Discussion Paper Series

IZA DP No. 19017

October 2026

Gender Differences in High-Stakes Performance Among Early Adolescents: Evidence from a Middle-School Exit Exam Reform

Annalisa Loviglio
University of Bologna
and IZA@LISER

Veronica Rattini
University of Bologna
and IZA@LISER

Federico Stronati
University College Dublin

The IZA Discussion Paper Series (ISSN: 2365-9793) ("Series") is the primary platform for disseminating research produced within the framework of the IZA@LISER Network, an unincorporated international network of labour economists coordinated by the Luxembourg Institute of Socio-Economic Research (LISER). The Series is operated by LISER, a Luxembourg public establishment (établissement public) registered with the Luxembourg Business Registers under number J57, with its registered office at 11, Porte des Sciences, 4366 Esch-sur-Alzette, Grand Duchy of Luxembourg.

Any opinions expressed in this Series are solely those of the author(s). LISER accepts no responsibility or liability for the content of the contributions published herein. LISER adheres to the European Code of Conduct for Research Integrity. Contributions published in this Series present preliminary work intended to foster academic debate. They may be revised, are not definitive, and should be cited accordingly. Copyright remains with the author(s) unless otherwise indicated.

IZA@LISER NETWORK
ISSN: 2365-9793

IZA DP No. 19017 October 2026

Gender Differences in High-Stakes Performance Among Early Adolescents: Evidence from a Middle-School Exit Exam Reform*

Abstract

We study whether high-stakes testing affects the performance of female and male students differently in early adolescence. Existing evidence from high-school, university, and admission settings suggests that female students often perform relatively worse when stakes are high, but evidence from earlier stages of the educational career remains limited. We exploit a 2017 reform of Italy's lower-secondary exit exam that removed the score on a nationwide standardized test from the calculation of the final grade while preserving mandatory participation. Using administrative data covering the full population of students and a gender-by-reform difference-in-differences specification, we compare the female–male performance gap across the high- and low-stakes regimes. Girls outperform boys in both regimes, and their relative advantage is modestly larger—by about 0.03–0.05 standard deviations—when the test carries direct grade consequences. Our results provide population-wide quasi-experimental evidence that the relative female losses documented in high-stakes settings at later educational stages are not yet present in early adolescence, suggesting that gender differences in responses to test stakes may vary across educational stages and institutional settings.

JEL classification

I21, I24, J16

Keywords

high-stakes testing, gender gaps, standardized exams, adolescence

Corresponding author

Annalisa Loviglio
annalisa.loviglio@unibo.it

* V. Rattini gratefully acknowledges funding by the European Union – NextGenerationEU, Mission 4, Component 2, in the framework of the GRINS – Growing Resilient, INclusive and Sustainable project (GRINS PE00000018 – CUP J33C22002910001).

1 Introduction

Standardized tests are widely used to measure students' competencies and to provide a common metric of achievement across individuals, schools, and cohorts. Some are purely diagnostic and carry no direct consequences for test takers; these are typically regarded as *low-stakes*.[1] Others have immediate implications for students' educational and career opportunities, including secondary-school exit exams and university admission tests, and are therefore *high-stakes*. Because these assessments often govern access to higher education and, indirectly, to later labor-market opportunities, they are expected to provide accurate measures of the competencies they are designed to capture. Yet when stakes are high, observed performance may reflect not only knowledge and skills, but also individuals' responses to pressure, anxiety, or competition.

A growing literature studies whether performance in high-stakes testing environments reflects underlying ability or, instead, differential behavioral responses to incentives, stress, and competition. A prominent result is that, in late adolescence and early adulthood, female students often perform relatively worse than male students in testing environments that are more consequential or competitive (Ors et al., 2013; Azmat et al., 2016; Cai et al., 2019; Schlosser et al., 2019; Iriberri and Rey-Biel, 2019; Montolio and Taberner, 2021).[2]

Evidence on gender differences in responses to test stakes in early adolescence remains much more limited. A notable exception is Azmat et al. (2016), who find a modest decline in girls' relative advantage in higher-stakes tests among students aged 12–14 at a private school in Barcelona. More generally, existing studies often compare assessments that differ not only in stakes but also in format, content, or the amount of material covered (Azmat et al., 2016; Ors et al., 2013; Montolio and Taberner, 2021), or examine settings in which higher stakes are intertwined with competition for a limited number of educational opportunities or prizes (Ors et al., 2013; Cai et al., 2019; Iriberri and Rey-Biel, 2019).

We study whether gender differences in performance under high-stakes conditions already emerge in early adolescence. We address this question by exploiting a national reform of Italy's lower-secondary exit exam that sharply reduced the stakes attached to one standardized component while leaving participation compulsory. This exam is the first high-stakes national examination faced by Italian students and marks a salient milestone in the transition from lower- to upper-secondary education. It certifies completion of lower-secondary schooling, enables progression to upper-secondary education, and carries broader educational and social significance: the final result is formally recorded, publicly

[1] Prominent examples include large-scale international assessments such as the OECD's PISA and the IEA's PIRLS and TIMSS.

[2] Related evidence shows that gender differences in performance may also respond to other features of the testing environment. In a field experiment with university students, De Paola and Gioia (2016) find that binding time constraints reduce female performance while leaving male performance largely unchanged.

displayed at the student's middle school, and transmitted to the receiving upper-secondary school. At the same time, exam results are not used to allocate students across schools or tracks. The setting therefore allows us to study how the gender gap changes when a national standardized test becomes less consequential, without simultaneously changing competition for a fixed number of places.

Our analysis exploits the fact that, until 2016/17, the grade-8 INVALSI score contributed directly to the lower-secondary exit-exam grade, accounting for one-seventh (just over 14 percent) of the final score.[3] Starting in 2017/18, a reform removed the INVALSI score from the calculation of the final grade while preserving participation in the assessment as a requirement for admission to the exit exam.[4] The reform therefore eliminated the test's direct grade consequences while leaving both the national assessment and the participation requirement in place. This institutional change provides quasi-experimental variation to study how the gender gap in test performance responds when the stakes attached to a standardized assessment are reduced. Accordingly, we classify INVALSI assessments administered before the reform as high-stakes and those administered after the reform as low-stakes.

Using administrative data on the full population of students, we estimate a gender-by-reform difference-in-differences model, following an empirical strategy similar to Arenas and Calsamiglia (2025).[5] We find that the gender difference in performance is largely stable across high- and low-stakes testing regimes. Girls outperform boys both before and after the reform and, if anything, their relative advantage is slightly larger when the test contributes to the exit-exam grade. The estimated change in the female–male gap is modest, ranging from about 0.03 to 0.05 standard deviations across the main specifications. Overall, the evidence points to at most a small difference between the high- and low-stakes regimes, slightly favorable to girls under high stakes.

