to estimates when the resource is unavailable, and the even-numbered column when it is available.

The estimates in the table display limited heterogeneity based on school resources. The availability of an auditorium or dining hall displays negative effects for all subgroups. On the one hand, the effect appears larger and statistically significant only for schools without an auditorium or dining hall. On the other hand, the effect for schools where these resources are available is negative and aligned with the point estimate in the baseline specification. The small sample size makes the estimates imprecise and therefore statistically non-significant. The case of the sports facility is slightly different as schools with gyms and/or pools display a larger negative effect of multigrading. The effect is non-significant for those schools without these facilities. A likely explanation is that availability of sports facilities is a proxy for a less peripheral location of the municipalities hosting these schools. In these contexts, students who have experienced multigrading during their primary school career can find more difficulties compared with students attending schools in less peripheral urban centers.

Overall, the analysis of municipality spending and school resources does not provide any clear-cut evidence of heterogeneous effects of multigrading. The lack of heterogeneous impacts seems to rule out the possibility that attending a multigrade class shapes the students' probability of ending up in a lower-quality school with fewer financial and non-financial resources. The analysis reinforces the idea that elements more strongly related to teachers' characteristics and the quality of implementation of the educational practice are responsible for the medium- and long-term effects of multigrading. Finally, it is important to recall that despite being measured in Grade 8, for many students in our sample the measure for school resources would have been the same in Grade 2. Indeed, as our sample comprises municipalities with a unique primary school, students usually undertake their primary (Grades 1 to 5) and lower-secondary (Grades 6 to 8) cycles of education in the same school. This means that our results also speak to the role of school financial and non-financial resources in the year in which a student enrolls in a multigrade class. Again, these resources seem marginal in shaping the effect of multigrading on students' achievements.

7 Conclusions

The practice of multigrading—mixing children of different ages in the same class—is widely used worldwide both to comply with resource constraints and for educational and pedagogical reasons. Despite its use, there remains no consensus on the role of this practice in shaping child development. The few causal studies on multigrading—mostly based on cross-sectional data—draw a mixed picture with some works highlighting a positive effect on students' achievements and others showing a detrimental effect of this practice.

This study reconciles contrasting findings in the literature by building on a new longitudinal data set on child cognitive achievements for Italian primary and lower-secondary school students. We leverage institutional rules on class formation to deal with the potential endogeneity underlying the attendance of a multigrade class at the beginning of the students' school career. These rules allow us also to isolate the effect of multigrading from the effect of class size.

Our findings show that multigrading in Grade 2 of primary school has beneficial short-term effects on students' cognitive achievements. However, this positive impact fades away over time and becomes negative on average in the long-term. As explained in previous studies such as Barbetta et al. (2021), the persistence of multigrading in the primary cycle of education—together with relative age within a multigrade class—explains why the effect fades away over time. In other words, ceteris paribus, the analysis of multigrading at different stages of the educational career might display different results as the treatment effect might derive from attending a multigrade class for a single versus multiple SYs. In early stages of education, the effect of multigrading should be considered as the pure effect of this educational practice. In the longer run, the effect needs to be interpreted as the cumulative effect of several years of multigrading. The switch from being part of the younger cohort(s) in the classroom (early grades) to belonging to the older cohort(s) shapes the attenuation of the positive initial impact of multigrading.

Our results suggest an important policy-relevant question: Is it possible to avoid the negative effect of multigrading in the long run? This question is even more important for those areas where multigrading represents the only possibility for a local school to operate. Investigating the mechanisms, we show that the quality of implementation of multigrade teaching is the key factor. Indeed, we show that in a context with high persistence of multigrade teaching, the negative long-term effect of multigrading only appears for those students who attended a multi-grade class in a school with high shares of teachers with temporary contracts. These teachers are likely less motivated and have lower experience with multi-grade teaching. On the contrary, multigrading is not associated with the choice of future lower-quality schools. Therefore, it seems unlikely that the long-term impact of multigrading is driven by factors such as the quality of schools in lower-secondary education.

