When science feels like math: Quantitative questions and the science gender gap

Divyanshu Jain Tarun Jain*

September 5, 2025

Abstract

This paper investigates the role of math-intensive evaluation in driving science gender gaps. Using question-level data from a large standardized exam conducted annually from grades 3 to 10 in India, we find that (i) boys outperform girls in mathematics in each grade, (ii) girls outperform boys in science in grades 3 to 6, and underperform in grades 7 to 10, and (iii) the gender gap in science is entirely explained by underperformance on math-intensive questions. Girls perform comparably to boys on non-quantitative science questions, but underperform on quantitative science questions across all grades. Further, conditional on previous year performance, girls face a larger penalty from increasing math intensity in subsequent science exams. These findings highlight the importance of how science is evaluated in explaining gender differences in science performance in school.

Keywords: Science achievement; Gender gap; Math-in-science; School education

JEL Classification: I21; I24; J16

*D. Jain: Economics Area, Indian Institute of Management Ahmedabad (email: phd21divyanshuj@iima.ac.in); T. Jain: Economics Area, Indian Institute of Management Ahmedabad (email: tj9d@virginia.edu). We appreciate the cooperation of Educational Initiatives, especially Pranav Kothari, Aarthi Muralidharan, and Pramod Krishna, in sharing the ASSET Exam data and offering insights into its operations. Comments by Pritha Dev, Ambrish Dongre, Abhijeet Singh and participants at several conferences and workshops helped to improve the paper substantially. We thank Shreya Sam for excellent research assistance. Views expressed are strictly those of the author, and may not be attributed to the Reserve Bank of India.

1 Introduction

Science education is important in the modern economy, with several studies pointing to the importance of scientific knowledge in better economic performance (Balart et al., 2018; Murphy et al., 1991), higher earnings (Altonji et al., 2012; Jain et al., 2022; Kuuppelomäki et al., 2025; Roychowdhury, 2021), and improved job quality (Black et al., 2021; Sahoo and Klasen, 2018). Despite these benefits, men are more likely to study science and allied subjects such as engineering and finance, and enter related professions compared to women (Bell et al., 2019).[1] A large fraction of elite scientists are male - for example, 627 men have received a Nobel Prize in the sciences compared to 26 women. This mismatch has both efficiency and equity implications. From the efficiency perspective, the economy does not benefit from women's scientific perspectives and contributions (Kahn and Ginther, 2017). Simultaneously, from an equity perspective, women miss the professional opportunities and rewards from studying science and related fields such as technology, engineering, mathematics, and economics (Delaney and Devereux, 2019; World Bank, 2012).

Where and how does the gender gap in science education emerge? Studies investigating the STEM ("Science, Technology, Engineering and Mathematics") achievement gender gap have pointed to three stylized facts. First, gender differences on tests of mathematics knowledge and skills are large and persistent (Bharadwaj et al., 2016; Dercon and Singh, 2013; Ellison and Swanson, 2010; Fryer and Levitt, 2010). Second, women's under-representation in science is primarily limited to the math-intensive science fields, including geosciences, engineering, economics, math/computer science, and physical science (Kahn and Ginther, 2017). Third, how students' knowledge is evaluated or questions are framed can significantly influence measures of their performance (Banerjee et al., 2025; Davison and Dustova, 2017; Royer et al., 1999; Reardon et al., 2018). These facts collectively suggest that if girls underperform boys in

[1]See the gender gaps in the Joint Entrance Examination in India (Sharma and Sengupta, 2024), Gaokao in China (Cai et al., 2019), and the Turkish University Entrance Exam (Akyol et al., 2024), which are used to admit students in top engineering and science courses. Also see the gender gaps in performance on the international PISA examination (Anaya and Zamarro, 2024; Griselda, 2024; World Bank, 2018).

mathematics, then quantitative framing of science questions could be a reason for the subsequent science gender gap.[2]

In this paper, we investigate the role of mathematics-intensive framing to explain why girls underperform (compared to boys) in tests of scientific knowledge and proficiency. Specifically, we document a significant gender gap in math achievement through grades 3 to 10. Next, we show the trend in gender differences in science scores from grades 3 to 10, and the flip from girls' overachievement in science in early grades to underachievement in the later grades. Most importantly, we pinpoint the role of math-intensive science questions as the key reason why girls underperform boys in science examinations.