This result is robust across a range of complementary exercises. First, the main conclusion remains unchanged when we account for the contemporaneous transition from paper-based to computer-based testing. We estimate an analogous specification for grade-

[3] INVALSI is a nationwide assessment in Italian and mathematics administered to the universe of students and designed to measure competencies on a common scale.

[4] The reform was introduced by Legislative Decree n. 62 of April 13, 2017 (*Decreto Legislativo 13 aprile 2017, n. 62*).

[5] Arenas and Calsamiglia (2025) exploit a reform in Catalonia that increased the contribution of the standardized university entrance examination to the composite admission grade which combines high-school GPA and entrance-exam performance from 40% to 57%. As in our setting, the reform applied to all students, and identification comes from the differential change in outcomes by gender. Their research question and outcome differ from ours, however, because they study the composite admission grade and its consequences for university allocation, whereas our outcome is performance on the standardized assessment itself. They find that the reform reduced women's relative admission grades and their representation in selective degree programs. Although separate high-school and entrance-exam scores are not observed before the reform, their decomposition suggests that most of this effect reflects the mechanical reweighting of the two achievement measures, with a smaller component consistent with a deterioration in women's relative entrance-exam performance as stakes increased.

10 students, who experienced the same change in test mode but no change in stakes. Complementary evidence from grade-8 teacher questionnaires shows no statistically significant association between classroom gender composition and reported difficulties with computer-based testing. Second, the result is replicated in the representative subsample of externally monitored classrooms, where concerns about score manipulation and cheating are substantially reduced. Third, the result is closely mirrored in within-student specifications comparing the same students across grades 5, 8, and 10 and absorbing time-invariant individual heterogeneity. We then examine heterogeneity by subject and find no evidence that higher stakes reduce girls' relative performance in either Italian or mathematics, although the modest positive interaction is concentrated in Italian. Taken together, these exercises point to the same conclusion: high stakes do not adversely affect girls' performance relative to boys in early adolescence. Finally, we discuss the potential role of gender differences in effort across testing regimes. Although our data do not allow us to disentangle changes in effort from responses to test pressure, evidence from previous research suggests that differential effort is unlikely to drive our results.

Our findings contribute to the literature in two main ways: by providing population-wide evidence on gender differences in responses to test stakes in early adolescence, and by offering comparatively direct evidence on test stakes as distinct from competition and other features of the assessment environment. First, we provide new evidence on gender differences in responses to test stakes in early adolescence. The closest existing evidence on test stakes at these ages comes from Azmat et al. (2016), who study a high-performing private school in Barcelona and find that girls' performance advantage is about 0.05 standard deviations smaller in higher-stakes tests already at ages 12–14. In our national grade-8 setting, by contrast, higher stakes do not reduce girls' relative performance and may slightly increase it. Our findings therefore show that relative female losses under high stakes are not a general feature of early adolescence and, together with the evidence for older students, suggest that responses to test stakes may vary across educational stages and institutional environments.

This first contribution also connects to a related literature on incentives and achievement earlier in the educational cycle. This small literature studies whether achievement is higher in grades in which performance has stronger consequences for tracking or progression (Bach and Fischer, 2020; Brunello and Kiss, 2022). In those studies, achievement is measured using low-stakes assessments that do not themselves determine students' grades or educational opportunities.[6] Our analysis addresses a distinct question: whether the

[6]Bach and Fischer (2020) exploit variation across German states in whether primary-school track recommendations are binding and find that stronger incentives raise achievement. Brunello and Kiss (2022) study primary- and lower-secondary-school students and find that mathematics scores on low-stakes assessments are higher when the assessments are administered in grades that are more consequential for track assignment or graduation. Both studies find larger achievement responses among girls, although in Brunello and Kiss (2022) the gender differences are not statistically significant.

gender gap in performance changes when the standardized assessment itself carries direct consequences. It therefore speaks to gender differences in performance when students sit a high-stakes test, rather than to incentives for learning and skill accumulation over the school year.

An adjacent experimental literature studies the emergence of gender differences in competitive behavior among children and adolescents. In Western samples, boys are often more willing than girls to enter competitive payment schemes, but evidence on gender differences in performance once competition is imposed is mixed and varies across tasks, ages, and cultural settings (Dreber et al., 2011, 2014; Cárdenas et al., 2012; Sutter and Glätzle-Rützler, 2015). The heterogeneity of these findings is consistent with our broader conclusion that gender differences in responses to stakes do not follow a universal pattern. Our setting is distinct, however, because all students are required to sit a nationwide standardized exam that is not competitive in nature: a student's outcome does not depend directly on how other students perform.

Second, the paper provides comparatively clean evidence on test stakes, as distinct from other features of the assessment environment. Existing studies often compare assessments that differ in format, content, or the amount of material covered (Azmat et al., 2016; Ors et al., 2013; Montolio and Taberner, 2021).[7] Even when the tasks are closely comparable, the low-stakes assessment may be an optional experimental section taken after the consequential examination, as in Schlosser et al. (2019). Moreover, in prominent admission and contest settings, higher stakes are intertwined with competition for a limited number of places or prizes (Jurajda and Munich, 2011; Cai et al., 2019; Ors et al., 2013; Iriberri and Rey-Biel, 2019). In our setting, the Italian reform changed whether performance on the grade-8 INVALSI national assessment contributed to the exit-exam grade, while the assessment continued to be administered nationally and participation remained compulsory. Because the assessment is not used to rank students for admission to capacity-constrained schools or tracks, the design speaks more directly to how the gender gap responds to the consequences attached to test performance, rather than to competition for scarce educational opportunities.

Our setting offers two additional advantages for identifying gender differences. First, before upper-secondary school, the curriculum is largely homogeneous nationwide and students are not tracked, reducing concerns that girls and boys have been exposed to systematically different educational environments before the assessment. Second, the data cover the universe of students and allow us to estimate gender gaps within the same schools and cohorts, reducing concerns about selective test participation or institution-

[7] More broadly, comparisons across standardized tests, anonymously graded external exams, and teacher-assigned evaluations may also reflect differences in grading practices or in the set of skills being measured, rather than the effect of stakes alone (Lavy, 2008; Calsamiglia and Loviglio, 2019; Saygin, 2020; Terrier, 2020; Angelo and Reis, 2021).

specific student composition. Together with the validation exercises described above, these features strengthen the interpretation of the estimates as evidence on responses to test stakes.

Our findings have implications for both the interpretation of standardized test scores and the design of assessment systems. If students from different groups respond differently to the incentives and pressure attached to an examination, observed achievement gaps may partly reflect behavioral responses to the testing environment rather than differences in underlying competence. In our setting, however, attaching direct grade consequences to the national assessment does not disadvantage girls at ages 13–14. More broadly, the results suggest that the effects of high-stakes testing on gender inequality may depend on the stage of schooling and on the institutional design of the assessment, including whether stakes are accompanied by competition for scarce opportunities.

The remainder of the paper is organized as follows. Section 2 describes the institutional setting, the reform, and the data. Section 3 presents the empirical design and the results, as well as robustness checks. Section 4 concludes.