Overall, this new evidence obtained through longitudinal data and a neat identification strategy suggests that multigrade teaching does not necessarily imply negative effects on medium- or long-term child development. Teachers appear as the most important actor when it comes to guaranteeing the effective provision of multigrading. As reinforced by anecdotal evidence and teachers' interviews, multigrade teaching requires particular skills, effort, and motivations to be effectively run. Low-motivated or low-experienced teachers face too many challenges to deal effectively with students from different grades at the same time.

Summing up, multigrading is an important resource, but it needs to be carefully implemented to avoid harming child development. Admittedly, our study neglects some potentially important dimensions of the impact of multigrading on a child's development. Due to data limitations, we only focus on students' cognitive development. However, child development is multidimensional and non-cognitive, behavioral, and socio-emotional skills are usually deemed as fundamental predictors of future life opportunities (Heckman and Mosso, 2014). Moreover, these soft skills are malleable (Alan et al., 2019, 2021a; Kosse et al., 2020; Sorrenti et al., 2020). Multigrading might play an important role in shaping these skills, e.g. by exposing children to more mature peers or fostering the sense of responsibility toward younger peers by older peers in the classroom. The analysis of the impact of multigrade teaching on a broad and multidimensional set of skills should be prioritized by future research on the topic.

References

S. Alan, T. Boneva, and S. Ertac. Ever Failed, Try Again, Succeed Better: Results from a Randomized Educational Intervention on Grit. *The Quarterly Journal of Economics*, 134(3):1121–1162, 2019.

S. Alan, C. Baysan, M. Gumren, and E. Kubilay. Building Social Cohesion in Ethnically Mixed Schools: An Intervention on Perspective Taking. *The Quarterly Journal of Economics*, 136(4):2147–2194, 2021a.

S. Alan, E. Duysak, E. Kubilay, and I. Mumcu. Social Exclusion and Ethnic Segregation in Schools: The Role of Teacher's Ethnic Prejudice. *The Review of Economics and Statistics*, 2021b.

J. D. Angrist and V. Lavy. Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement. *The Quarterly Journal of Economics*, 114(2): 533–575, 1999.

G. P. Barbetta, G. Sorrenti, and G. Turati. Multigrading and child achievement. *Journal of Human Resources*, 56(3):940–968, 2021.

D. Checchi and M. De Paola. The Effect of Multigrade Classes on Cognitive and Non-cognitive Skills. Causal Evidence Exploiting Minimum Class Size Rules in Italy. *Economics of Education Review*, 67:235–253, 2018.

R. Chetty, J. N. Friedman, and J. E. Rockoff. Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates. *American Economic Review*, 104(9):2593–2632, 2014.

M. Chingos. Class Size and Student Outcomes: Research and Policy Implications. *Journal of Policy Analysis and Management*, 32:411–438, 2013.

F. Cunha, J. J. Heckman, and S. M. Schennach. Estimating the technology of cognitive and noncognitive skill formation. *Econometrica*, 78(3):883–931, 2010.

J. J. Heckman and S. Mosso. The economics of human development and social mobility. *Annual Review of Economics*, 6(1):689–733, 2014.

INDIRE. Tracce di sostenibilità. Comunità, Pluriclassi e Tecnologie nella Pratica della Scuola. *I Quaderni delle Piccole Scuole*, 1/2019, 2019.

INDIRE. Apprendimento Intervallato per Competenze. *I Quaderni delle Piccole Scuole*, 7/2020, 2020.

C. K. Jackson, J. E. Rockoff, and D. O. Staiger. Teacher effects and teacher-related policies. *Annual Review of Economics*, 6(1):801–825, 2014.

F. Kosse, T. Deckers, P. Pinger, H. Schildberg-Hörisch, and A. Falk. The Formation of Prosociality: Causal Evidence on the Role of Social Environment. *Journal of Political Economy*, 128(2):434–467, 2020.

E. Leuven and M. Rønning. Classroom grade composition and pupil achievement. *The Economic Journal*, 126(593):1164–1192, 2016.

A. W. Little. Multigrade Teaching: Towards an International Research and Policy Agenda. *International Journal of Educational Development*, 21(6):481–497, 2001.

