ISSN 2282-6483

Alma Mater Studiorum - Università di Bologna
DEPARTMENT OF ECONOMICS

Do High-Stakes Exams Widen Gender Gaps in Early Adolescence? Evidence from a Middle-School Exit Exam Reform

Annalisa Loviglio
Veronica Rattini
Federico Stronati

Quaderni - Working Paper DSE N°1223

Do High-Stakes Exams Widen Gender Gaps in Early Adolescence? Evidence from a Middle-School Exit Exam Reform

Annalisa Loviglio[*] Veronica Rattini[†] Federico Stronati[‡]

Abstract

Do high-stakes exams widen gender gaps in academic performance? We study this question in early adolescence using Italy's lower-secondary exit exam. A 2017 reform removed the national standardized test from the final grade while preserving the requirement to take the same test, reducing stakes without changing content. Using administrative data on the full population of students and a difference-in-differences design, we compare gender gaps before and after the reform. We find no evidence that higher stakes disadvantage girls. If anything, girls perform slightly better relative to boys when the test has exam stakes, suggesting gender differences under pressure emerge later in education.

Keywords: high-stakes testing, gender gaps, standardized exams, adolescence.

JEL codes: I21, I24, J16.

[*]University of Bologna, Department of Economics, Piazza Scaravilli 2, 40126 Bologna, Italy, and IZA - Institute of Labor Economics. E-mail: annalisa.loviglio@unibo.it

[†]University of Bologna, Department of Economics, Piazza Scaravilli 2, 40126 Bologna, Italy, and IZA - Institute of Labor Economics. E-mail: veronica.rattini2@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).

[‡]University College Dublin, School of Economics, Belfield, Dublin 4, Ireland; e-mail: federico.stronati@ucdconnect.ie

1 Non-technical summary

This paper asks whether high-stakes exams make gender gaps in academic performance worse during early adolescence. The authors study this question using an education reform in Italy. Before 2017/18, Italian eighth-grade students took the national INVALSI standardized test as part of their lower-secondary school exit exam, and the INVALSI score contributed to their final graduation mark. After the reform, students still had to take the same INVALSI test, but the score no longer counted toward the final exam grade. This created a useful comparison: the test itself stayed largely the same, but its importance for students changed sharply.

The main concern the paper addresses is that high-pressure exams may not measure only what students know. They may also reflect how students respond to stress, competition, and pressure. Previous research, especially on older students in high school, university, or college-admission settings, has often found that women perform relatively worse than men when exams are very high-stakes. This paper asks whether that pattern already exists at ages 13–14, when Italian students complete lower-secondary school.

Using administrative data covering the full population of students, the authors compare girls' and boys' test performance before and after the reform. They use a statistical approach that looks at whether the female–male performance gap changed when the INVALSI test became lower-stakes. Their main finding is clear: there is no evidence that high-stakes testing disadvantaged girls at this age. Girls performed slightly better than boys both before and after the reform. In fact, if anything, girls' relative advantage was a little larger when the test counted toward the final exam grade, although the size of this difference was small.

The authors run several checks to make sure the finding is not driven by other factors. They repeat the analysis using classrooms monitored by external observers, where cheating or manipulation is less likely. They also compare nearby grades to account for other changes that happened around the same time, such as the move from paper-based to computer-based testing. In addition, they follow the same students across different grades to compare their performance in lower- and higher-stakes settings. These checks all point to the same conclusion: higher stakes did not widen the gender gap against girls in early adolescence.

The paper also looks separately at Italian and mathematics. Girls had a clear advantage in Italian, while boys had a small advantage in mathematics. The slight overall benefit for girls under higher stakes appears to come mainly from the Italian section, not from math. In mathematics, high stakes did not meaningfully change the gender gap.

Overall, the paper suggests that gender differences in responses to exam pressure are not fixed from an early age. The female disadvantage seen in some later high-stakes

settings does not appear at the end of middle school in Italy. This matters for education policy because it shows that the effects of high-stakes testing may depend on students' age and stage of schooling. It also suggests that later gender gaps under pressure may develop through later social, educational, or institutional experiences rather than being an inherent difference in how girls and boys perform on important tests.