2 Institutional Setting and Data

2.1 The Italian Education System

Basic education in Italy begins at age 6 and comprises five years of primary school followed by three years of lower-secondary school. The curriculum is largely homogeneous nationwide, and students are not tracked before the end of lower-secondary education.[8] At the end of grade 8, students sit a national exit exam, which they must pass to complete basic education and progress to upper-secondary education. We describe the structure of the exit exam and the role of the INVALSI assessment in greater detail in Section 2.3.

Grade repetition is formally possible throughout basic education but is uncommon. Using ISTAT data, we calculate average pre-COVID retention rates of 0.32% in primary school and 2.53% in lower-secondary school.[9] Consequently, students are usually 13 or 14 years old when they sit the grade-8 exit exam.

After completing lower-secondary school, students must enroll in upper-secondary education, choosing among three broad tracks: vocational, technical, and academic (*licei*). Admission is not formally determined by prior academic performance. Lower-secondary teachers provide non-binding track recommendations, but students' actual choices are

[8] Minor differences may arise in lower-secondary school with respect to the second foreign language taught, typically Spanish, French, or German.

[9] Students may repeat a grade because of insufficient attendance, failure to meet minimum learning standards, or failure to pass the national exit exam. During the COVID period, grade retention was strongly discouraged and remained lower afterward.

strongly associated with these recommendations, as well as with prior achievement, family background, and other characteristics (Carlana et al., 2022). Compulsory schooling continues until age 16, but this age threshold does not coincide with the completion of a school cycle or with any national exit exam.

2.2 INVALSI tests

Student achievement in Italy is monitored by the National Institute for the Evaluation of the Education System (INVALSI), which administers standardized national assessments each year.[10] During the period relevant for our analysis, INVALSI tests were administered in Italian and mathematics to the full population of students in grades 2, 5, 8, 10, and 13.[11] Tests are administered at school during the spring term. Participation is compulsory for schools, and students present on the test day are required to sit the assessment. In grade 8, moreover, taking the INVALSI test has been a formal requirement for admission to the lower-secondary exit exam since 2009/10. Students absent on the test day in grade 8 are required to take the assessment in a later make-up session, while in other grades absent students do not have to make up the test.

The grade-8 INVALSI assessment consists of two sections, Italian and mathematics, each lasting up to 75 minutes. Both include a combination of multiple-choice and open-ended questions. The Italian section assesses reading comprehension, lexical proficiency, and grammar, while the mathematics section evaluates numerical and logical skills. As detailed in Subsection 2.3, the key institutional feature is that the role of the grade-8 INVALSI test changed sharply over time. Until school year 2016/17, the INVALSI score entered the final lower-secondary exit-exam grade. Starting in 2017/18, sitting the test remained compulsory for admission to the exit exam, but the score no longer contributed to the final graduation mark.

2.3 Grade-8 Exit Exam and Policy Change

At the end of grade 8, students take the national lower-secondary exit exam, held each year in June over about three weeks. Its overall structure remained broadly stable throughout the period relevant for our analysis.[12] Students first receive an admission grade based on their performance during the three years of middle school. They then take written exams

[10] INVALSI data have been increasingly used in international research in the economics of education (e.g., Angrist, Joshua D. and Battistin, Erich and Vuri, Daniela (2017); Carlana (2019); Hanushek et al. (2026)).

[11] An English section was introduced later: in 2017/18 for grades 5 and 8, and in 2018/19 for grade 13. English scores are not used in this paper since they were not available throughout the analysis period.

[12] Relevant legal references include: Ministerial Decree of August 26, 1981; Legislative Decree n. 59 of February 19, 2004, Chapter IV, as subsequently amended by Law n. 176 of October 25, 2007; Presidential Decree n. 122 of June 22, 2009, Section 3; and Ministerial Circular n. 46 of May 26, 2011.

in Italian, mathematics, and foreign languages, followed by an oral examination covering the broader curriculum. Until school year 2016/17, the grade-8 INVALSI assessment was also a component of the exit exam and was the only component administered nationally in a common format. While the Ministry of Education sets the general framework, the remaining written and oral components are designed and graded internally by school-level examination committees.

The final graduation mark is assigned on a 1–10 scale, with 6 as the minimum passing grade, and combines the admission grade with performance across the exam components. Although it does not formally determine access to upper-secondary tracks, the exam is widely regarded as high-stakes in the Italian context. It is the first national exit exam students face, and its overall result is both institutionally salient and socially visible. The final graduation mark is transmitted to the receiving upper-secondary school and, by law, publicly posted at the student's lower-secondary school.[13]

Until school year 2016/17, the INVALSI score entered the final graduation mark and accounted for one seventh, i.e., more than 14%, of the final score. Contemporaneous coverage by leading Italian news outlets provides suggestive evidence that students perceived the pre-reform INVALSI component as a source of considerable pressure, consistent with its high-stakes role in the exit exam. In 2016, *La Stampa* reported results from a web survey of approximately 2,700 students taking the grade-8 exit exam: 27% identified INVALSI as the component they feared most, second only to the oral examination. In the same year, ANSA reported substantial discussion of the test on social media, with students expressing apprehension before and after the test and the hashtag #invalsi2016 appearing among the day's trending topics. In 2017, *Corriere della Sera* described INVALSI as the most feared written component of the exam.[14]

Questionnaires administered to grade-8 teachers provide complementary evidence on the role of INVALSI in schools. Among teachers who provided open-ended comments in the 2013/14 INVALSI Teacher Questionnaire, the inclusion of the national test in the grade-8 exit exam was one of the most frequently discussed topics (Martini and Papini, 2015). Preparation for the assessment was also widespread before the reform. In the 2016 and 2017 INVALSI Teacher Questionnaires, 99% of sampled grade-8 teachers reported undertaking at least one activity specifically intended to prepare students for the test, with 79% reporting that they practiced with tests from previous years.[15]

Starting in 2017/18, Legislative Decree n. 62/2017 removed the INVALSI score entirely from the calculation of the graduation mark, while preserving participation in the test as

[13] See Presidential Decree n. 122 of June 22, 2009, Section 3, Clause 9.

[14] See *La Stampa* (June 12, 2016), ANSA (June 17, 2016), and *Corriere della Sera* (June 17, 2017). All webpages were accessed on August 6, 2026.

[15] Other activities included using dedicated preparation books (50%), assigning INVALSI-like tests as homework (43%), and administering similar tests in class (33%). Figures are based on our calculations using the INVALSI Teacher Questionnaires described in Section 2.4.

a prerequisite for admission to the exit exam. The reform therefore did not merely reduce the weight of a component that remained consequential; it eliminated altogether the direct contribution of INVALSI performance to the final mark. In this sense, it operated at the extensive margin of performance-based incentives: before the reform, performing better on INVALSI could raise the graduation mark, whereas afterward, conditional on participation, performance had no effect on it. Because the assessment had previously determined one seventh of the overall mark, this represented the removal of a quantitatively meaningful incentive.[16]

Although the reform removed the direct grade consequences of INVALSI performance, the assessment retained a meaningful role in grade-8 schooling. Participation remained a prerequisite for admission to the exit exam, and teachers continued to prepare students for the test. In the 2018 and 2019 INVALSI Teacher Questionnaires, 91% of sampled grade-8 teachers reported using the practice tests provided by INVALSI at least sometimes during the school year.[17] This evidence suggests that the removal of performance-based incentives did not render the assessment irrelevant to classroom practices, supporting its characterization as a low-stakes test that remained salient within schools.

