Multigrading and Child Achievement

ONLINE APPENDIX

Gian Paolo Barbetta* Giuseppe Sorrenti† Gilberto Turati‡

August 2019

*Università Cattolica del Sacro Cuore, Milan (IT). E-mail: gianpaolo.barbetta@unicatt.it

†University of Zürich and Jacobs Center for Productive Youth Development, Zurich (CH). E-mail: giuseppe.sorrenti@econ.uzh.ch

‡Università Cattolica del Sacro Cuore, Rome (IT). E-mail: gilberto.turati@unicatt.it

I. Data Construction Process

In this section we describe the process we used to: (a) identify students attending multigrade classes, (b) identify the classroom grade composition of multigrade classes, and (c) track the career path of fifth-grade students.

A. Students in Multigrade Classes

The INVALSI data do not contain information concerning which type of class a student attended, so students enrolled in a multigrade class cannot be directly identified. To obtain this information, we merged three administrative archives.

The first data set (the INVALSI data from now on) contains information about children's performance on the INVALSI test in school year (SY) 2012/2013. For each student, the test score 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 this data set. However, each individual record also includes a class and a 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 the 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 on) provided by the Italian Ministry of Education (MIUR) contains detailed information about the characteristics of each Italian primary school in 2012/2013 SY as well as the five previous years. All the Italian regions are covered in this data except for Valle d'Aosta and Trentino Alto Adige. The School Register includes information such as 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 of the 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 second-grade students attend a multigrade class; (ii) schools where second-grade students attend one second-grade single-grade class; (iii) schools where more than one second-grade single-grade class is offered; and (iv) schools with no second-grade 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 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. Last, we match the INVALSI+ISTAT data with the School Register data based on municipality names. With this last matching, we are able to uniquely identify only schools located in municipalities hosting no more than one school. We repeated the same procedure to obtain the data for fifth-grade students.

B. 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 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:

- one multigrade class;
- no first- and second-grade single classes;
- first- and second-grade students;
- third-, fourth-, and fifth-grade single classes;
- third-, fourth-, and fifth-grade students.

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

C. School Career Path of Fifth-Grade Students in SY 2012/2013

To interpret the effect of multigrading on fifth-graders, we tracked back their school career path. We repeated the procedure described in Section I.B for fourth-grade students in SY 2011/2012, third-grade students in SY 2010/2011, second-grade students in SY 2009/2010, and first-grade students in SY 2008/2009. This procedure allowed us to determine how many years (and which years) the students attending fifth-grade in SY 2012/2013 spent in a multigrade or in a single-grade class (assuming that they did not change schools during their primary education).

II. REDUCED-FORM ESTIMATES

Table A.1 provides the reduced-form estimates of our model.

Table A.1: Reduced-Form Estimates

	Combined Math-Language			
	OLS (1)	OLS (2)	OLS (3)	OLS (4)
$I[Students < 10]$	0.17***	0.23***	0.02	0.05
	(0.04)	(0.04)	(0.04)	(0.03)
$I[10 \leq Students < 15]$	0.05	0.11***	0.03	0.06*
	(0.04)	(0.03)	(0.03)	(0.03)
$I[15 \leq Students < 27]$	0.05*	0.04	0.02	0.01
	(0.03)	(0.03)	(0.02)	(0.02)
Class size	-0.01***		-0.00*	
	(0.00)		(0.00)	
Sample	2^{nd} Grade	2^{nd} Grade	5^{th} Grade	5^{th} Grade
Observations	92,469	92,469	89,780	89,780

Notes: Reduced-form estimates of the effect of the instruments on a child's test score. Dependent variable: Combined Math-Language test score. Columns (1) and (2) refer to second-grade students. Columns (3) and (4) refer to fifth-grade students. The world *Students* refers to the number of (enrolled) second-grade (columns 1 and 2) and fifth-grade students (columns 3 and 4). The reference category for the number of students is $I[Students \geq 27]$. All models include controls for child's gender, age, nationality, father's and mother's educational level, and father's and mother's profession. All models also include variables for altitude and population of the municipality, geographical macro-area, and road distance in time to the closest alternative school. Standard errors are clustered at the school level and reported in brackets. *, **, *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

III. SENSITIVITY TESTS

In this section we present several sensitivity tests for our results. In Section III.A, instead of instrumenting both multigrade and class size with the same four instruments, we construct different sets of instruments based on first-stage evidence for the two endogenous variables of interest. Results are unaffected by the use of this different set of instruments.