Researchers investigating these questions face several data challenges. First, understanding when the gender gap in science and mathematics achievement begins requires data from similar evaluations conducted over several grades (which means that data from one-shot exams such as the SAT, JEE, and Gaokao is not useful). Second, the data should contain results from science, mathematics, and language tests so that spillovers from one subject can be observed in another. Third, the researcher must be able to observe precise questions used in the tests to classify their mathematics intensity.

We overcome these challenges using granular, administrative data from the Assessment of Scholastic Skills through Educational Testing (ASSET) exam conducted among grade 3 to 10 students from 591 private schools across 147 cities in India. We create a dataset from the universe of ASSET exam data from 2022 to 2024 for science, mathematics, and English. We classify the quantitative intensity of each question on the science exam using AI models, keyword-based algorithms, and manual coding, as well as the combination of all three methodologies, and match that to the performance of each student on each question. Using this data and a fixed effects specification that controls for time-invariant student and school characteristics, time fixed effects and several other question-level factors, we estimate the gender gap in mathematics, gen-

[2]Benbow and Minor (1986) observed that "sex differences in mathematical reasoning ability may explain some of the sex difference in science participation and achievement."

der differences in science scores, and the gender gap in performance on math-intensive versus non-math-intensive science questions through grades 3 to 10.

Our main findings are as follows. First, boys outperform girls in mathematics in each year from grades 3 to 10, indicating that the gender gap in mathematics starts as early as primary school and sustains through secondary school. Second, girls outperform boys in science in grades 3 to 6 (when around one-twelfth of questions are math-intensive), but boys outperform girls in science in grades 7 to 10, with significantly better scores each year, when approximately one-fifth of the science questions are math-intensive. This finding echoes results from other studies that report that the science gap typically emerges in middle school (Kahn and Ginther, 2017).

Third and most significantly, we find that girls consistently underperform boys on math-intensive science questions. On average, girls have a 2.8% lower probability of correctly answering a math-intensive science question compared to boys. With the base probability of a correct answer at 43.8%, the gender gap on math-intensive science questions translates to a 6.4% disadvantage for girls. Unpacking science into its constituent branches (i.e., physics, chemistry, and biology) reveals physics (which is the most math-intensive) to be the main driver of the gender gap, while girls' relative advantage in biology and chemistry is also undermined once mathematical intensity increases.

Overall, on science exams across all grades, girls underperform boys in science by 0.17 percentage points. On non-math-intensive questions, girls actually perform 0.3 percentage points *better* than boys. However, this advantage is offset by the disadvantage for girls on math-intensive science questions (which are 14% of science questions across all grades). Thus, math-intensive framing of science questions accounts for the entire observed gender gap in science performance. In grades 7 to 10, when the science gender gap is most pronounced, boys outperform girls by 1.08 percentage points. In these grades, performance on math-intensive questions (which are nearly 17% of a typical test) explains approximately 45% of the science gender gap. We also find that, conditional on previous year performance, girls face a steeper decline (compared to boys) in response to increasing math intensity of science exams. Thus, observed gender differences do not simply reflect pre-existing ability differences; rather, they arise from differential progress within science.

Many studies in this domain combine science and mathematics under the STEM acronym, which ignores the possibly different learning dynamics between these subjects. We unpack this dynamic by considering the potential spillovers of mathematics performance on science learning. Thus, we bridge the literature on gender gaps in mathematics and science achievement.[3]

By highlighting the role of math-intensive evaluation in science, our paper investigates a novel reason why gender gaps in science arise. Prior research has explored stereotype internalization (Nosek et al., 2009), differential response to competitive tests (Ors et al., 2013), teacher and family stereotypes (Kahn and Ginther, 2017), teacher gender (Carrell et al., 2010), peer gender (Brenøe and Zölitz, 2020), and exposure to hands-on lab activities (Burkam et al., 1997) as possible reasons why boys outperform girls in science. By offering a new mechanism for these gender gaps, we draw the attention of researchers and policymakers to how deficiencies in one subject can spill over into another. Thus, remedial early mathematics education among girls might have larger positive spillovers in facilitating science achievement, and possibly STEM-based careers.