A contemporaneous change is also relevant for our empirical design. Starting in 2017/18, the grade-8 INVALSI test moved from paper-based to computer-based administration; the same transition occurred in grade 10. In the empirical analysis, we exploit adjacent grades and other validation exercises to distinguish the role of stakes from broader changes in the testing environment.

2.4 Data description

For each school year and grade, INVALSI provides two student-level datasets, one for Italian and one for mathematics. Each contains the subject-specific score—measured on a Rasch scale—together with rich background information.[18] In particular, we observe students' gender, origin, grade-retention history, parental education and occupation. Starting in school year 2011/12, the data also include a longitudinal student identifier (SIDI INVALSI), which allows us to link students across waves. Our grade-8 data cover school years 2014/15–2021/22. The 2014/15 cohort is the first whose grade-5 assessment falls

[16] For comparison, Arenas and Calsamiglia (2025) study a reform that increased the contribution of a standardized university entrance examination to a composite admission grade from 40% to as much as 57%. Their reform increased the weight of an examination that remained consequential in both regimes, whereas the Italian reform eliminated altogether the direct contribution of INVALSI performance to the final graduation mark.

[17] The post-reform questionnaire is less detailed than its pre-reform counterpart: it asks only about the use of practice tests provided by INVALSI, without covering other forms of preparation.

[18] The Rasch model is a standard psychometric framework used to scale performance in assessments based on categorical responses. Similar approaches are used in large-scale international surveys such as OECD's PISA and IEA's TIMSS.

within the period covered by the longitudinal identifier.

The INVALSI records do not include the grade-8 final graduation mark or other components of the exit exam. Our analysis therefore focuses exclusively on performance in the standardized INVALSI assessment.

We first merge the Italian and mathematics files and construct our main outcome as the average of the two subject scores. This reflects the fact that, in grade 8, the INVALSI assessment entered the lower-secondary exit exam as a single component and provides a summary measure of overall performance. The within-year match between the two subject files is successful for virtually the entire sample. We then standardize this average score within school year and grade to mean zero and standard deviation one, following Azmat et al. (2016). Finally, using the longitudinal identifier, we link grade-8 records with the same students' observations in grade 5 and in grade 10. The match with grade 5 covers 78% of the grade-8 sample, while the three-wave grade-5, grade-8, and grade-10 panel covers 67%. Incomplete longitudinal matching may reflect occasional SIDI inconsistencies, the absence of make-up sessions outside grade 8, grade retention, recent immigration among students who did not attend primary school in Italy, and dropout before grade 10.

For additional descriptive evidence, we also use INVALSI Teacher Questionnaires from 2015/16 to 2018/19, administered to a sample of roughly 2,000 grade-8 teachers each year. These data provide the evidence on test preparation reported in Section 2.3 and are used in Section 3.3 to examine teachers' perceptions of the transition to computer-based testing.

3 Empirical Strategy and Results

3.1 Empirical Strategy

We study whether test stakes affect the gender gap by exploiting the reform described in Section 2.3, which removed the contribution of the grade-8 INVALSI score to the lower-secondary exit-exam grade starting in 2017/18. We therefore treat pre-reform grade-8 INVALSI tests as *high-stakes* and post-reform tests as *low-stakes*.[19] Our empirical strategy compares the female–male performance gap across the high- and low-stakes regimes using a gender-by-reform difference-in-differences specification similar in spirit to Arenas and Calsamiglia (2025). Because the reform applied to all students, boys do not constitute an untreated control group. Rather, identification comes from the differential change in performance by gender: the first difference is between the high- and low-stakes regimes,

[19] A similar definition of stakes is used in Azmat et al. (2016) and Montolio and Taberner (2021).

and the second is between girls and boys. In particular, we estimate:

$$Y_{it} = \alpha + \beta_1 Female_i + \beta_2 HighStake_t + \gamma\big(Female_i \times HighStake_t\big) + C_{it}'\Delta + \varepsilon_{it}, \quad (1)$$

where Y_{it} is student i's standardized average INVALSI score in year t, $Female_i$ identifies girls, $HighStake_t$ equals one in the pre-reform period, and C_{it} includes student controls.[20]

The coefficient of interest is γ, which measures the difference between the female–male performance gap in the high-stakes regime and the corresponding gap in the low-stakes regime. A positive coefficient indicates that girls perform relatively better when the INVALSI score contributes to the exit-exam grade. The causal interpretation requires that, in the absence of the reform, the female–male gap would not have changed systematically at the reform date because of other contemporaneous factors. We assess the plausibility of this assumption through prereform trends and the validation exercises described below. Moreover, compositional changes in test participation are unlikely to threaten identification: lower-secondary education is compulsory, the INVALSI assessment covers the full population of grade-8 students, and participation—including attendance at a make-up session for students absent on the scheduled test day—is required both before and after the reform.

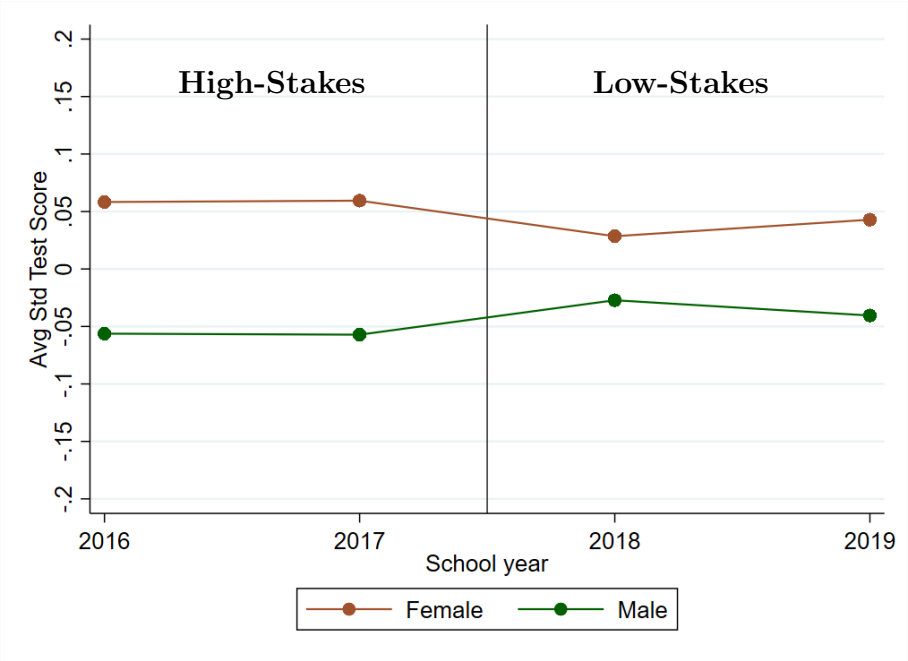
3.2 Main Results

We first report the baseline estimates and then turn to robustness and validation exercises. Our main analysis focuses on the four school years surrounding the reform, from 2015/16 to 2018/19. This narrow window maximizes comparability across cohorts and avoids contamination from the COVID-19 shock to learning outcomes; accordingly, we exclude the 2020/21 and 2021/22 school years.[21] Figure A1 and Table A1 in the Appendix reports results for the full available 2014/15–2021/22 sample and shows that the main conclusions are not driven by the choice of the narrower window.