2 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*.[2] Because these assessments often govern access to higher education and, indirectly, to later labor-market opportunities, they are expected to provide an accurate measure of the competencies they are designed to capture. Yet when stakes are high, observed performance may reflect not only knowledge and skill, but also individuals' responses to pressure, anxiety, and competition. A large literature studies whether performance in such environments reflects underlying ability or, instead, differential behavioral responses to the incentives and stress induced by the testing situation. A prominent result is that, in late adolescence and early adulthood, female students often perform relatively worse than male students when stakes are high (Azmat et al., 2016; Cai et al., 2019; Schlosser et al., 2019; Arenas and Calsamiglia, 2025). Yet we know much less about when in the life cycle this pattern emerges.

This paper studies whether high-stakes testing widens the gender gap already in early adolescence. We focus on Italy's lower-secondary exit exam, the first national exam students face and a salient milestone in the transition from middle school to upper secondary education.[3] This exam is relevant not only because it marks the completion of compulsory lower-secondary schooling, but also because it is closely linked to students' subsequent educational trajectories. Although upper-secondary track assignment is not formally based on the exam score, track choice is strongly associated with prior achievement, teacher recommendations, and family background (Carlana et al., 2022a).[4] Our analysis exploits the fact that, until 2016/17, the grade-8 INVALSI standardized test score entered the lower-secondary exit-exam grade.[5] Starting in 2017/18, a reform removed this contribution while preserving participation in the test as a requirement for admission to the exit exam.[6] The reform therefore generated a sharp reduction in the stakes attached to the

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

[2] In some countries, admission depends almost entirely on standardized test scores, as in China, South Korea, India, and Brazil. In others, test scores are combined with additional measures of prior preparation, such as school-leaving grades, as in Spain, Australia, Denmark, Norway, and Sweden, or with institution-specific admission criteria, as in the United States and the United Kingdom.

[3] The lower-secondary exit exam is called *Esame di Stato conclusivo del primo ciclo*.

[4] In addition, exam results are publicly recorded and transmitted to the receiving upper-secondary school, which makes the exam socially and educationally salient.

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

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

standardized test while leaving the test in place. This institutional change provides a useful source of quasi-experimental variation to study how the gender gap responds when the incentives attached to a common standardized assessment are weakened.

Using administrative data on the full population of students, we estimate a difference-in-differences model that compares the female–male performance gap before and after the reform. We find no evidence that lowering test stakes meaningfully reduces the gender gap in favor of girls. Girls outperform boys both before and after the reform and, if anything, their relative advantage is slightly larger when the test carries exam stakes. 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, if anything slightly favorable to girls under high stakes.

This result is robust across a range of complementary exercises. First, it is replicated in the representative subsample of externally monitored classrooms, where concerns about manipulation and cheating are substantially reduced. Second, it remains qualitatively unchanged in placebo analyses using adjacent grades, which help isolate the effect of stakes from other contemporaneous changes in the testing environment. In particular, because the reform coincided with a transition from paper-based to computer-based administration in grade 8, we compare grade-8 students to nearby cohorts that experienced the same change in test mode but no change in stakes. Third, the result is confirmed in within-student specifications that compare the same students across grades 5, 8, and 10, thereby leveraging the longitudinal structure of the data while absorbing time-invariant individual heterogeneity. Taken together, these exercises point to the same conclusion: we find no evidence that higher stakes widen the gender gap against girls in early adolescence.

The paper contributes to the literature in two main ways. First, it provides new evidence on when gender differences in performance under pressure emerge. To the best of our knowledge, this is the first paper to provide causal evidence on gender differences in a genuinely high-stakes standardized exam in early adolescence. Existing evidence on female relative underperformance under pressure is concentrated in high-school, university, and college-admission settings (Azmat et al., 2016; Cai et al., 2019; Schlosser et al., 2019; Arenas and Calsamiglia, 2025; Iriberri and Rey-Biel, 2019; De Paola and Gioia, 2016; Montolio and Taberner, 2021). By contrast, the small literature on younger students studies a related but different question: whether students accumulate more skills in school years in which future educational opportunities depend more strongly on performance, using low-stakes standardized tests that do not themselves affect grades or progression (Bach and Fischer, 2020; Brunello and Kiss, 2022). That evidence is informative about learning and effort over the school year, but not about gender differences in performance

2017, n. 62).

on a single high-stakes exam. Our paper instead isolates how the gender gap changes when the same standardized test carries high stakes. Focusing on students aged 13–14, we show that the relative female losses observed in later educational stages are not yet present at the end of lower secondary school. This finding is important: if the female disadvantage found in later high-stakes environments does not emerge at age 13–14, it is less likely to reflect a fixed feature of performance under pressure, and more likely to arise later as developmental, social, and institutional forces become more salient (Ors et al., 2013; Pekkarinen, 2015; Iriberri and Rey-Biel, 2019; Saygin, 2019; Delaney and Devereux, 2021). More broadly, our evidence suggests that gender differences in responses to high-stakes assessment are not constant over the life cycle, but instead depend on the stage of schooling at which pressure is experienced.