We also contribute to the literature on science and STEM major choice in college, as well as the subsequent transition to scientific careers. This literature examines several factors that contribute to the gender gap in STEM-based majors and jobs. These include preferences for major-specific attributes and confidence in major-specific abilities (Dasgupta and Sharma, 2022; Shi, 2018; Zafar, 2013), competitiveness (Buser et al., 2014), value of home and leisure (Gemici and Wiswall, 2014), female role models (Porter and Serra, 2020), and culture (Friedman-Sokuler and Justman, 2020). By high-

[3]The STEM gender gap is discussed by Black et al. (2021); Card and Payne (2021); Cimpian et al. (2020); Delaney and Devereux (2019); Goulas et al. (2024); Nosek et al. (2009), and Park et al. (2018). See Bonnot and Croizet (2007); Bharadwaj et al. (2016); Buser and Yuan (2019); Contini et al. (2017); Delaney and Devereux (2021); Ellison and Swanson (2010); Fryer and Levitt (2010); Niederle and Vesterlund (2010); Nollenberger et al. (2016) on the gender gap in mathematics.

lighting specific deficiencies in math and science performance among girls compared to boys in secondary school, we offer a reason why girls might feel less STEM-ready compared to boys, and hence choose science and engineering majors at lower rates than boys.

We also contribute to the economics of education literature by analyzing a new standardized examination dataset. The ASSET exam has been conducted in India since 2001, with over 15 million individual assessments conducted so far. It is the only multi-subject (mathematics, science, English, Hindi, social science, and computational thinking) and multi-grade (grades 3 to 10) examination anywhere in the world conducted at such scale. Since students and schools are tracked over time, the ASSET Exam database (2022-24) permits precise analysis of research questions that are difficult to do with other data sources.

2 Data

2.1 Data source: ASSET exam

The Assessment of Scholastic Skills through Educational Testing (ASSET) exam is a computer (or OMR) based test conducted annually since 2001 by Educational Initiatives (Ei), an Indian education firm. The tests cover English, mathematics, science, and computational thinking across grades 3 to 10, social science for grades 5 to 10, and Hindi for grades 4 to 8. The primary purpose of the exam is for schools to benchmark their students' learning from early grades (large-scale standardized testing otherwise starts only in grade 10) and to diagnose school-level deficiencies in learning. Thus, Ei reports results by student, by teacher, and by school to permit precise benchmarking and diagnosis.

The exam is developed using unfamiliar and thought-provoking questions to assess students' understanding and application of concepts. The questions are designed by subject matter experts based on the National Curriculum Framework's (NCF) guidelines. Each question goes through a peer review process that checks for clarity, age appropriateness, technical correctness, and diagnostic value by five independent experts before being included in the ASSET exam. Since the start of the ASSET exam, Ei has conducted over 15 million individual assessments.

2.2 Sample description

In this study, we use the universe of ASSET exam data for mathematics, science, and English for the period 2022-24, covering 591 schools across 147 cities in 26 states (see Figure A1 for the geographical coverage of schools taking the ASSET exam). Our sample includes de-identified question-level details of tests for grades 3 to 10. Table 1 shows the summary statistics of our sample for the period 2022-24. The number of questions on each exam differs based on the subject and grade. With each question on an exam as the unit of observation, our ASSET Exam database (2022-24) sample comprises 20,478,379 observations at the question level for mathematics, 20,450,072 observations for science, and 28,958,825 observations for English.

Boys are approximately 54% and girls are 46% of the total sample. This distribution remains consistent across all the subjects and grades. However, this dataset is not a balanced panel, as not all students are observed from grades 3 to 10. Moreover, some students and schools drop out or enter late into the sample and are not observed in all three years from 2022-24.[4] We observe a drop in the sample size in grades 9 and 10, suggesting that fewer schools and students took the ASSET exam in these grades. Table 7 in the Appendix shows the number of students and schools observed in each year. We check for the effects of possible attrition and its potential influence on our main results in section 5.2.

The mean probability of a correct answer for girls, boys, and the difference between them is shown in Columns (7), (8) and (9) of Table 1, with the difference also shown in Figure 1. These simple differences in means imply that boys outperform girls in all grades in mathematics, and girls outperform boys in all grades in English. In science, girls perform better than boys in grades 3 to 6, but this pattern is reversed in grades 7

[4]For example, in science (which is our main subject of analysis), there are 353,437 students across grades 3 to 10 during 2022-24, but only 514,196 student-exam observations in total.

to 10.