[20] Individual controls include students' origin (i.e., native, second-generation immigrant or first-generation immigrant), academic regularity (i.e., whether they have repeated a year), and parents' education and occupation. In some specifications, we additionally control for the previous INVALSI test score (grade 5). This variable is available only for the subset of students who can be linked longitudinally across grades.

[21] The INVALSI assessment was not administered in school year 2019/20 because of the COVID-19 pandemic. We do not include 2014/15 in the baseline window because grade-8 students in that cohort can be linked only imperfectly to their grade-5 records. In that year, the longitudinal identifier is unavailable for an unusually large number of students, making specifications that control for prior achievement less reliable. We therefore use the same four-year sample across all baseline specifications.

Figure 1: Temporal evolution of the outcome



Notes: Figure 1 illustrates the evolution of average test scores by gender from 2016 to 2019. In 2017, male students had an average score of -0.057, while female students scored 0.059. Following the reform, average scores for males and females were -0.027 and 0.029 in 2018, and -0.040 and 0.043 in 2019, respectively. To ensure comparability across years, the average score is standardized at year level to have mean equal to 0 and standard deviation equal to 1.

Figure 1 plots average standardized scores by gender around the reform. Two facts stand out. First, gender differences are modest throughout the period. Second, girls outperform boys both before and after the reform. In the pre-reform high-stakes period, the female advantage is about 0.10 standard deviations; in the post-reform low-stakes period, it narrows to roughly 0.07 standard deviations.[22]

Table 1 reports the corresponding OLS estimates based on Equation 1. Across all specifications, the interaction coefficient γ is positive, precisely estimated, and modest in magnitude, ranging from about 0.03 to 0.05 standard deviations. The estimated difference between the high- and low-stakes regimes is therefore modest: girls' relative advantage is slightly larger when the INVALSI score contributes to the exit-exam grade. Thus, we find no evidence that higher test stakes adversely affect girls' performance relative to boys in early adolescence.[23]

These results show that the relative female losses documented in several high-stakes

[22] A gap of 0.10 standard deviations corresponds to roughly one third of a school year of learning (Woessmann, 2016).

[23] These estimates are stable to the inclusion of school fixed effects, individual controls such as parental education, migration background, and grade-retention history, and previous performance.

Table 1: Main Analysis

	(1)	(2)	(3)
Female × High-Stakes	0.046***	0.048***	0.028***
	(0.003)	(0.003)	(0.003)
Female	0.069***	0.044***	0.057***
	(0.002)	(0.002)	(0.002)
High-Stakes	-0.023***	-0.015***	-0.039***
	(0.005)	(0.005)	(0.005)
Controls	No	Yes	Yes
School FE	No	Yes	Yes
Score G5	No	No	Yes
Observations	2,143,036	2,143,036	1,666,568
R^2	0.002	0.205	0.480

Notes: The dependent variable Y_{it} is the average test score of student i at time t, obtained in the Italian and Mathematics sections of the INVALSI test. The average score is standardized at the year level to have mean equal to 0 and standard deviation equal to 1. Control variables include students' origin and regularity in their studies (i.e., whether they have repeated a grade), parental education and occupation. Standard errors, clustered at school level, are in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

settings involving older students are not present in this national setting at ages 13–14. More broadly, they indicate that relative female losses under high stakes are not a general feature of early adolescence, suggesting that gender differences in responses to test stakes may not yet be systematically established at this age. The findings also complement the small literature on incentives earlier in the educational cycle: when the standardized assessment itself carries direct consequences, rather than merely being administered in a grade that is consequential for tracking or progression, we still find no evidence that higher stakes disadvantage girls. Finally, Appendix Table A1 shows that the same conclusion holds in the full 2014/15–2021/22 sample of more than 3.7 million students.

3.3 Robustness checks and discussion

We assess the robustness of the baseline findings using three complementary empirical exercises. First, we use grade-10 placebo comparisons and evidence on grade-8 teachers' perceptions to assess whether the estimates partly reflect the contemporaneous transition from paper- to computer-based testing. Second, we address concerns about manipulation and cheating using a representative subsample of externally monitored classrooms. Third, we exploit within-student variation across grades 5, 8, and 10 to compare the same students across assessments that differ in stakes. We next examine whether the aggregate results mask heterogeneity between Italian and mathematics. Finally, we discuss

how gender differences in effort across testing regimes may affect the interpretation of our estimates. Overall, the evidence presented in this section leaves the main conclusion unchanged: we find no evidence that higher stakes reduce girls' performance relative to boys.

Table 2: Robustness Checks

	(1)	(2)	(3)	(4)
Female × High-Stakes		0.032***	0.065***	0.065***
		(0.012)	(0.003)	(0.002)
Female × Pre-Reform	0.025***			
	(0.005)			
Female	-0.038***	0.048***	-0.000	
	(0.004)	(0.008)	(0.002)	
High-Stakes		-0.003	0.051***	0.051***
		(0.013)	(0.002)	(0.001)
Pre-Reform	-0.047***			
	(0.005)			
Controls	Yes	Yes	Yes	No
School FE	Yes	No	No	No
Student FE	No	No	No	Yes
Observations	1,727,633	114,245	2,032,380	2,032,380
R^2	0.432	0.149	0.075	0.741