Second, the paper makes a methodological contribution by bringing comparatively clean evidence on the role of stakes. A common challenge in this literature is that assessments compared across high- and low-stakes environments often differ simultaneously in format, timing, grading regime, or the population taking them (Azmat et al., 2016; Montolio and Taberner, 2021).[7] In our setting, the Italian reform changed whether the same national standardized test score entered the lower-secondary exit-exam grade, while leaving the test in place for the full population of students. This makes it possible to identify the effect of stakes *per se*, because the comparison is not across different assessments, institutions, or student groups, but across the same test before and after a change in stakes. The setting offers two further advantages for identifying gender differences. First, before upper secondary school, the curriculum is nationally homogeneous and students are not tracked, which reduces concerns that girls and boys are exposed to systematically different educational environments before the exam. Second, the data cover the universe of students, allowing us to compare girls and boys within the same schools and cohorts rather than across institutions with different student compositions. Finally, although the reform coincided with other changes in the testing environment, our validation exercises help isolate the role of stakes from these concurrent changes, allowing for a substantially cleaner interpretation of the resulting estimates as evidence on how test stakes shape the gender gap in performance.

These findings matter for both economics and education policy. From an economics perspective, they speak to whether standardized test scores can be interpreted as comparable signals of skill when incentives differ across testing environments. If pressure affects some groups more than others, observed achievement gaps may partly reflect responses to incentives rather than differences in underlying competence. From a policy perspective, the results suggest that the implications of high-stakes testing for gender inequality

[7] More broadly, comparisons across standardized tests, blind 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; Terrier, 2020; Calsamiglia and Loviglio, 2019).

may depend critically on the stage of schooling at which such assessments are introduced. Overall, the paper highlights the importance of studying not only whether test stakes matter on average, but also when in the life cycle, and for whom, they matter.

The remainder of the paper is organized as follows. Section 3 reviews the related literature. Section 4 describes the institutional setting, the reform, and the data. Section 5 presents the empirical design and the results, as well as robustness checks. Section 6 concludes.

3 Related Literature

This paper relates to three strands of literature. First, it contributes to the literature on gender differences in performance under pressure, competition, and high-stakes evaluation.[8] A large body of evidence from secondary school, university, and college-admission settings shows that female students often perform relatively worse than male students when evaluation is high-stakes or more competitive. For example, Azmat et al. (2016) show that, although girls outperform boys in all tests, they do so by relatively more in low-stakes assessments than in high-stakes exams in Spain. Saygin (2019) finds that girls perform relatively better in high-school GPA than in Turkey's centralized university entrance examination, and Iriberri and Rey-Biel (2019) show that the gender gap widens as competitive pressure increases in a two-stage mathematics competition. Evidence from admission settings points in the same direction: Cai et al. (2019) find that female students underperform on China's high-stakes *Gaokao* relative to low-stakes mock examinations,[9] while Jurajda and Munich (2011) document a male relative advantage in admission scores for selective university programs in the Czech Republic. Related evidence from university settings also points to gender differences in response to evaluative pressure and timing constraints (Ors et al., 2013; De Paola and Gioia, 2016; Montolio and Taberner, 2021). Comparisons between low- and high-stakes testing environments, including the GRE, similarly suggest that measured performance may reflect not only knowledge and skill, but also differential responses to incentives and pressure (Schlosser et al., 2019).[10] Closely related to our paper, Arenas and Calsamiglia (2025) exploit a reform in Spanish university admission exams that increased the weight of the centralized test and show that higher stakes negatively affect female scores, especially among top students.

Second, the paper relates to the methodological challenge of isolating the effect of stakes from other features of the assessment environment. A recurring difficulty in this

[8] Delaney and Devereux (2021) provides a broad overview of the literature on gender and educational achievement.

[9] *Gaokao* is China's national undergraduate admission exam.