In Section III.B, we test our baseline model for possible residual endogeneity due to parental preferences for single versus multigrade classes. To do so, we estimate a model with travel time to the closest school as an additional instrumental variable. The analysis does not display any relevant difference with respect to the baseline results.

A further source of concern in our analysis comes from differences in school size when comparing schools with multigrade classes with schools with single-grade classes. We replicate our main analysis by restricting the sample to schools with a cohort size of second or fifth graders of no more than 30 (two times the minimum number of students required to form a single-grade class), or 60 students. Section III.C reports the IV estimates and shows that all the main conclusions hold on these two different subsamples.

In the baseline analysis, we considered the combined math-language test score as the main outcome of interest. In Section III.D, we focus on its two components separately, namely, mathematics and language. The main results are consistent with those obtained with the combined measure.

In Section III.E, we discuss concerns related to possible opportunistic behavior on standardized tests. By restricting the sample to geographical areas with no (or little) evidence of opportunistic behavior, we find remarkably similar results to those of the baseline analysis.

Finally, in Section III.F, we address concerns about parents' missing information, and we also investigate the sensitivity of our results to the use of different geographical levels of aggregations (Nomenclature of Territorial Units for Statistics, NUTS 2 and NUTS 3). Results remain unaltered.

A. Exploiting Different Sets of Instruments for Multigrading and Class Size

The four instruments identified by Italian law DPR 81/2009 differently affect the two endogenous variables of our model, namely multigrading and class size. According to first-stage estimates in Tables 4 and 5 of the paper, the first two intervals determine the probability of attendance in a multigrade class, while the third and fourth intervals mainly affect class size.

In this section we re-estimate our baseline model considering both $Multigrade$ and $ClassSize$ as endogenous variables, but using two different sets of instruments for each of the two variables.

The two sets of instruments are the ones suggested by first-stage estimates. Specifically, to instrument multigrade, we exploit the first three intervals in terms of enrolled students determined by the 10-student cutoff and the 15-student cutoff (fewer than 10 students, 10–14 students, and more than 14 students). We consider the intervals of 1–14 students, 15–26 students, and more than 26 students as instruments for class size.

Table A.2 reports results for this exercise. First-stage estimates and diagnostic tests for both endogenous variables testify to a very precise first stage. Results are remarkably similar to those in the baseline models in Table 6 of the paper. We record a positive impact of multigrading on the combined math-language test score for second-grade students, while the effect for fifth graders is statistically insignificant. Larger class size negatively affects the performance of both second and fifth graders.

Table A.2: Different Sets of Instruments for Multigrading and Class Size

	Combined Math-Language	
	IV (1)	IV (2)
Multigrade	0.19*** (0.05)	-0.01 (0.05)
Class size	-0.01 (0.01)	-0.01 (0.00)
	First stage Multigrade	
$I[Students < 10]$	0.86*** (0.01)	0.79*** (0.02)
$I[10 \leq Students < 15]$	0.15*** (0.01)	0.10*** (0.01)
F-stat ($I[Students < 10]$)	> 100	> 100
F-stat ($I[10 \leq Students < 15]$)	> 100	87.46
	First stage Class size	
$I[Students < 15]$	-5.34*** (0.18)	-5.00*** (0.17)
$I[15 \leq Students < 27]$	0.97*** (0.17)	1.09*** (0.18)
F-stat ($I[Students < 15]$)	> 100	> 100
F-stat ($I[15 \leq Students < 27]$)	31.11	38.82
Instrumented variables	Multigrade, Class size	Multigrade, Class size
Sample	2^{nd} Grade	5^{th} Grade
Observations	92,469	89,780