The dataset includes additional question-level variables. These include the difficulty level of each question (easy = 1, medium = 2, and hard = 3), as decided by subject matter experts at Ei. Also included is the question number that depicts where in the sequence the question appears in each exam. By including this in regressions, we are able to control for any differential performance due to the length of the exam or the order of the questions. Additionally, motivated by evidence that language proficiency is linked to achievement in both arithmetic and science domains (Reed et al., 2017; Wei et al., 2012), we include the language complexity of each question. The language complexity variable is created using OpenAI's GPT-4o model. We asked the model to rate the language complexity of each question, according to the respective grade, on a scale of 1-10 (increasing in complexity).

Finally, the dataset includes several categorical variables that represent the skills tested by each question. For example, each science question is categorized into one of thirteen skills, such as definition or description, hypothesis formulation and prediction, knowledge of scientific instruments and tools, recollection or recognition of scientific facts and concepts, application of information to identify trends and properties, etc. Similarly, each mathematics question is categorized into one of twenty-nine skills, such as arithmetic operations, shapes and geometry, fractions, area and perimeter, algebra, data interpretation, etc. Descriptive statistics of these additional question-level variables are shown in Table A1 in the Appendix.

2.3 Question classification

The central challenge in studying the gender gap in math-intensive science questions is classifying each question into "math-intensive" or not. In the absence of any established method of doing this, the essential solution is to develop and implement a classification methodology. We augment the ASSET exam (2022-24) database with variables that classify each science question according to its mathematics intensity. To generate these variables, we develop and implement four methodologies.

1. **Generative Artificial Intelligence (GenAI):** In this approach, we ask OpenAI's GPT-4o model (release date: 13 May 2024), which is a neural network-based large language prediction model, to do a binary classification of each of the 1950 science questions into "math-intensive" or not. We first ask the GenAI model whether it can understand if answering a question requires application of mathematics skills or knowledge. Then we understand the model's basis for classification. The primary criteria include inferring the concepts involved in the question and the quantitative skills required to answer the question. If the model perceives that mathematical and/or quantitative reasoning is required, then it classifies the question as "math-intensive", otherwise not. Figure A2 in the Appendix shows examples of the GPT model's prompts and responses during the classification exercise. Column 3 of Table 2 shows the share of science questions classified as math-intensive by grade using the GenAI methodology.

2. **Keyword-based algorithm (KW):** This approach classifies questions as "math-intensive" if certain mathematical or quantitative terms (such as "graphs", "units", and "calculate") appear in the question's text. The list of these terms and their respective frequencies for each grade is given in Appendix Table A2. This approach is independent of inferring the text of the question and avoids subjective judgment.[5] Column 4 of Table 2 shows the share of science questions classified as math-intensive for each grade by the KW methodology.

3. **Manual coding:** This approach involves two individuals (Ind 1 and Ind 2, both with at least Master's degrees) carefully evaluating the math intensity of each question by reading the text of the questions. Classification based on this method relies on human judgment and interpretation and is independent of computer algorithms and models. The advantage of this approach is that individuals can determine the actual applicability of math skills instead of relying only on the presence of mathematical terms. Thus, the manual coding approach is more

[5] This list of keyword families was created using the help of artificial intelligence, where the GPT model checked standard science curriculum and previous public exams to come up with a list of word families that indicate quantitative or mathematical concepts.

conservative compared to the keyword method. Columns 5 and 6 of Table 2 show the share of science questions classified as math-intensive for each grade by Ind 1 and Ind 2 separately.

4. **Average (Avg):** We create a non-binary measure of math intensity by averaging the binary responses (1 if math-intensive, 0 otherwise) from the previous three methods, hence incorporating information from all the classification methods into one value for each question (Avg $\in$ [0, 0.25, 0.5, 0.75, 1] for each question). The advantage of a non-binary measure of math intensity is finer classification, which avoids extreme variance. Column 7 of Table 2 shows the mean value of math intensity by grade using the Avg measure.[6]

To validate our classification methodology, we report Cronbach's alpha of the four binary methods of classification (GenAI, KW, Ind 1, and Ind 2). Cronbach's alpha is 0.81, indicating a high level of reliability and consistency. Further, the interrater agreement coefficients in Table A4 in the Appendix corroborate the moderate to high reliability of the various classification methodologies (Klein, 2018).