[10] The Graduate Record Examination (GRE) is a standardized test used in admissions to many graduate programs.

literature is that exams compared across high- and low-stakes settings often differ simultaneously in format, timing, grading regime, workload, or the population taking them. For instance, Azmat et al. (2016) compare assessments that vary not only in stakes, but also in format—multiple choice, open-question, and oral—and in timing within the school year.[11] Differences in measured performance may therefore reflect features of the assessment itself, rather than stakes alone. Our setting mitigates these concerns because the reform changed whether the same national standardized test score entered the lower-secondary exit-exam grade, while leaving the test in place for the full population of students. Before upper secondary school, the curriculum is homogeneous nationwide and students are not tracked, limiting concerns that girls and boys are exposed to systematically different educational environments before the exam. Since the data cover the universe of students, we can also compare girls and boys within the same schools and cohorts, reducing the scope for compositional differences to drive the results. Together with our validation exercises, these features allow for a comparatively clean interpretation of the role of stakes *per se*.

Third, our paper relates to the smaller literature on incentives and achievement earlier in the education cycle, while differing from it in a central respect. Existing papers for younger students, mainly from Germany, study whether students perform better in school years in which achievement has stronger consequences for subsequent educational opportunities. Bach and Fischer (2020) exploit variation in binding track recommendations across German states and show that stronger incentives raise primary-school achievement. Brunello and Kiss (2022) study primary- and lower-secondary-school students and find that math scores on low-stakes assessments are higher when those tests are taken in grades that are more consequential for track assignment or graduation. In both cases, however, achievement is measured using low-stakes standardized tests that do not themselves carry direct consequences for students. These papers therefore provide evidence on incentives for learning and skill accumulation over the school year, rather than on gender differences in performance when students sit a high-stakes exam under pressure. Our paper instead exploits a reform that changed the stakes attached to the same national standardized test. To the best of our knowledge, it is the first to provide causal evidence for younger students on high-stakes performance that is directly comparable to the literature on later adolescence and early adulthood.

[11] A related literature compares standardized external exams with teacher-assigned grades during the school year, but interpretation is not straightforward because these measures may capture different skills and may also differ in the scope for teacher discretion or grading bias (Lavy, 2008; Calsamiglia and Loviglio, 2019; Terrier, 2020; Angelo and Reis, 2021).

4 Institutional Setting and Data

Basic education in Italy starts at age 6 and consists of 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.[12] At the end of grade 8, students must pass a national exit exam to complete the first cycle of education and enroll in upper secondary school.

After lower secondary school, students choose among three broad upper-secondary tracks: vocational, technical, and academic (*licei*). Admission is not formally determined by prior academic performance. Middle-school teachers provide non-binding track recommendations, but actual choices are strongly associated with these recommendations as well as with prior achievement, family background, and other student characteristics (Carlana et al., 2022b). Upper secondary education lasts four or five years, although compulsory schooling ends at age 16. Grade repetition is possible at all stages but remains uncommon in basic education.[13] As a result, most students are 13 or 14 years old when they sit the grade-8 exit exam.[14]

4.1 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.[15] 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.[16] 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.[17]

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-

[12] Minor differences may arise in middle school with respect to the second European language taught, typically Spanish, French, or German.

[13] Using ISTAT data, we calculate average pre-COVID retention rates of 0.32% in primary school and 2.53% in lower secondary school. During the COVID period, grade retention was strongly discouraged and remained lower afterward.

[14] Students may repeat a grade because of insufficient attendance, failure to meet minimum learning standards, or failure of the national exam.

[15] INVALSI data have been increasingly used in international research in the economics of education (e.g., Angrist et al. (2017); Hanushek et al. (2026)).

[16] An English section was introduced later: in 2017/18 for grades 5 and 8, and in 2018/19 for grade 13. Since this study focuses on earlier cohorts, English scores are not used in the analysis.

[17] Students absent on the test day in grade 8 can take the exam in a later make-up session. In other grades, absent students are not required to retake the test.

ended questions. The Italian section assesses reading comprehension, lexical proficiency, and grammar, while the mathematics section evaluates numerical and logical skills. As detailed below, 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.

4.2 Grade 8 National 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.[18] Students first receive an admission grade based on their performance during the three years of middle school. They then take written exams in Italian, mathematics, and foreign languages, followed by an oral examination covering the broader curriculum. While the Ministry of Education sets the general framework, the 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 in the written and oral 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 exam students face, its outcome is reported to the receiving upper-secondary school and, by law, is publicly posted at the school entrance.[19] The exam is therefore both institutionally salient and socially visible.