Notes: Analysis of the effect of multigrading on a child's test score. Dependent variable: Combined Math-Language test score. Column (1) refers to second-grade students. Column (2) refers to fifth-grade students. The word *Students* refers to the the number of (enrolled) second-grade (column 1) and fifth-grade students (column 2). The reference category in the first stage for multigrade is $I[Students \geq 15]$. The reference category in the first stage for class size is $I[Students \geq 27]$. All models include controls for child's gender, age, nationality, father's and mother's educational level, and father's and mother's profession. All models also include variables for altitude and population of the municipality, geographical macro-area, and road distance in time to the closest alternative school. Standard errors are clustered at the school level and reported in brackets. *, **, *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

B. Testing Parental Preferences: Travel Time to the Closest School

We further discuss the possible existence of parents' preference for single versus multigrade classes. Such preference is a potential source of endogeneity underlying individual attendance in a multigrade class. First-stage evidence (see Section V.A of the paper) regarding the role of parental education in shaping multigrade class attendance signals that this is unlikely to be a threat to the reliability of our findings. However, to be even more cautious, we estimate an additional IV specification in which travel time to the closest school is also used as an instrument. Assuming that parental preferences about grade composition play a role in choosing a school for their children, we have to consider that these preferences are constrained by the time needed to reach the closest alternative school.

Table A.3 shows the first-stage estimates of the model. First-stage estimates suggest that travel time to the closest school plays a modest role in determining the individual probability of attendance in a multigrade class. For second-grade students, an additional one-minute distance causes an increase in the probability of attendance in a multigrade class of 0.1 percentage points. The coefficient is statistically significant at the 5 percent level. The effect is statistically insignificant for fifth-grade students. No effect of travel time on class size is detected.

Table A.4 reports second-stage estimates. Main results are unaffected by the inclusion of travel time to the closest school as an additional instrument. Similar to our baseline analysis, the effect of multigrading on child cognitive achievement ranges between 16 and 19 percent of a standard deviation for second graders, while the effect for fifth graders is close to zero.

Table A.3: First-Stage Estimates With Distance to the Closest School as Instrument

	Model (1)	Model (2)		Model (3)	Model (4)	
	Multigrade	Multigrade	Class size	Multigrade	Multigrade	Class size
	OLS	OLS	OLS	OLS	OLS	OLS
	(1)	(2)	(3)	(4)	(5)	(6)
$I[Students < 10]$	0.90***	0.86***	-5.19***	0.83***	0.79***	-5.06***
	(0.01)	(0.01)	(0.20)	(0.01)	(0.02)	(0.20)
$I[10 \leq Students < 15]$	0.19***	0.15***	-5.44***	0.15***	0.10***	-4.97***
	(0.02)	(0.01)	(0.18)	(0.01)	(0.01)	(0.16)
$I[15 \leq Students < 27]$	-0.01***	-0.00	0.97***	-0.01***	-0.00	1.09***
	(0.00)	(0.00)	(0.17)	(0.00)	(0.00)	(0.18)
Time distance	0.00**	0.00**	-0.02	0.00	0.00	-0.01
	(0.00)	(0.00)	(0.02)	(0.00)	(0.00)	(0.02)
Class size	0.01***			0.01***		
	(0.00)			(0.00)		
SW Chi-sq. (UId)	> 100	> 100	> 100	> 100	> 100	> 100
p-value	0.00	0.00	0.00	0.00	0.00	0.00
SW F (WId)	> 100	> 100	> 100	> 100	> 100	> 100
p-value	0.00	0.00	0.00	0.00	0.00	0.00
KP (WId)	> 100	> 100	> 100	> 100	> 100	> 100
F-stat ($I[Students < 10]$)	> 100	> 100	> 100	> 100	> 100	> 100
F-stat ($I[10 \leq Students < 15]$)	> 100	> 100	> 100	> 100	86.66	> 100
F-stat ($I[15 \leq Students < 27]$)	10.23	0.14	31.28	25.06	2.58	38.64
F-stat (Time distance)	5.24	4.30	0.92	0.67	0.51	0.15
Instrumented variable(s)	Multigrade	Multigrade, Class size		Multigrade	Multigrade, Class size	
Sample	2^{nd} Grade	2^{nd} Grade	2^{nd} Grade	5^{th} Grade	5^{th} Grade	5^{th} Grade
Observations	92,469	92,469	92,469	89,780	89,780	89,780