Notes: The dependent variable Y_{it} is the average test score of student i at time t, obtained in the Italian and Mathematics sections of the INVALSI test. The average score is standardized at the year level to have mean equal to 0 and standard deviation equal to 1. Control variables include students' origin and regularity in their studies (i.e., whether they have repeated a grade), parental education and occupation. Column (1) reports placebo results for grade 10, which help assess whether the estimates reflect other reform-period shifts. Column (2) reports estimates using the representative subsample of externally monitored classrooms to address concerns regarding manipulation or cheating. Columns (3) and (4) exploit within-student variation across grades 5, 8, and 10 using a longitudinal panel, with Column (4) including student fixed effects to absorb time-invariant student heterogeneity. Standard errors, clustered at the school level in Columns (1)–(2) and robust to heteroskedasticity in Columns (3)–(4), are reported in parentheses.* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Transition to Computer-Based Testing. The reform coincided with a shift in the grade-8 INVALSI assessment from paper-based to computer-based administration, raising the possibility that the baseline estimates partly capture a gender-specific response to the change in test mode. To assess this possibility, we estimate an analogous specification for grade-10 students, who experienced the same transition to computer-based testing but whose INVALSI scores did not contribute to their grades and therefore remained

low-stakes throughout.[24] Specifically, we estimate:

$$Y_{it} = \alpha + \beta_1 Female_i + \beta_2 PreReform_t + \gamma(Female_i \times PreReform_t) + C_{it}'\Delta + \varepsilon_{it}, \quad (2)$$

where $PreReform_t$ equals one in the years before the transition to computer-based testing. Column (1) of Table 2 reports a small positive interaction of 0.025 standard deviations. This estimate suggests that the change in test mode may account for part of the baseline grade-8 interaction, but does not eliminate it.[25]

We also use the INVALSI Teacher Questionnaire to examine whether grade-8 teachers' perceptions of computer-based testing vary with classroom gender composition. We focus on 2017/18, pooling responses from Italian and mathematics teachers, because three relevant questions were available in that year. These asked whether computer-based testing made test questions easier to understand, whether the testing platform was easy for students to use, and whether students' limited familiarity with the computer keyboard slowed test completion.[26] For each question, we construct an indicator equal to one if the teacher agreed or strongly agreed. If the new format disadvantaged girls relative to boys, teachers in classes with more girls might report greater difficulties with computer-based testing. Table A2 provides no evidence of such a pattern: all three associations are statistically insignificant, with point estimates indicating slightly more positive assessments of comprehension and ease of use, and fewer reported keyboard difficulties, in classes with more girls. Although this evidence is indirect, it complements the grade-10 placebo results in suggesting that the transition to computer-based testing is unlikely to be the main driver of our findings.

Externally monitored classrooms. Previous research documents manipulation of INVALSI scores and shows that external monitoring substantially reduces it (Angrist, Joshua D. and Battistin, Erich and Vuri, Daniela, 2017). A potential concern for our analysis is that the extent or nature of manipulation may have differed between the high- and low-stakes regimes in ways that affected girls' and boys' recorded scores differently. For example, if girls benefited relatively more from teacher assistance or other irregularities under the high-stakes regime, the positive interaction coefficient would overstate their genuine relative performance under higher stakes. In addition to the full-population data, INVALSI provides test scores for a representative sample of externally monitored class-

[24] This exercise relies on the additional assumption that the gender-specific effect of moving from paper-based to computer-based testing is comparable in grades 8 and 10.

[25] Under the assumption that the gender-specific effect of the paper-to-computer transition is the same in grades 8 and 10, a back-of-the-envelope adjustment would reduce the baseline estimate of 0.048 in column (2) of Table 1 to 0.023 standard deviations.

[26] Only the question on whether computer-based testing made test questions easier to understand was also included in the 2018/19 questionnaire. Results for that year are fully aligned with those reported here.

rooms selected through a two-stage sampling procedure.[27] Because test administration in these classrooms is overseen by an external observer, concerns about manipulation and cheating are substantially reduced. Column (2) of Table 2 reports estimates of Equation 1 for this subsample. The interaction coefficient remains positive and modest in magnitude, although somewhat smaller than in the full sample, suggesting that manipulation is unlikely to drive the baseline result.

Within-student comparisons. We exploit the longitudinal structure of the data to compare the same students across grades 5, 8, and 10. For students who reached grade 8 before the reform, the grade-8 INVALSI assessment was high-stakes, whereas their grade-5 and grade-10 assessments were low-stakes. We therefore restrict the sample to students who took the grade-8 assessment in 2015/16 or 2016/17 and link their records across grades using the SIDI INVALSI identifier. This yields a three-wave panel of 677,460 students, corresponding to approximately 65% of the relevant grade-8 population.

We estimate both pooled specifications analogous to Equation 1 and a student fixed-effects model of the form:

$$Y_{it} = \alpha_i + \beta HighStake_t + \gamma\big(Female_i \times HighStake_t\big) + \varepsilon_{it}, \tag{3}$$

where α_i absorbs time-invariant student heterogeneity and $HighStake_t$ equals one for the grade-8 observation and zero for the grade-5 and grade-10 observations. The results, reported in Columns (3)–(4) of Table 2, are positive and stable across specifications. The interaction coefficient is 0.065 standard deviations and remains stable after the inclusion of student fixed effects, indicating that the within-student pattern is not driven by student composition or other time-invariant characteristics. Because stakes and grade level vary together in this comparison, however, this exercise complements rather than replaces the reform-based design. Its results nevertheless closely mirror the baseline estimates.

Heterogeneity by subject. We next examine whether the aggregate result masks heterogeneity between Italian and mathematics, a natural dimension given the well-documented gender differences in achievement across language and mathematical skills (OECD, 2023, 2024). These differences are also visible in our data: as shown by the *Female* coefficients in Table 3, girls perform better in Italian, whereas boys have a small advantage in mathematics. The interaction estimates show that the modest positive effect for the aggregate score is driven primarily by Italian. In this subject, the interaction between *Female* and *HighStake* is positive and statistically significant, ranging from approximately 0.06 to 0.08 standard deviations. In mathematics, by contrast, the interaction

[27] Schools are first sampled, and, usually, two classes are then sampled and monitored within each selected school. The resulting sample is representative at the national level.

is close to zero and not statistically significant at conventional levels. Thus, higher stakes do not reduce girls' performance relative to boys in either subject, while the modest relative female advantage in the aggregate score is concentrated in Italian.

Table 3: Analysis by Subject

Panel A: Italian

	(1)	(2)	(3)
Female × High-Stakes	0.082***	0.083***	0.057***
	(0.003)	(0.003)	(0.002)
Female	0.223***	0.198***	0.147***
	(0.002)	(0.002)	(0.002)
High-Stakes	-0.049***	-0.043***	-0.054***
	(0.004)	(0.004)	(0.004)
Controls	No	Yes	Yes
School FE	No	Yes	Yes
Score G5	No	No	Yes
Observations	2,143,036	2,143,036	1,718,107
R^2	0.018	0.200	0.418

Panel B: Mathematics

	(1)	(2)	(3)
Female × High-Stakes	-0.000	0.001	-0.004
	(0.003)	(0.003)	(0.003)
Female	-0.094***	-0.116***	-0.049***
	(0.002)	(0.002)	(0.002)
High-Stakes	0.008	0.016***	-0.011**
	(0.005)	(0.005)	(0.005)
Controls	No	Yes	Yes
School FE	No	Yes	Yes
Score G5	No	No	Yes
Observations	2,143,036	2,143,036	1,733,513
R^2	0.002	0.173	0.393