Until school year 2016/17, the INVALSI test was one of the components entering the final graduation mark and accounted for more than 14% of the final score, making it a quantitatively meaningful component of the exit exam.[20] Starting in 2017/18, Legislative Decree n. 62/2017 removed the contribution of the INVALSI score to the final graduation mark, while preserving participation in the test as a prerequisite for admission to the exit exam.[21] This reform sharply reduced the stakes attached to the standardized test without eliminating the test itself or its formal role within the examination process.

[18] 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.

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

[20] For comparison, Arenas and Calsamiglia (2025) study a reform that increased the weight of a centralized admission test by 17 percentage points.

[21] Legislative Decree n. 62 of April 13, 2017 also introduced other changes that are not central to our analysis, including a higher weight for the admission grade and a lower weight for the language components.

A second 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 separate the role of stakes from these broader changes in the testing environment.

4.3 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.[22] 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.

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 merge grade-8 records with the same students' observations in grade 5 and, when needed, in grade 10. These longitudinal matches are less complete than the within-year subject match, but still cover most students.[23]

5 Empirical Strategy and Results

We study whether test stakes affect the gender gap by exploiting the reform described in Section 4.2, 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*.[24] Our empirical strategy compares how the female–male gap changed after the reform using a difference-

[22] 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-PISA and IEA-TIMSS.

[23] Matching with grade 5 alone is successful for 78% of the sample, and matching with both grade 5 and grade 10 for 67%. The lower match rate likely reflects 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.

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

in-differences design similar in spirit to Arenas and Calsamiglia (2025). 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.[25] The coefficient of interest is γ, which captures whether the gender gap differs between the high- and low-stakes regimes. Under the usual DiD assumptions, it identifies the causal effect of higher stakes on the gender gap in test performance.

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.[26] We also omit 2014/15 because of slightly lower data quality.[27]

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.[28]

Table 1 reports the corresponding OLS estimates based on Eq. (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 estimates therefore indicate that lowering test stakes does not meaningfully reduce the gender gap in favor of girls. If anything, girls perform slightly better relative to boys when the test carries exam stakes.[29]

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

[25] 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.

[26] The INVALSI assessment was not administered in school year 2019/20 because of the COVID-19 pandemic.

[27] Appendix Section A1 reports results for the full 2015–2022 sample and shows that the main conclusions are not driven by this sample restriction.

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

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

settings are not yet present at age 13–14. More broadly, they help locate more precisely the stage of the educational career at which gender differences in performance under pressure begin to emerge. They also complement the smaller literature on younger students by showing that, when the test itself becomes consequential rather than merely being taken in an incentive-relevant school year, there is still no evidence of a female disadvantage. Finally, Appendix Table A1 shows that the same conclusion holds in the full 2015–2022 sample of more than 3.7 million students.

5.1 Robustness

We assess the robustness of the baseline findings along three margins. First, we address concerns about manipulation and cheating using the representative subsample of externally monitored classrooms. Second, we use adjacent-grade placebo comparisons to gauge whether the estimates may partly reflect contemporaneous changes in the testing environment rather than changes in stakes alone. Third, we exploit within-student variation across grades 5, 8, and 10 to compare the same students across assessments that differ in stakes.

We begin with the representative subsample of externally monitored classrooms. In addition to the population census, INVALSI provides standardized test score data for a representative sample of monitored classrooms selected through a two-stage sampling procedure.[30] Because test administration in this sample is overseen by an external observer, concerns about manipulation or cheating are substantially reduced. Table 2, Column (1) reports estimates of equation Eq. (1) for this subsample. The interaction coefficient remains positive and small, although somewhat smaller than in the full sample. This suggests that cheating account for a limited share of the baseline estimates, and it does not alter the main conclusion: if anything, girls retain a slight relative advantage when the test carries higher stakes.