D. Sims. A Strategic Response to Class Size Reduction: Combination Classes and Student Schievement in California. *Journal of Policy Analysis and Management*, 27:457–478, 2008.

G. Sorrenti, U. Zölitz, D. Ribeaud, and M. Eisner. The Causal Impact of Socio-Emotional Skills Training on Educational Success. *CESifo Working Paper*, 2020.

UNESCO. Educating rural people: A low priority. *The Newsletter of UNESCO Education Sector*, 9, 2004.

D. Xu and F. X. Ran. The Impacts of Different Types of College Instructors on Students' Academic and Labor Market Outcomes. *Journal of Policy Analysis and Management*, 40(1):225–257, 2020.

Table 1: Summary Statistics

	Mean	St.Dev	Min	Max
Combined Test Score Grade 2	0.09	0.96	-4.86	4.002
Combined Test Score Grade 5	0.08	0.96	-4.81	4.244
Combined Test Score Grade 8	0.08	0.98	-4.29	3.590
Multigrade (MG) Grade 2	0.06	0.23	0	1
Years in MG Class	3.64	1.56	1	5
Class Size Grade 2	19.39	4.11	5	30
Class Size Grade 5	19.06	4.07	4	29
Female	0.50	0.50	0	1
Italian	0.92	0.28	0	1
Immigrant 1st Generation	0.02	0.14	0	1
Immigrant 2nd Generation	0.06	0.24	0	1
Father University	0.07	0.26	0	1
Mother University	0.11	0.31	0	1
Nord West	0.47	0.50	0	1
Nord East	0.17	0.38	0	1
Center	0.11	0.31	0	1
South	0.18	0.39	0	1
Islands	0.06	0.25	0	1
N	70,688			

Notes: This table shows summary statistics for the sample analyzed in this paper.

Table 2: Comparison between Multigrade and Single-Grade Classes

	Multigrade Class		Single-Grade Class		Difference	
	Mean	St.Dev.	Mean	St.Dev.	Δ	t
Combined Test Score Grade 2	0.25	1.03	0.08	0.95	0.17^{***}	(10.18)
Combined Test Score Grade 5	0.12	1.01	0.08	0.96	0.04^{**}	(2.54)
Combined Test Score Grade 8	-0.04	0.98	0.09	0.98	-0.13^{***}	(-8.15)
Years in MG Class	4.22	1.19	1.68	0.94	2.54^{***}	(65.17)
Class Size Grade 2	13.88	2.65	19.73	3.94	-5.85^{***}	(-131.58)
Class Size Grade 5	13.07	3.17	19.42	3.83	-6.35^{***}	(-122.16)
Female	0.48	0.50	0.50	0.50	-0.02^{**}	(-2.09)
Italian	0.93	0.25	0.91	0.28	0.02^{***}	(4.33)
Immigrant 1st Generation	0.03	0.16	0.02	0.14	0.01^{**}	(2.36)
Immigrant 2nd Generation	0.04	0.20	0.07	0.25	-0.02^{***}	(-7.25)
Father University	0.05	0.22	0.08	0.26	-0.03^{***}	(-7.54)
Mother University	0.08	0.28	0.11	0.31	-0.02^{***}	(-5.50)
Nord West	0.44	0.50	0.47	0.50	-0.03^{***}	(-4.20)
Nord East	0.08	0.28	0.18	0.38	-0.09^{***}	(-20.54)
Center	0.12	0.32	0.11	0.31	0.01	(1.52)
South	0.29	0.45	0.18	0.38	0.11^{***}	(15.15)
Islands	0.07	0.26	0.06	0.24	0.01^{**}	(2.32)
N	4,033		66,655		70,688	

Notes: This table compares the group of students in multigrade versus single-grade classes. The last two columns provide the difference and the t-statistic for the difference.
Significance levels: $^{*}\ p < 0.1$, $^{**}\ p < 0.05$, $^{***}\ p < 0.01$.