Notes: First-stage estimates with distance to the closest school as an additional instrument. Dependent variable: Attendance in a multigrade class (columns 1, 2, 4, and 5), class size (columns 3 and 6). Models (1) and (2) refer to second-grade students. Models (3) and (4) refer to fifth-grade students. The word *Students* refers to the the number of (enrolled) second-grade (models 1 and 2) and fifth-grade students (models 3 and 4). The reference category for the number of students is $I[Students \geq 27]$. Road distance (in minutes) to the closest alternative school is used as an additional instrument for attendance in a multigrade class and class size. All models include controls for child's gender, age, nationality, and father's and mother's profession. All models also include variables for altitude and population of the municipality, geographical macro-area, and road distance in time to the closest alternative school. Standard errors are clustered at the school level and reported in brackets. *, **, *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

Table A.4: Second-Stage Estimates With Distance to the Closest School as Instrument

	Combined Math-Language			
	IV (1)	IV (2)	IV (3)	IV (4)
Multigrade	0.16***	0.19***	0.00	-0.02
	(0.04)	(0.05)	(0.04)	(0.05)
Class size	-0.01***	-0.01	-0.01**	-0.01
	(0.00)	(0.01)	(0.00)	(0.00)
Instrumented variable(s)	Multigrade	Multigrade, Class size	Multigrade	Multigrade, Class size
Sample	2^{nd} Grade	2^{nd} Grade	5^{th} Grade	5^{th} Grade
Observations	92,469	92,469	89,780	89,780

Notes: Analysis of the effect of multigrading on a child's test score with distance to the closest school as an additional instrument. Dependent variable: Combined Math-Language test score. Columns (1) and (2) refer to second-grade students. Columns (3) and (4) refer to fifth-grade students. Road distance (in minutes) to the closest alternative school is used as an additional instrument for attendance in a multigrade class and class size. All models include controls for child's gender, age, nationality, father's and mother's educational level, and father's and mother's profession. All models also include variables for altitude and population of the municipality and geographical macro-area. Standard errors are clustered at the school level and reported in brackets. *, **, *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

C. School Size Restriction: The Cases of 30 and 60 Students

One potential concern with the baseline analysis relates to difference in school size when comparing schools with multigrade classes with schools with single-grade classes. This concern is mitigated by the evidence that a considerable fraction of students in schools with fewer than 10 students enrolled attend a single-grade class instead of the predicted-by-the-law multigrade class. Although the law, in principle, prevents the creation of single-grade classes with fewer than 10 students, in practice they exist. In our sample, about 25 percent of students enrolled in schools with fewer than 10 second graders attend a single-grade class. In the case of fifth graders, about 21 percent of students enrolled in schools with fewer than 10 fifth graders attend a single-grade class. This evidence, on the one hand, makes the use of an instrumental variable approach essential to deal with possible endogeneity. On the other hand, single-grade classes with fewer than 10 students are crucial for our analysis as they allow us to separately estimate the class size effect and the effect of grade composition.

To be even more cautious about the definition of our (single-grade) control group, we re-estimate the baseline specifications on two different subsamples. Specifically, we limit the analysis to schools with a cohort size of second or fifth graders of no more than 30 or 60 students. The choice is driven by the law, which fixes the minimum number of students needed to form a single-grade class at 15.

Table A.5 reports IV estimates. All the main conclusions hold on these two different subsamples: the impact of multigrading for second-grade students is bounded between 15 and 21 percent of a standard deviation; for fifth-grade students, the impact is close to zero and statistically insignificant. Class size displays the same results as in the baseline analysis.