3 Empirical analysis

3.1 Gender gap in mathematics and science performance

We analyze the gender gap in mathematics and science performance with the following empirical specification estimated on the mathematics questions and science questions subsample separately.

$$\begin{aligned} y_{iqegst} \;=\; & \beta_0 + \beta_1(\text{Girl}_i \times \text{Grade}_g) + \beta_2\text{Girl}_i + \beta_3\text{Grade}_g \\ & + \beta_4\text{Difficulty}_{qe} + \beta_5\text{QNo}_{qe} + \beta_6\text{LangComplexity}_{qe} \\ & + \text{SkillTestedFE}_{qe} + \text{SchoolFE}_s + \text{YearFE}_t + \epsilon_{iqegst} \qquad (1) \end{aligned}$$

[6]See Table A3 for examples of science questions classified as non-math-intensive and math-intensive by all classification methods unanimously (i.e., Avg = 0 and Avg = 1, respectively).

In equation (1), y_{iqegst} is the outcome variable for student i on question q in exam e of grade g from school s in year t. y_{iqegst} is a binary variable which is 1 if the question is answered correctly, and 0 otherwise. Girl_i is 1 when the student i is a girl and 0 if boy. Grade_g indicates the grade of the student giving the exam e. Thus, β_1 is the key coefficient of interest representing the gender gap in performance on mathematics and science, with $\beta_1 > 0$ indicating better performance by girls on the specific question compared to boys with grade 3 as the base. We include the difficulty level of each question (Difficulty_{qe}), the order of a question in an exam (QNo_{qe}), the language complexity of each question ($\text{LangComplexity}_{qe}$), and the skills tested ($\text{SkillTestedFE}_{qe}$) as control variables. Finally, we control for all common year characteristics using year fixed effects (YearFE_t), and the time-invariant school characteristics using school fixed effects (SchoolFE_s). These school, grade, and year fixed effects control for contextual and cohort-level variation in instruction, classroom dynamics, or differences in test-taking environments. The unobservable characteristics are represented by ϵ_{iqegst}.

One shortcoming of the dataset and, therefore, the analysis in this section is that we do not observe student characteristics other than gender. Hence, we cannot control for other factors such as access to tutoring, teacher inputs, and household characteristics, as well as behavioral factors such as ability, confidence, and motivation that might influence student performance on the ASSET exam. Another key assumption is that the performance on mathematics or science on the ASSET exam does not influence grade progression or the propensity to take the exam in a subsequent year. This assumption is plausible since the exam is used as a benchmarking tool by schools, not to decide students' grade progression. Nonetheless, we check the influence of selective attrition in section 5.2.

Using equation (1), Figure 2 shows the estimated gender gap for mathematics by grade. The consistently negative coefficient, ranging between 1% and 3.5%, implies that girls perform worse than boys in mathematics for each grade from 3 to 10. This translates to a gender gap of 0.02 to 0.07 σ in the probability of a correct answer. That the gender gap in mathematics achievement pre-dates grade 3 is consistent with Martinot et al. (2025)'s findings from French data that the mathematics gender gap emerges rapidly in first grade, thus well established by third grade. Further, the consistently growing gender gap from lower to higher grades is consistent with Bharadwaj et al. (2016)'s finding from Chilean data that the gender gap in mathematics doubles between grades 4 and 8.

Our estimate (0.07 σ for grades 8 and 10) can be compared to Hyde et al. (2008)'s estimate of a 0.05 σ gender gap in grade 10 math performance, and Pope and Sydnor (2010)'s the 0.06 σ gap in grade 8 performance on the National Assessment of Educational Progress (NAEP) in the USA. Our finding is smaller than Fryer and Levitt (2010)'s estimate of a 0.2 σ gap by the fifth grade, again in the USA. Stoet and Geary (2013) also reports larger math achievement gender gaps than our estimates for 15-year-olds: 0.25 σ in South Korea and 0.21 σ in Greece. However, Iceland (0.17 σ) and Thailand (0.05 σ) have gender gaps favoring girls.