We next turn to adjacent-grade placebo exercises, which help assess whether the baseline estimates may partly reflect other changes introduced around the reform. Specifically, the reform introduced a transition from paper-based to computer-based administration in grade 8 starting in 2017/18. To gauge whether this change affected girls and boys differently, we estimate the same specification on grade-10 students, who experienced the same change in test mode but whose INVALSI test never contributed to the final grade

[30] Schools are first sampled, and, usually, two classes are then sampled and monitored within selected schools.

and therefore remained low-stakes throughout.[31] Specifically, we estimate:

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

where $PreReform_t$ equals one in the pre-reform years. Table 2, Column (2) reports a small positive interaction, around 0.02 standard deviations in the richer specifications. This suggests that the paper-to-computer transition may account for a limited share of the baseline grade-8 estimate, but the magnitude is modest and leaves the main conclusion unchanged: even after accounting for differences related to test mode, there is still no evidence that higher stakes widen the gender gap against girls. If anything, this exercise suggests that the effect of stakes on the gender gap is even closer to zero than the baseline estimates imply.[32] We also estimate equation 1 on grade-5 students. Since grade 5 experienced no change in either stakes or test mode, this exercise is informative about broader reform-period shifts unrelated to the grade-8 treatment itself. As shown in Table 2 Column (3), the estimated interaction is statistically significant but very small in magnitude. Taken together with the grade-10 placebo, this evidence supports the same overall conclusion: contemporaneous changes in the testing environment may account for a limited share of the estimated variation, but they do not overturn the main result.

Finally, we exploit the longitudinal structure of the data to compare the same students across grades 5, 8, and 10. For students observed before the reform, grade 8 is the only stage at which the INVALSI test was high-stakes, whereas the tests taken in grades 5 and 10 were low-stakes. We therefore restrict attention to students who took grade 8 before the reform, in 2015/16 or 2016/17, and match their records across grades using the SIDI INVALSI identifier. This yields a panel of 677,460 students, about 65% of the relevant population.[33]

We estimate both pooled specifications analogous to equation Eq. (1) and a student fixed-effects model of the form:

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

where α_i absorbs time-invariant student heterogeneity. The results, reported in Table 2 Columns (4)-(5), are again positive and stable across specifications: the interaction coefficient is about 0.06 standard deviations and remains precisely estimated with student fixed effects. This within-student evidence closely mirrors the baseline results and reinforces the main interpretation of the paper. Across all robustness exercises, we find

[31] This approach relies on the additional assumption that the effect of moving from paper-based to computer-based testing is comparable between 10^{th} and 8^{th} grade students.

[32] Under the assumption that the paper-to-computer effect is the same in grades 8 and 10, a back-of-the-envelope adjustment would reduce a baseline estimate of 0.05 to about 0.03 standard deviations.

[33] As in the main analysis, we exclude 2014/15 because of lower data quality.

no evidence that higher stakes widen the gender gap against girls in early adolescence. If anything, the estimated effect remains small and mildly favorable to girls.

As a final exercise, we examine whether the aggregate result masks meaningful heterogeneity between Italian and mathematics, a natural margin given well-documented gender differences in achievement across language and mathematical skills (OECD, 2024, 2023). The same pattern is 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. Moreover, the modest positive interaction in the aggregate score is driven primarily by the Italian component (Table 3). In Italian, the interaction between *Female* and *HighStake* is positive and statistically significant, with estimates ranging from about 0.06 to 0.08 standard deviations. In mathematics, by contrast, the corresponding interaction is close to zero and not statistically different from zero at conventional levels. Thus, higher stakes do not widen the gender gap against girls in either subject, and the small relative female advantage in the aggregate score appears to come mainly from language performance rather than mathematics. We interpret this evidence cautiously, since each subject-specific score accounts for only part (about 7%) of the high-stakes INVALSI component. Nonetheless, the subject-level results are consistent with the main findings.

Taken together, these results reinforce the central conclusion of the paper: at age 13–14, higher stakes do not generate the female relative losses observed in later high-stakes settings.

6 Conclusions

This paper studies whether high-stakes testing widens the gender gap in academic performance in early adolescence. Exploiting a reform that removed the contribution of the grade-8 INVALSI national standardized test from Italy's lower-secondary exit-exam grade, we compare the female–male performance gap before and after this sharp reduction in test stakes. Using administrative data on the universe of students, we find no evidence that lower stakes improve girls' relative performance. If anything, girls retain a small relative advantage when the test carries exam stakes, although the effect is modest.

This conclusion is robust across a range of checks. The baseline result is replicated in the sample of externally monitored classrooms, remains qualitatively unchanged in adjacent-grade placebo analyses, and is confirmed in within-student specifications comparing the same students across grades 5, 8, and 10. We also show that the modest aggregate effect is driven primarily by the Italian component of the test, while the corresponding estimates in mathematics are close to zero. Taken together, these results point to a clear message: at age 13–14, higher stakes do not generate the female relative losses documented in later high-stakes settings.