Table A.5: Multigrading and Cohort Size

	Combined Math-Language							
	IV (1)	IV (2)	IV (3)	IV (4)	IV (5)	IV (6)	IV (7)	IV (8)
Multigrade	0.15***	0.19***	0.19***	0.21***	0.02	-0.01	0.00	0.00
	(0.04)	(0.05)	(0.05)	(0.05)	(0.04)	(0.05)	(0.04)	(0.05)
Class size	-0.01***	-0.01	-0.01***	-0.01	-0.00**	-0.01	-0.01***	-0.01*
	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
Instrumented variable(s)	Multigrade	Multigrade, Class size	Multigrade	Multigrade, Class size	Multigrade	Multigrade, Class size	Multigrade	Multigrade, Class size
Cohort size	$\leq$ 60 Stud.	$\leq$ 60 Stud.	$\leq$ 30 Stud.	$\leq$ 30 Stud.	$\leq$ 60 Stud.	$\leq$ 60 Stud.	$\leq$ 30 Stud.	$\leq$ 30 Stud.
Sample	2^{nd} Grade	2^{nd} Grade	2^{nd} Grade	2^{nd} Grade	5^{th} Grade	5^{th} Grade	5^{th} Grade	5^{th} Grade
Observations	69,629	69,629	37,399	37,399	68,891	68,891	36,980	36,980

Notes: Analysis of the effect of multigrading on a child's test score with sample restricted to small schools. Dependent variable: Combined Math-Language test score. Columns (1) to (4) refer to second-grade students. Columns (5) to (8) refer to fifth-grade students. Columns (1) and (2) only include schools with at most 60 second-grade (enrolled) students. Columns (3) and (4) only include schools with at most 30 second-grade (enrolled) students. Columns (5) and (6) only include schools with at most 60 fifth-grade (enrolled) students. Columns (7) and (8) only include schools with at most 30 fifth-grade (enrolled) students. All models include controls for child's gender, age, nationality, father's and mother's educational level, and father's and mother's profession. All models also include variables for altitude and population of the municipality, geographical macro-area, and road distance in time to the closest alternative school. Standard errors are clustered at the school level and reported in brackets. *, **, *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

D. Multigrading and Performance in Mathematics and Language

The baseline analysis of this paper is based on a combined math-language test score on the standardized INVALSI test. In this Appendix we separately test the effect of multigrading on the two single tests, namely, the test in mathematics and the test in language.

Table A.6 shows that attendance in a multigrade class positively affects both the performance in mathematics and in language. The multigrade effect on mathematics scores seems slightly higher in magnitude (17–20 percent of a standard deviation) compared to the effect on language scores (12–14 percent of a standard deviation), although the difference between the two is not statistically significant. As in the baseline analysis, there is no significant multigrade effect on test scores for fifth graders.

Table A.6: Math and Language Test Scores

	IV (1)	IV (2)	IV (3)	IV (4)
	Panel (a): Math			
Multigrade	0.17***	0.20***	0.02	-0.00
	(0.04)	(0.05)	(0.04)	(0.05)
Class size	-0.01***	-0.01	-0.01**	-0.01*
	(0.00)	(0.01)	(0.00)	(0.01)
	Panel (b): Language			
Multigrade	0.12***	0.14***	0.00	-0.01
	(0.04)	(0.05)	(0.03)	(0.04)
Class size	-0.01***	-0.01	-0.00**	-0.01
	(0.00)	(0.01)	(0.00)	(0.00)
Instrumented variable(s)	Multigrade	Multigrade, Class size	Multigrade	Multigrade, Class size
Sample	2^{nd} Grade	2^{nd} Grade	5^{th} Grade	5^{th} Grade
Observations	92,469	92,469	89,780	89,780

Notes: Analysis of the effect of multigrading on a child's test score in Mathematics and in Language. Dependent variable: Math test score (Panel (a)), Language test score (Panel (b)). Columns (1) and (2) refer to second-grade students. Columns (3) and (4) refer to fifth-grade students. All models include controls for child's gender, age, nationality, father's and mother's educational level, and father's and mother's profession. All models also include variables for altitude and population of the municipality, geographical macro-area, and road distance in time to the closest alternative school. Standard errors are clustered at the school level and reported in brackets. *, **, *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