Table 3: Balancing Test for Second-Grade Students

	Below (-2)	On and Above (+2)	Diff	p-Value
	(1)	(2)	(3)	(4)
	Cut-Off: Cohort-Size 10 Students			
Age	7.41	7.40	0.01	(0.40)
Female	0.49	0.50	-0.01	(0.56)
Italian	0.94	0.93	0.01	(0.40)
First Gen Immigrant	0.02	0.02	-0.00	(0.82)
Second Gen Immigrant	0.04	0.05	-0.01	(0.39)
Father University	0.06	0.05	0.01	(0.33)
Mother University	0.08	0.09	-0.00	(0.74)
	Cut-Off: Cohort-Size 15 Students			
Age	7.40	7.42	-0.02	(0.09)
Female	0.48	0.51	-0.03	(0.06)
Italian	0.93	0.93	0.00	(0.93)
First Gen Immigrant	0.02	0.02	-0.01	(0.15)
Second Gen Immigrant	0.06	0.05	0.01	(0.40)
Father University	0.06	0.06	0.00	(0.76)
Mother University	0.10	0.11	-0.01	(0.51)
	Cut-Off: Cohort-Size 27 Students			
Age	7.41	7.40	0.01	(0.54)
Female	0.50	0.48	0.02	(0.27)
Italian	0.93	0.92	0.01	(0.08)
First Gen Immigrant	0.02	0.02	0.00	(0.37)
Second Gen Immigrant	0.04	0.06	-0.02**	(0.01)
Father University	0.06	0.08	-0.02*	(0.05)
Mother University	0.10	0.11	-0.01	(0.24)

Notes: This table compares observable characteristics around the cut-off points of second-grade (enrolled) students identified by DPR 81/2009. See text for details on these cut-offs. The comparison is performed around the 10-, 15-, and 27-students cut-offs. For each cut-off, interval widths are defined by two students below versus above the cut-off. Columns (3) and (4) report the difference in means between students "just below" and "just above" the cut-off and the relative p-value. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 4: First-Stage Estimates: Determinants of Multigrade Attendance (and Class Size)

Dependent Variable:	Multigrade Attendance		Class Size (CS)
	(1)	(2)	(3)
CohortSize $\leq$ 10	0.851***	0.887***	-5.164***
	(0.016)	(0.015)	(0.213)
11 $\leq$ CohortSize $\leq$ 14	0.147***	0.185***	-5.342***
	(0.016)	(0.018)	(0.181)
15 $\leq$ CohortSize $\leq$ 26	-0.001	-0.008***	1.074***
	(0.002)	(0.002)	(0.183)
Class Size (Grade 2)		0.007***	
		(0.001)	
Model	CS Exogeneous	CS Endogenous	CS Endogenous
Individual Controls	Yes	Yes	Yes
Parent Controls	Yes	Yes	Yes
Regional FE	Yes	Yes	Yes
F I[CohortSize $\leq$ 10]	> 100	> 100	> 100
F I[11 $\leq$ CohortSize $\leq$ 14]	81.37	> 100	> 100
F I[15 $\leq$ CohortSize $\leq$ 26]	0.75	11.31	34.35
N	70,688	70,688	70,688

Notes: This table shows the first-stage estimates of the model. In columns (1) and (2), the dependent variable is an indicator variable for attendance of a multigrade class in Grade 2 of primary school. In column (3), the dependent variable is class size in Grade 2 of primary school. The model in column (1) considers class size as exogenous and includes it a standard control variable in the model. The model in columns (2) and (3) considers both multigrade attendance and class size as endogenous variables of the model. See text for further details. The reference category for the number of second-grade students is $I[CohortSize \geq 27]$. All models include controls for the child's gender, age, nationality, father's and mother's education and profession, as well as the population and altitude of the municipality. Standard errors are clustered at the school level and reported in brackets. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 5: IV Estimates: Multigrading and Educational Performance over Time

	OLS			IV: CS Exogenous			IV: CS Endogenous		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Dependent Variable:	Score 2	Score 5	Score 8	Score 2	Score 5	Score 8	Score 2	Score 5	Score 8
Multigrading Grade 2	0.080**	-0.005	-0.067***	0.133***	0.007	-0.101***	0.159***	-0.019	-0.094**
	(0.036)	(0.032)	(0.021)	(0.045)	(0.041)	(0.026)	(0.057)	(0.053)	(0.034)
Class Size Grade 2	-0.010***	-0.005	0.000	-0.009***	-0.005	-0.000	-0.005	-0.008	0.001
	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)	(0.002)	(0.006)	(0.006)	(0.004)
Individual Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Parent Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Regional FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	70,688	70,688	70,688	70,688	70,688	70,688	70,688	70,688	70,688