More broadly, the paper contributes to the literature by providing, to the best of our knowledge, the first causal evidence on gender differences in performance in a genuinely high-stakes standardized exam in early adolescence. Existing evidence from high school, university, and college-admission settings often finds that female students perform relatively worse under higher stakes (Azmat et al., 2016; Cai et al., 2019; Iriberri and Rey-Biel, 2019; De Paola and Gioia, 2016; Montolio and Taberner, 2021; Arenas and Calsamiglia, 2025). Our results help identify more precisely when gender differences in performance under pressure begin to emerge and distinguish gender differences in skill accumulation from gender differences in responses to high-stakes testing conditions.

These findings matter for how standardized test scores are interpreted in both economics and education policy. If test pressure affects some groups more than others, observed achievement gaps may partly reflect differences in response to incentives rather than differences in underlying competence. Our results suggest that this concern may depend importantly on the stage of schooling at which high-stakes assessments are introduced. An important direction for future research is to identify the mechanisms through which the gender gap under pressure emerges later in the educational career. Understanding when these differences arise and why they arise is crucial for designing effective policy responses.

References

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

Angrist, J. D., E. Battistin, and D. Vuri (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., E. La Ferrara, and P. Pinotti (2022a, May). Implicit stereotypes in teachers' track recommendations. *AEA Papers and Proceedings 112*, 409–14.

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

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.

Delaney, J. M. and P. J. Devereux (2021, 08). The economics of gender and educational achievement: Stylized facts and causal evidence.

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.

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.

Pekkarinen, T. (2015). Gender differences in behaviour under competitive pressure: Evidence on omission patterns in university entrance examinations. *Journal of Economic Behavior & Organization 115*, 94–110. Behavioral Economics of Education.

Saygin, P. O. (2019, August). Gender bias in standardized tests: evidence from a centralized college admissions system. *Empirical Economics 59*(2), 1037–1065.

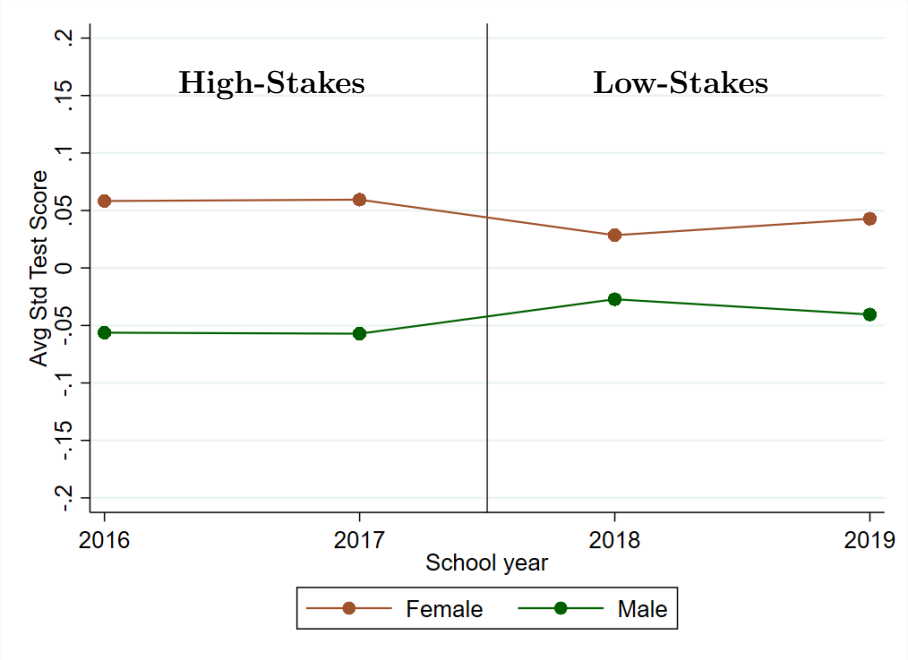
Schlosser, A., Z. Neeman, and Y. Attali (2019, 05). Differential Performance in High Versus Low Stakes Tests: Evidence from the Gre Test*. *The Economic Journal 129*(623), 2916–2948.

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.

7 Tables and Figures

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.