E. Opportunistic Behavior on the INVALSI Test

The use of scores on standardized tests to assess individuals' skills is common in social sciences such as economics, sociology, and psychology. However, given that standardized tests are useful tools to compare different schools, classes, and teachers, this produces potential incentives for opportunistic behavior by principals, teachers, and even students. Bertoni, Brunello and Rocco (2013) provide the first empirical evidence of possible cheating on the INVALSI test. According to Lucifora and Tonello (2015), cheating mainly occurs when teachers shirk or decrease monitoring. In analyzing class size and score manipulation in the southern regions, Angrist, Battistin and Vuri (2017) find that cheating largely reflects teacher behavior, motivated by moral hazard in their grading effort. According to their estimates, roughly 5 percent of Italian scores are biased because of cheating.

Several factors suggest that cheating should not be a major concern for our analysis. The existence of cheating in our setting would imply that our outcome of interest is the real score of the test plus some noise. On the one hand, if noise is stochastic, this would only affect our estimates by lowering precision, and all the coefficients would remain consistently estimated. On the other hand, if noise is correlated with our variable of interest (attendance in a multigrade class) the multigrade coefficient estimate would potentially be biased.

As also confirmed by our discussions with administrators, principals, primary school teachers, and members of INVALSI, it is difficult to believe that the probability of observing opportunistic and cheating behavior directly depends on considering a single versus a multigrade class. Other elements suggested by the literature, such as teachers' unobserved characteristics, should be considered as the main determinants of possible cheating (Angrist, Battistin and Vuri, 2017). This anecdotal evidence is also confirmed by intraschool variability in cheating patterns.

We empirically deal with possible cheating-induced bias in our estimates with the analysis of geographical patterns underlying our baseline results. As shown by Bertoni, Brunello and Rocco (2013), Angrist, Battistin and Vuri (2017), and other qualitative studies, cheating is a major concern in southern Italy and much less so in northern regions. According to the score manipulation index developed by Angrist, Battistin and Vuri (2017) cheating only accounts for 2 percent of scores in the northern and central regions of Italy; this percentage is even lower for only northern regions.

With this evidence in mind, in Table A.7, we replicate our baseline analysis focusing on regions in northern Italy. Despite the reduced sample size, results for the northern regions are similar to the ones we obtain for the whole country. We observe a positive and significant effect of multigrading only for second-grade students (9–15 percent of a standard deviation). For fifth graders, the impact is statistically insignificant. The coefficients of our relevant covariates in the sample from the northern regions are never statistically different from those in the sample that includes the whole set of Italian regions.

Table A.7: Multigrading and Child Achievement: The Case of Northern Regions

	Combined Math-Language					
	OLS (1)	IV (2)	IV (3)	OLS (4)	IV (5)	IV (6)
Multigrade	0.09**	0.11**	0.15**	-0.05	-0.05	-0.05
	(0.04)	(0.05)	(0.06)	(0.03)	(0.04)	(0.06)
Class size	-0.01***	-0.01***	-0.00	-0.01**	-0.01**	-0.01
	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.01)
Instrumented variable(s)		Multigrade	Multigrade, Class size		Multigrade	Multigrade, Class size
Regions	Only North	Only North	Only North	Only North	Only North	Only North
Sample	2^{nd} Grade	2^{nd} Grade	2^{nd} Grade	5^{th} Grade	5^{th} Grade	5^{th} Grade
Observations	58,329	58,329	58,329	54,402	54,402	54,402

Notes: Analysis of the effect of multigrading on a child's test score with sample restricted to regions in northern Italy. Dependent variable: Combined Math-Language test score. Columns (1) to (3) refer to second-grade students. Columns (4) to (6) refer to fifth-grade students. The analysis is only based on northern Italian regions. All models include controls for child's gender, age, nationality, father's and mother's educational level, and father's and mother's profession. All models also include variables for altitude and population of the municipality and road distance in time to the closest alternative school. Standard errors are clustered at the school level and reported in brackets. *, **, *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

F. Further Sensitivity Tests

In this section we perform additional tests of our results. In the first test (Table A.8, Panel (a)) we tackle one of the limitations of the INVALSI data: missing information about parents or about some of their characteristics, such as education or profession. Although in our baseline analysis we introduced residual groups for students with missing parental information to allow us to consider single-head households, here we restrict our sample to include only students for whom information about both parents' profession and educational level are available. Results are remarkably similar to those in the baseline analysis.