Notes: This table shows the OLS and second-stage estimates for the effect of multigrading on students' test scores. The dependent variable is the student's test score in Grade 2 (columns 1,4, and 7), Grade 5 (columns 2, 5, and 8), and Grade 8 (columns 3, 6, and 9). Columns (1) to (3) report OLS estimates. Columns (4) to (6) report IV estimates with only multigrading treated as endogenous variable and class size included as a standard control variable. Columns (7) to (9) report IV estimates with both multigrading and class size treated as endogenous variables. See text for further details. All models include controls for the child's gender, age, nationality, father's and mother's education and profession, as well as the population and altitude of the municipality. Standard errors are clustered at the school level and reported in brackets. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 6: Heterogeneity in the Effect of Multigrading by Share of Temporary Teachers

Dependent Variable:	Score 2		Score 5		Score 8	
	(1)	(2)	(3)	(4)	(5)	(6)
Multigrading Grade 2	0.220**	0.114	0.005	-0.062	-0.047	-0.162***
	(0.089)	(0.080)	(0.081)	(0.078)	(0.049)	(0.049)
Class Size Grade 2	-0.006	-0.004	-0.004	-0.013	0.007	-0.003
	(0.009)	(0.009)	(0.008)	(0.009)	(0.005)	(0.005)
Individual Controls	Yes	Yes	Yes	Yes	Yes	Yes
Parent Controls	Yes	Yes	Yes	Yes	Yes	Yes
Regional FE	Yes	Yes	Yes	Yes	Yes	Yes
Temporary Teacher Share	Low	High	Low	High	Low	High
N	32,972	35,021	32,972	35,021	32,972	35,021

Notes: This table shows whether the multigrade effect is heterogeneous by the share of temporary teachers in the school. The dependent variable is the student's test score in Grade 2 (columns 1 and 2), Grade 5 (columns 3 and 4), and Grade 8 (columns 5 and 6). The sample is split according to the share of temporary teacher in the school during Grade 2 of primary school. *Low* indicates that the primary school has below-the-median temporary teachers. *High* indicates that the primary school has above-the-median temporary teachers. The table report IV estimates with both multigrading and class size treated as endogenous variables. See text for further details. All models include controls for the child's gender, age, nationality, father's and mother's education and profession, as well as the population and altitude of the municipality. Standard errors are clustered at the school level and reported in brackets.
Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 7: Alternative Classification for the Share of Temporary Teachers

Dependent Variable:	Score 8		
	(1)	(2)	(3)
Multigrading Grade 2	-0.013	-0.127**	-0.177***
	(0.066)	(0.060)	(0.055)
Class Size Grade 2	0.007	-0.000	-0.001
	(0.006)	(0.007)	(0.006)
Individual Controls	Yes	Yes	Yes
Parent Controls	Yes	Yes	Yes
Regional FE	Yes	Yes	Yes
Temporary Teacher Share	Low	Medium	High
N	20,337	25,460	22,196

Notes: This table tests the robustness of heterogeneous effects by share of temporary teachers in the school. The dependent variable is the student's test score in Grade 8. The analysis mimics the one in Table 6. *Low* indicates that the primary school belongs to the first (or bottom) tertile of the distribution of schools according to their share of temporary teachers. *Medium* indicates that the primary school belongs to the second tertile of the distribution of schools according to their share of temporary teachers. *High* indicates that the primary school belongs to the third (or top) tertile of the distribution of schools according to their share of temporary teachers. All models include controls for the child's gender, age, nationality, father's and mother's education and profession, as well as the population and altitude of the municipality. Standard errors are clustered at the school level and reported in brackets.
Significance levels: * $p < 0.01$, ** $p < 0.05$, *** $p < 0.01$.