Figure 3 shows that in science, girls outperform boys in grades 3 to 6, whereas boys outperform girls in grades 7 to 10, with significantly better scores each year (0.5% difference in grade 7, 1.5% difference in grades 8 and 9, and 2% difference in grade 10). Our point estimate (0.03 σ in grade 8) is smaller than Nosek et al. (2009)'s estimate of a 1.1 σ gap in grade 8 in the 2003 Trends in International Mathematics and Science Study (TIMSS) in the USA. Over all grades, our results are consistent with Kahn and Ginther (2017)'s review that the gender gap in science emerges around middle school, i.e., grades 6 to 8, and not earlier.

3.2 Gender gap in performance on math-intensive science questions

To estimate the role of math-intensive science questions, we analyze the performance of student i on question q in exam e of grade g from school s at time t with the following specification.

$$
\begin{aligned}
y_{iqegst} \quad = \quad & \beta_0 + \beta_1(\text{Girl}_i \times \text{MathIntensive}_{qe}) + \beta_2\text{MathIntensive}_{qe} \\
& +\beta_3\text{Difficulty}_{qe} + \beta_4\text{QNo}_{qe} + \beta_5\text{LangComplexity}_{qe} \\
& +\text{SkillTestedFE}_{qe} + \text{StudentFE}_i + \text{GradeFE}_g \\
& +\text{SchoolFE}_s + \text{YearFE}_t + \epsilon_{iqegst} \qquad (2)
\end{aligned}
$$

The common variables in equation (2) are the same as in (1). We introduce $\text{MathIntensive}_{qe}$, which is a binary variable equal to 1 if question q is classified as math-intensive and 0 if not.[7] We estimate equation (2) for each classification methods separately. β_1 is the key coefficient of interest representing the gender differences in the performance on math-intensive science questions. In addition to the control variables and fixed effects in (1), we control for any time-invariant student characteristics using student fixed effects (StudentFE_s) and any grade-specific factor using grade fixed effects (GradeFE_g). Unobservable characteristics (ϵ_{iqegst}) are clustered at the student level.

In equation (2), we assume that the assignment of math-intensive questions within a test is independent of unobserved factors that influence performance but vary by gender. In this setting, tests are externally designed, standardized, and uniformly administered, and students do not select specific questions. Further, the inclusion of student fixed effects allows us to control for time-invariant student characteristics (such as ability, confidence, or motivation, as well as teacher inputs, access to tutors, and household characteristics). Thus, we interpret β_1 precisely as the difference in how girls (compared to boys) perform on math-intensive versus non-math-intensive questions on the ASSET examination.

Table 3 presents estimates from equation (2) using each classification method separately. Across all methods, the interaction term coefficient indicates that girls perform worse than boys on math-intensive science questions when pooling grades 3 to 10. Considering the average classification measure for interpretation, girls have a 2.8%

[7] In case of the Avg measure of math intensity, $\text{MathIntensive}_{qe}$ is a non-binary variable that indicates the level of math intensity on that question.

lower probability of correctly answering a math-intensive science question than boys — a difference equivalent to 0.06 σ of the overall mean probability of a correct answer.

Figure 4, based on equation (2), illustrates predicted probabilities of a correct answer for girls and boys by question type.[8] When pooling all grades, girls slightly outperform boys on non-math-intensive science questions by 0.3 percentage points, but fall behind by 3 percentage points on math-intensive questions. Given that math-intensive items comprise approximately 14% of the typical science exam, these gender differences result in an overall gender gap of 0.17 percentage points in favor of boys.

Table 4 reports estimates of equation (2) separately for each grade. The disadvantage for girls on math-intensive science questions is evident and persistent from grades 3 through 10, with gender gaps ranging from 1.2 to 5.8 percentage points (or 0.02 σ to 0.12 σ).

Figure 5 shows combined results for grades 7 to 10 to focus on the grades where the overall science gender gap emerges. In these grades, girls underperform boys by 2.9 percentage points on math-intensive science questions and also lose their advantage on non-math-intensive questions, suggesting that underperformance in math-intensive content may spill over to overall science performance. Since 17.3% of science questions in these grades are math-intensive, this component alone accounts for roughly 45% of the total science gender gap in grades 7 to 10, which is 1.08 percentage points (0.02 σ) in aggregate.