Notes: In Panels A and B, the dependent variable Y_{it} represents the test score of student i at time t for the Italian and Mathematics sections of the INVALSI test, respectively. These scores are standardized at the year level to have mean equal to 0 and standard deviation equal to 1. Control variables include students' origin and regularity in their studies (i.e., whether they have repeated a grade), parental education and occupation. Standard errors, clustered at school level, are in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Effort as a potential mechanism. A potential issue in interpreting our estimates is that changes in stakes may affect how much effort students exert as well as how they perform under pressure. If effort responds differently by gender, part of the estimated

change in the gender gap may reflect differential engagement across testing regimes. This distinction concerns mechanisms rather than identification: effort is itself one channel through which stakes may affect observed performance, and our estimates capture the overall effect but cannot disentangle effort from responses to test pressure or other mechanisms. We can nevertheless draw on previous research to assess whether differential effort could plausibly account for the direction of our findings. Evidence from students of comparable ages generally indicates that boys exert less effort than girls in low-stakes settings. Most directly relevant to our context, Anaya and Zamarro (2024) document this pattern among 15-year-old PISA test takers, who are only slightly older than the students in our sample. Soland (2018) finds a similar gender difference among students in grades 5, 7, and 9.[28] If the same pattern holds in our setting, boys' relatively lower effort under the low-stakes regime would make the gender gap more favorable to girls in that regime, thereby attenuating the positive interaction between *Female* and *HighStake*. Differential effort of the kind documented in the literature would therefore work against our finding rather than generate it.

Taken together, the robustness and additional analyses reinforce the central conclusion of the paper. Although the transition to computer-based testing may account for part of the positive baseline interaction, we find no evidence that higher stakes reduce girls' performance relative to boys in early adolescence; if anything, the estimates point to a small relative female advantage, concentrated in Italian.

4 Conclusions

This paper studies whether high-stakes testing affects the performance of female and male students differently in early adolescence. We exploit a reform that removed the grade-8 INVALSI score entirely from the calculation of Italy's lower-secondary exit-exam grade while preserving participation in the assessment as a requirement for admission to the exit exam. Using administrative data covering the full population of students, we compare the female–male performance gap across the high- and low-stakes regimes. We find no evidence that removing the test's direct grade consequences improves girls' relative performance. Girls outperform boys in both regimes, and their advantage is modestly larger—by approximately 0.03 to 0.05 standard deviations across the main specifications—when the INVALSI score contributes to the exit-exam grade.

This conclusion is robust across a range of validation exercises. Grade-10 placebo comparisons, complemented by evidence from grade-8 teacher questionnaires, support

[28] Lower effort among male students in low-stakes settings is also documented among GRE test takers (Schlosser et al., 2019).

the robustness of the main conclusion to the contemporaneous transition from paper- to computer-based testing. The baseline pattern is also replicated in the representative subsample of externally monitored classrooms and closely mirrored by within-student specifications comparing the same students across grades 5, 8, and 10. Finally, the modest positive interaction in the aggregate score is driven primarily by Italian, whereas the corresponding estimates for mathematics are close to zero. Across all exercises, we find no evidence that higher stakes adversely affect girls' performance relative to boys at ages 13–14.

The paper contributes population-wide quasi-experimental evidence on gender differences in responses to test stakes at an educational stage for which the existing evidence remains limited. Several studies involving older students find that female performance deteriorates relative to male performance in more consequential or competitive testing environments. Our findings show that such relative female losses are not a general feature of early adolescence.

These findings matter for the interpretation of standardized test scores and the design of assessment systems. If students from different groups respond differently to the incentives and pressure attached to an examination, observed achievement gaps may partly reflect behavioral responses to the testing environment rather than differences in underlying competence. In our setting, however, attaching direct grade consequences to a standardized assessment does not disadvantage girls. More broadly, the results suggest that gender differences in responses to high-stakes testing may depend on the stage of schooling and on the institutional design of the assessment. An important direction for future research is therefore to identify when and why gender differences in responses to high-stakes testing emerge, rather than assuming that patterns documented among older students necessarily extend to children at earlier stages of education.

Declaration of generative AI and AI-assisted technologies in the writing process During the preparation of this manuscript in 2026, the authors used ChatGPT (OpenAI) for proofreading and to improve the flow and readability of the text. All AI-generated suggestions were reviewed, revised, and approved by the authors, who take full responsibility for the accuracy and integrity of the work.

References

Anaya, L. M. and G. Zamarro (2024, 04). The role of student effort on performance in PISA: revisiting the gender gap in achievement. *Oxford Economic Papers 76*(2), 533–560.

Angelo, C. and A. Reis (2021). Gender gaps in different grading systems. *Education Economics 29*(1), 105–119.

Angrist, Joshua D. and Battistin, Erich and Vuri, Daniela (2017). In a small moment: Class size and moral hazard in the italian mezzogiorno. *American Economic Journal: Applied Economics 9*(4), 216–249.

Arenas, A. and C. Calsamiglia (2025). Gender differences in high-stakes performance and college admission policies. *Management Science 71*(10), 8413–8429.

Azmat, G., C. Calsamiglia, and N. Iriberri (2016, 08). Gender Differences in Response to Big Stakes. *Journal of the European Economic Association 14*(6), 1372–1400.

Bach, M. and M. Fischer (2020). Understanding the response to high-stakes incentives in primary education. IZA Discussion Papers 13845, Institute of Labor Economics (IZA).

Brunello, G. and D. Kiss (2022). Math scores in high stakes grades. *Economics of Education Review 87*, 102219.

Cai, X., Y. Lu, J. Pan, and S. Zhong (2019, 05). Gender Gap under Pressure: Evidence from China's National College Entrance Examination. *The Review of Economics and Statistics 101*(2), 249–263.

Calsamiglia, C. and A. Loviglio (2019). Grading on a curve: When having good peers is not good. *Economics of Education Review 73*, 101916.

Carlana, M. (2019, 08). Implicit stereotypes: Evidence from teachers' gender bias. *The Quarterly Journal of Economics 134*(3), 1163–1224.

Carlana, M., E. La Ferrara, and P. Pinotti (2022). Implicit stereotypes in teachers' track recommendations. *AEA Papers and Proceedings 112*, 409–414.

Cárdenas, J.-C., A. Dreber, E. von Essen, and E. Ranehill (2012). Gender differences in competitiveness and risk taking: Comparing children in Colombia and Sweden. *Journal of Economic Behavior & Organization 83*(1), 11–23. Gender Differences in Risk Aversion and Competition.

De Paola, M. and F. Gioia (2016). Who performs better under time pressure? results from a field experiment. *Journal of Economic Psychology 53*, 37–53.

Dreber, A., E. von Essen, and E. Ranehill (2011). Outrunning the gender gap—boys and girls compete equally. *Experimental Economics 14*(4), 567–582.

Dreber, A., E. von Essen, and E. Ranehill (2014). Gender and competition in adolescence: task matters. *Experimental Economics 17*(1), 154–172.