In Panels (b) and (c) of Table A.8, we test different levels of geographical aggregation for our data. In our main analysis, we use the five macro-regions (NUTS 1) to capture macro-regional fixed effects. Here we estimate two alternative models by considering regional fixed effects (NUTS 2, Panel (b)) and provincial fixed effects (NUTS 3, Panel (c)). Results remain unchanged in both specifications, suggesting that the choice of geographical level of aggregation does not affect the size and significance of our findings.

Table A.8: Sensitivity Analysis

	Combined Math-Language			
	IV (1)	IV (2)	IV (3)	IV (4)
	Panel (a): Parents' missing information			
Multigrade	0.16*** (0.04)	0.17*** (0.05)	-0.02 (0.04)	-0.05 (0.05)
Class size	-0.01*** (0.00)	-0.01 (0.01)	-0.01*** (0.00)	-0.01** (0.01)
	Panel (b): Regional (NUTS 2) FE			
Multigrade	0.18*** (0.04)	0.20*** (0.05)	0.02 (0.04)	-0.01 (0.05)
Class size	-0.01*** (0.00)	-0.01 (0.01)	-0.00** (0.00)	-0.01* (0.00)
	Panel (c): Provincial (NUTS 3) FE			
Multigrade	0.20*** (0.04)	0.21*** (0.05)	0.02 (0.04)	-0.01 (0.05)
Class size	-0.01*** (0.00)	-0.01* (0.01)	-0.01** (0.00)	-0.01* (0.00)
Instrumented variable(s)	Multigrade	Multigrade, Class size	Multigrade	Multigrade, Class size
Sample	2^{nd} Grade	2^{nd} Grade	5^{th} Grade	5^{th} Grade

Notes: Sensitivity analysis for baseline estimates. Dependent variable: Combined Math-Language test score. Columns (1) and (2) refer to second-grade students. Columns (3) and (4) refer to fifth-grade students. All models include controls for child's gender, age, nationality, father's and mother's educational level, and father's and mother's profession. All models also include variables for altitude and population of the municipality, geographical macro-area (except Panels (b) and (c)), and road distance in time to the closest alternative school. Standard errors are clustered at the school level and reported in brackets. *, **, *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

IV. CLASSROOM GRADE COMPOSITION: FIRST-STAGE ESTIMATES

In this section we provide the first-stage estimates for the IV analysis of classroom grade composition and the effect of multigrade on a second-grade student's test score. Refer to Section VI.A of the paper for further details on the analysis.