Table 1: DiD Estimation – 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 year level to have mean equal to 0 and standard deviation equal to 1. Control variables includes students' origin and regularity in their studies (i.e., if they have failed a year or not), parents' education and profession. Standard errors, clustered at school level, are in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 2: Robustness Checks

	(1)	(2)	(3)	(4)	(5)
Female × High-Stakes	0.032***			0.065***	0.065***
	(0.012)			(0.003)	(0.002)
Female × Pre-Reform		0.025***	-0.021***		
		(0.005)	(0.002)		
Female	0.048***	-0.038***	0.029***	-0.000	
	(0.008)	(0.004)	(0.002)	(0.002)	
High-Stakes	-0.003			0.051***	0.051***
	(0.013)			(0.002)	(0.001)
Pre-Reform		-0.047***	0.035***		
		(0.005)	(0.004)		
Controls	Yes	Yes	Yes	Yes	No
School FE	No	Yes	Yes	No	No
Student FE	No	No	No	No	Yes
Observations	114,245	1,727,633	3,638,297	2,032,380	2,032,380
R^2	0.149	0.432	0.174	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 year level to have mean equal to 0 and standard deviation equal to 1. Control variables includes students' origin and regularity in their studies (i.e., if they have failed a year or not), parents' education and profession. Column (1) reports estimates using the representative subsample of externally monitored classrooms to address concerns regarding manipulation or cheating. Columns (2) and (3) report placebo results for grades 10 and 5, which help isolate the effect of changes in stakes from other reform-period shifts. Columns (4) and (5) exploit within-student variation across grades 5, 8, and 10 using a longitudinal panel, with Column (5) including student fixed effects to absorb time-invariant student heterogeneity. Standard errors, clustered at the school level in Columns (1)–(3) and robust to heterogeneity in Columns (4)–(5), are reported in parentheses.* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3: DiD Estimation – Analysis by Subject

Panel A: Italiano

	(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 the above tables, the dependent variable Y_{it} represents 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 includes students' origin and regularity in their studies (i.e., if they have failed a year or not), parents' education and profession. Standard errors, clustered at school level, are in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Appendix

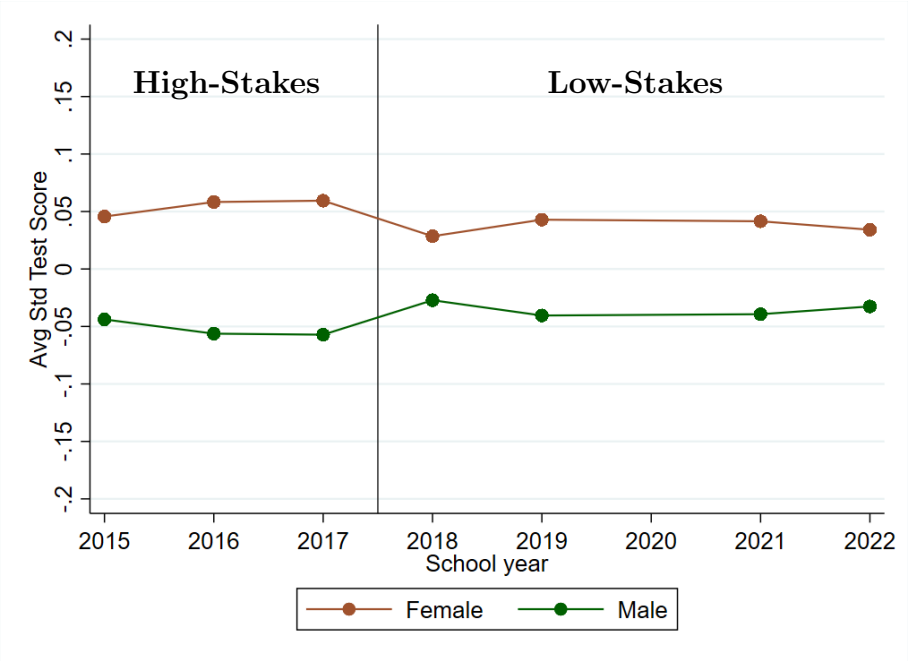
A1 Full Sample

Table A1: DiD Estimation – 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 year level to have mean equal to 0 and standard deviation equal to 1. Control variables includes students' origin and regularity in their studies (i.e., if they have failed a year or not), parents' education and profession. Standard errors, clustered at school level, are in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

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 didn't take place). 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.

Alma Mater Studiorum - Università di Bologna

DEPARTMENT OF ECONOMICS

Strada Maggiore 45
40125 Bologna - Italy
Tel. +39 051 2092604
Fax +39 051 2092664
http://www.dse.unibo.it