Hanushek, E. A., L. Kinne, P. Sancassani, and L. Woessmann (2026). Patience and Subnational Differences in Human Capital: Regional Analysis with Facebook Interests. *The Economic Journal 136*(673), 335–350.

Iriberri, N. and P. Rey-Biel (2019). Competitive pressure widens the gender gap in performance: Evidence from a two-stage competition in mathematics. *The Economic Journal 129*(620), 1863–1893.

Jurajda, S. and D. Munich (2011, May). Gender gap in performance under competitive pressure: Admissions to czech universities. *American Economic Review 101*(3), 514–18.

Lavy, V. (2008). Do gender stereotypes reduce girls' or boys' human capital outcomes? evidence from a natural experiment. *Journal of Public Economics 92*(10), 2083–2105.

Martini, A. and M. Papini (2015). Che cosa ne pensano gli insegnanti delle prove invalsi. Working Paper 24/2015, INVALSI.

Montolio, D. and P. A. Taberner (2021). Gender differences under test pressure and their impact on academic performance: A quasi-experimental design. *Journal of Economic Behavior & Organization 191*, 1065–1090.

OECD (2023). *Gender, Education and Skills: The Persistence of Gender Gaps in Education and Skills*. Paris: OECD Publishing.

OECD (2024). *PISA 2022 Results*. Paris: OECD Publishing.

Ors, E., F. Palomino, and E. Peyrache (2013). Performance gender gap: Does competition matter? *Journal of Labor Economics 31*(3), 443–499.

Saygin, P. O. (2020, August). Gender bias in standardized tests: evidence from a centralized college admissions system. *Empirical Economics 59*(2), 1037–1065.

Schlosser, A., Z. Neeman, and Y. Attali (2019, 05). Differential Performance in High Versus Low Stakes Tests: Evidence from the Gre Test. *The Economic Journal 129*(623), 2916–2948.

Soland, J. (2018). Are achievement gap estimates biased by differential student test effort? putting an important policy metric to the test. *Teachers College Record 120*(12), 1–26.

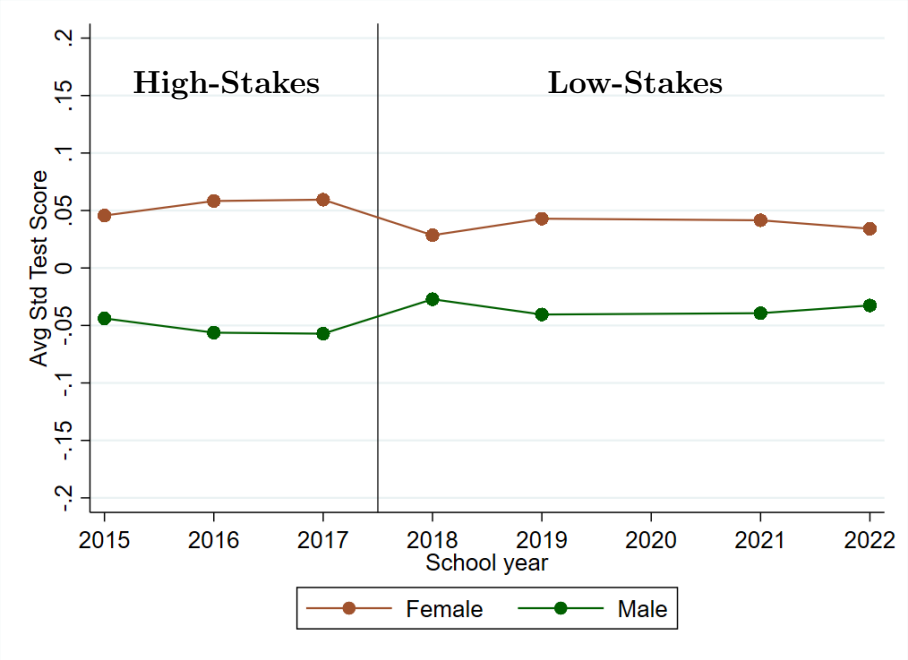
Sutter, M. and D. Glätzle-Rützler (2015). Gender differences in the willingness to compete emerge early in life and persist. *Management Science 61*(10), 2339–2354.

Terrier, C. (2020). Boys lag behind: How teachers' gender biases affect student achievement. *Economics of Education Review 77*, 101981.

Woessmann, L. (2016, September). The importance of school systems: Evidence from international differences in student achievement. *Journal of Economic Perspectives 30*(3), 3–32.

Appendix

Figure A1: Temporal evolution of the outcome – Full Sample (2015-2022)



Notes: Figure A1 illustrates the evolution of average test scores by gender from 2015 to 2022 (excluding 2020, when the test was not administered). In 2017, male students had an average score of -0.057, while female students scored 0.059. Following the reform, average scores for males and females were -0.027 and 0.029 in 2018, and -0.040 and 0.043 in 2019, respectively. To ensure comparability across years, the average score is standardized at year level to have mean equal to 0 and standard deviation equal to 1.

Table A1: Full Sample (2015-2022)

	(1)	(2)	(3)
Female × High-Stakes	0.035***	0.032***	0.025***
	(0.002)	(0.002)	(0.002)
Female	0.072***	0.050***	0.048***
	(0.002)	(0.001)	(0.001)
High-Stakes	-0.018***	0.014***	-0.021***
	(0.005)	(0.004)	(0.004)
Controls	No	Yes	Yes
School FE	No	Yes	Yes
Score G5	No	No	Yes
Observations	3,717,584	3,717,584	2,804,499
R^2	0.002	0.198	0.490

Notes: The dependent variable Y_{it} is the average test score of student i at time t, obtained in the Italian and Mathematics sections of the INVALSI test. The average score is standardized at the year level to have mean equal to 0 and standard deviation equal to 1. Control variables include students' origin and regularity in their studies (i.e., whether they have repeated a grade), parental education and occupation. Standard errors, clustered at school level, are in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A2: Gender Composition and Teacher Perceptions of Computer-Based Assessment

	(1)	(2)	(3)
Share Female	0.046	0.027	-0.006
	(0.099)	(0.053)	(0.088)
Mathematics	-0.014	0.002	0.020
	(0.023)	(0.013)	(0.020)
Constant	0.402***	0.896***	0.273***
	(0.051)	(0.027)	(0.045)
Observations	1,904	2,021	2,013
R^2	0.0003	0.0001	0.0005

Notes: Regression results are based on the INVALSI Teacher Questionnaire administered in 2018. In all specifications, the dependent variable is a binary indicator equal to 1 if the teacher "agrees" or "strongly agrees" with the statement, and 0 otherwise. Columns (1)–(3) refer, respectively, to the statements: "The transition from paper-based to computer-based testing facilitated question comprehension"; "The assessment platform is easy for students to use"; and "Students' lack of familiarity with the PC keyboard slowed down the execution of the test." Italian is the reference category for subject fixed effect. Robust standard errors are reported in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$