Table A.9: First-Stage Estimates for the Analysis of Classroom Grade Composition

	Model (1)	Model (2)		Model (3)	Model (4)	
	Multigrade	Multigrade	Class size	Multigrade	Multigrade	Class size
	OLS	OLS	OLS	OLS	OLS	OLS
	(1)	(2)	(3)	(4)	(5)	(6)
$I[2^{nd}G < 10]$ & $I[AdjG < 15]$	0.79***	0.77***	-5.65***	0.67***	0.65***	-6.39***
	(0.02)	(0.03)	(0.24)	(0.03)	(0.04)	(0.26)
$I[2^{nd}G < 10]$ & $I[15 \leq AdjG < 27]$	0.07**	0.03	-9.59***	0.25***	0.22***	-8.50***
	(0.03)	(0.03)	(0.37)	(0.07)	(0.07)	(0.50)
$I[2^{nd}G < 10]$ & $I[AdjG \geq 27]$	(Not observed in the sample)			0.04***	0.00	-9.95***
				(0.01)	(0.00)	(0.21)
$I[10 \leq 2^{nd}G < 15]$ & $I[AdjG < 10]$	0.39***	0.37***	-5.13***	0.20***	0.18***	-5.34***
	(0.04)	(0.04)	(0.23)	(0.04)	(0.04)	(0.32)
$I[10 \leq 2^{nd}G < 15]$ & $I[10 \leq AdjG < 15]$	0.07***	0.05***	-5.84***	0.05***	0.02***	-5.96***
	(0.02)	(0.01)	(0.20)	(0.01)	(0.01)	(0.19)
$I[10 \leq 2^{nd}G < 15]$ & $I[AdjG \geq 15]$	0.02***	0.00	-5.83***	0.06***	0.04***	-5.56***
	(0.00)	(0.00)	(0.17)	(0.01)	(0.01)	(0.19)
$I[15 \leq 2^{nd}G < 27]$ & $I[AdjG < 10]$	0.07*	0.07*	-0.56*	0.01***	0.00	-1.09***
	(0.04)	(0.04)	(0.33)	(0.00)	(0.00)	(0.33)
$I[15 \leq 2^{nd}G < 27]$ & $I[AdjG \geq 10]$	-0.00**	0.00	1.02***	-0.00***	0.00	1.05***
	(0.00)	(0.00)	(0.17)	(0.00)	(0.00)	(0.17)
Class size	0.00***			0.00***		
	(0.00)			(0.00)		
SW Chi-sq. (UId)	> 100	> 100	> 100	> 100	> 100	> 100
p-value	0.00	0.00	0.00	0.00	0.00	0.00
SW F (WId)	> 100	> 100	> 100	62.86	> 100	> 100
p-value	0.00	0.00	0.00	0.00	0.00	0.00
KP (WId)	> 100	> 100	> 100	62.86	> 100	> 100
Instrumented variable(s)	Multigrade	Multigrade, Class size		Multigrade	Multigrade, Class size	
In multigrade with	Younger peers	Younger peers	Younger peers	Older peers	Older peers	Older peers
Sample	2^{nd} Grade	2^{nd} Grade	2^{nd} Grade	2^{nd} Grade	2^{nd} Grade	2^{nd} Grade
Observations	89,461	89,461	89,461	88,428	88,428	88,428

Notes: First-stage estimates for second-grade students by classroom grade composition. Dependent variable: Attendance in a multigrade class with younger (columns 1 and 2) or older peers (columns 4 and 5), and class size (columns 3 and 6). Models (1) and (3) include class size as a control variable. Models (2) and (4) treat both multigrade and class size as endogenous variables. Younger peers (models 1 and 2) means that only first-grade students attend the same multigrade class of second-grade students. Older peers (models 3 and 4) means that children of higher grades (third, fourth, and fifth grades) attend the same multigrade class of second-grade students. $2^{nd}G$ stands for number of second-grade (enrolled) students. $AdjG$ stands for number of (enrolled) students in the adjacent grade. The adjacent grade is the first grade in models (1) and (2), and the third grade in models (3) and (4). The reference category for the set of instruments is $I[2^{nd}G \geq 27]$. All models include controls for child's gender, age, nationality, father's and mother's educational level, and father's and mother's profession. All models also include variables for altitude and population of the municipality, geographical macro-area, and road distance in time to the closest alternative school. Standard errors are clustered at the school level and reported in brackets. *, **, *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

REFERENCES

Angrist, Joshua D., Erich Battistin, and Daniela Vuri. 2017. "In a Small Moment: Class Size and Moral Hazard in the Italian Mezzogiorno." *American Economic Journal: Applied Economics* 9(4):216–49.

Bertoni, Marco, Giorgio Brunello, and Lorenzo Rocco. 2013. "When the Cat is Near, the Mice Won't Play: The Effect of External Examiners in Italian Schools." *Journal of Public Economics* 104:65–77.

Lucifora, Claudio, and Marco Tonello. 2015. "Cheating and Social Interactions. Evidence from a Randomized Experiment in a National Evaluation Program." *Journal of Economic Behavior & Organization* 115:45–66.