Table 8: Heterogeneity in the Effect of Multigrading by Financial Resources of the Municipality

	Score 8			
	(1)	(2)	(3)	(4)
Multigrading Grade 2	-0.0942*	-0.0933	-0.0984**	-0.0834
	(0.0544)	(0.0581)	(0.0501)	(0.0533)
Class Size Grade 2	-0.00312	0.00842	0.00428	0.00112
	(0.00577)	(0.00625)	(0.00514)	(0.00530)
Individual Controls	Yes	Yes	Yes	Yes
Parent Controls	Yes	Yes	Yes	Yes
Regional FE	Yes	Yes	Yes	Yes
Subgroup	Less Service	More Service	Less Expenditures	More Expenditures
N	30,894	26,091	37,492	25,814

Notes: This table shows the effect of multigrading on students' test scores by municipality resources of the municipality where students attend Grade 8. The dependent variable is the student's test score in Grade 8. The sample is split according to the level of services offered (columns 1 and 2) and expenditures (columns 3 and 4) by the municipality where the student attends Grade 8. Less indicates the municipality offer services or report expenditures below the sample median. High indicates the municipality offer services or report expenditures above the sample median. The table reports IV estimates with both multigrading and class size treated as endogenous variables. See text for further details. All models include controls for the child's gender, age, nationality, father's and mother's education and profession, as well as the population and altitude of the municipality. Standard errors are clustered at the school level and reported in brackets. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 9: Heterogeneity in the Effect of Multigrading by Availability of PC

	Score 8		
	(1)	(2)	(3)
Multigrading Grade 2	-0.111*	-0.0798	-0.112
	(0.0620)	(0.0485)	(0.0748)
Class Size Grade 2	0.00163	0.00111	0.00579
	(0.00580)	(0.00514)	(0.00897)
Individual Controls	Yes	Yes	Yes
Parent Controls	Yes	Yes	Yes
Regional FE	Yes	Yes	Yes
Share of Rooms with Computer	Low	Medium	High
N	23,586	29,665	15,741

Notes: This table shows the effect of multigrading on students' test scores by the share of rooms with personal computers in the school where students attend Grade 8. The dependent variable is the student's test score in Grade 8. The sample is split according to the availability of personal computers and other IT services available in the school where the student attends Grade 8. *Low* indicates that the school belongs to the first (or bottom) tertile of the distribution of schools according to their availability of personal computers and other IT services. *Medium* indicates that the school belongs to the second tertile of the distribution of schools according to their availability of personal computers and other IT services. *High* indicates that the school belongs to the third (or top) tertile of the distribution of schools according to their availability of personal computers and other IT services. The table reports IV estimates with both multigrading and class size treated as endogenous variables. See text for further details. All models include controls for the child's gender, age, nationality, father's and mother's education and profession, as well as the population and altitude of the municipality. Standard errors are clustered at the school level and reported in brackets. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 10: Heterogeneity in the Effect of Multigrading by School Environment

	Score 8					
	(1)	(2)	(3)	(4)	(5)	(6)
Facility:	Auditorium		Dining Hall		Pool	
	No	Yes	No	Yes	No	Yes
Multigrading Grade 2	-0.207***	-0.0775	-0.193***	-0.137	-0.0897	-0.248***
	(0.0650)	(0.0737)	(0.0603)	(0.0883)	(0.0836)	(0.0577)
Class Size Grade 2	0.00401	0.00428	0.00263	0.00372	0.00909	-0.00208
	(0.00572)	(0.00880)	(0.00525)	(0.0103)	(0.00730)	(0.00592)
Individual Controls	Yes	Yes	Yes	Yes	Yes	Yes
Parent Controls	Yes	Yes	Yes	Yes	Yes	Yes
N	33,822	17,426	38,982	12,266	24,168	27,080

Notes: This table shows the effect of multigrading on students' test scores by availability of certain facilities in the school where students attend Grade 8. The dependent variable is the student's test score in Grade 8. The sample is split according to the availability of an auditorium (columns 1 and 2), a dining hall (columns 3 and 4), and a pool (columns 5 and 6). The table reports IV estimates with both multigrading and class size treated as endogenous variables. See text for further details. All models include controls for the child's gender, age, nationality, father's and mother's education and profession, as well as the population and altitude of the municipality. Standard errors are clustered at the school level and reported in brackets. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

A Data Construction Process

In this section, we describe the process we used to identify: (a) students attending multigrade classes, and (b) the grade composition of multigrade classes.

A.1 Students in Multigrade Classes

Before SY 2018/2019, the INVALSI data set did not contain information concerning which type of class a student attended, so students enrolled in a multigrade class cannot be directly identified. Therefore, to obtain this information, we merged three administrative archives.

The first data set (the INVALSI data from now on) is obtained merging information about children's performance on the INVALSI second-grade test in SY 2012/2013, the fifth-grade test in SY 2015/2016 and the eighth-grade test in SY 2018/2019. For each student, the test scores in both mathematics and language as well as background information such as gender, age, nationality, attendance at preparatory schools, and parents' education and profession are available. Neither school names nor school characteristics and location are available in the INVALSI data set. However, each individual record—identified by a unique INVALSI-code—also includes a class and school code, as well as geographical and demographic information about the municipality where the student's school is located. This piece of information is fundamental for our matching procedure and includes: (i) the province where the school is located, (ii) the population (in the 2001 and 2011 census) of the municipality, (iii) the size (in square km) of the municipality, and (iv) the altitude of the municipality where the school is located.

A second administrative data set (School Register data from now onwards) provided by the Italian Ministry of Education (MIUR) contains detailed information about the characteristics of each Italian primary school in SYs 2011/2012 through 2015/2016. All of the Italian regions are covered in this data except for Valle d'Aosta and Trentino Alto Adige. The School Register includes information such as the school name, municipality, number of students (total and in each grade), number of classes (total and in each grade), and number of multigrade classes. Based on this information, we analyzed all possible combinations of grade composition at the school level to identify different types of schools. For example, if a school shows a positive number of second-grade students, but no second-grade single-grade classes and at least one multigrade class, we can assume that second-grade students attend a multigrade class. We ended up with: (i) schools where students in the grades of our interest attend a multigrade class; (ii) schools where students attend one single-grade class; (iii) schools where more than one single-grade class is offered; and (iv) schools with no students.

Note that we found no evidence of primary schools with both single and multigrade classes for the same grade.

Unfortunately, the INVALSI data and the School Register data cannot be matched directly. In fact, the first data set only identifies each primary school with an anonymous code. The only way to overcome this problem is to identify (at least) the names of the municipalities where the schools included in the INVALSI data set are located. Once identified, it would be possible to match the data set with the School Register, with the municipality as the matching variable.

The Municipality Register data set provided by ISTAT (National Institute of Statistics) is the last piece of information needed to complete the data construction process. The Municipality Register contains geographical and demographic information for each Italian municipality. This information (province, population in the 2001 and 2011 census, size and altitude of the municipality) is the same as that contained in the INVALSI data, therefore making the merger of the INVALSI data set with the Municipality Register data set possible. We use geographical and demographic information as key identifying variables in the matching process to obtain the INVALSI+ISTAT data. Finally, we match the INVALSI+ISTAT data with the School Register data based on municipality names. With this last matching, we can uniquely identify only schools located in municipalities hosting no more than one school. We repeated the same procedure for SY 2012/2013 and 2015/2016 to obtain the grade composition of the classes of each student attending a school in a municipality hosting only one school in both years.

A.2 Classroom Grade Composition of Multigrade Classes

As mentioned in the paper, no data identify the classroom grade composition of multigrade classes. We use the data built in the previous paragraph and apply a wide set of rules to identify the grade composition of multigrade classes. These rules are based on the information originally included in the School Register. For example, we define the following Rule 1 to identify a multigrade class whose students are first and second graders only (therefore, second graders are the older

peers in the multigrade class). According to Rule 1, the school has:
(i) one multigrade class;
(ii) no first- and second-grade single classes;
(iii) first- and second-grade students;
(iv) third-, fourth-, and fifth-grade single classes;
(v) third-, fourth-, and fifth-grade students.

We consider about 40 of such rules to enumerate all of the possible combinations of students of different grades and describe the classes in our data.