4 Additional analysis

4.1 Unpacking science: patterns in physics, chemistry, and biology

While the main analysis treated science as a single subject, it might be useful to unpack it into its constituent branches, i.e., physics, chemistry, and biology, because these differ in content and skill demands. In this subsection, we assess whether the over-

[8]This is based on the average measure of classification, but without the student fixed effects to enable estimation of all coefficients for visualization.

all gender patterns in science are uniform or whether they vary systematically across branches.

We highlight three main findings. First, the baseline gender gaps differ across branches (see Figure 7). Girls consistently outperform boys in biology and chemistry in the lower and middle grades, while the advantage reverses in the later grades, particularly in chemistry. In contrast, physics shows a persistent disadvantage for girls throughout, which widens with grade progression.

Second, the distribution of questions and the share of math-intensive items (see Table 5) provide some explanation for these divergent patterns. Physics not only accounts for the largest share of science questions, but also has the highest proportion of math-intensive ones (around 19%). Biology has the fewest math-intensive questions (around 9%), and chemistry lies in between (around 14%). The branches where girls initially do better, i.e., biology and chemistry, are precisely those with lower exposure to math-intensive content, while the branch where they underperform the most (physics), places the greatest mathematical demands.

Finally, we replicate the main analysis of gender differences on math-intensive versus non-math-intensive questions separately for each branch (Figure 6).[9] Girls perform better on non-math-intensive questions (except in physics) but do worse on math-intensive ones. Taken together, these results underscore that gender gaps in science are not uniform. They depend both on the branch and on the degree of mathematical intensity within it. Physics seems to be the main driver of the gender gap in science, while girls' relative advantage in biology and chemistry is undermined once mathematical intensity increases.

4.2 Alternate specification: exam level

We examine the role of math-intensive science questions in the science gender gap using an alternate specification where the unit of analysis is an exam instead of a question. We analyze the score percentage (y_{iegst}) of student i on exam e of grade g from

[9] This is based on equation (2) without student fixed effects for estimation of all coefficients for visualization.

school s in year t.

$$
\begin{aligned}
y_{iegst} &= \beta_0 + \beta_1(\text{Girl} \times \text{MathIntensiveShare})_{ie} + \beta_2\text{MathIntensiveShare}_e \\
&\quad + \text{StudentFE}_i + \text{SchoolFE}_s + \text{YearFE}_t + \epsilon_{iegst} \qquad (3)
\end{aligned}
$$

In equation (3), $\text{MathIntensiveShare}_e$ is the share of math-intensive questions in exam e. As earlier, this equation is estimated for each classification method separately, as well as using a binary indicator (high math intensity = 1 if the share of math-intensive questions (Avg measure) is more than the median; 0 otherwise) to study the effect of math intensity and performance on science exams. All other variables are as defined earlier.

Table 6 presents estimates from equation (3) using the six measures of math intensity for all grades put together. Across all measures, the interaction between the girl indicator and exam-level math intensity is negative and statistically significant, indicating that as the share of math-intensive questions increases, the relative performance of girls declines more sharply than that of boys.

To illustrate the magnitude and nature of this relationship, Figure 8 plots the predicted science exam scores of boys and girls at varying levels of math intensity (using the Avg measure). At low levels of math intensity (below 10%), girls outperform boys. As the math intensity of an exam increases, this advantage erodes and eventually reverses. At 20% math-intensive questions and above, boys score significantly higher than girls, with the gender gap widening further as math intensity increases.

Taken together, these findings provide strong evidence that test content, particularly math intensity, plays a key role in shaping gender gaps in science achievement and may help explain why aggregate gender gaps emerge more sharply in later grades, when exams become increasingly math-intensive.

To assess whether the observed gender differences simply reflect pre-existing achievement gaps or instead arise from differential progress within science, we estimate a value-added specification that conditions on students' prior-year performance.[10] Fig-

[10]We estimate the following regression specification: $y_{iegst} = \beta_0 + \beta_1(y_{iegst-1}) + \beta_2(\text{Girl} \times$

ure 9 plots the estimated gender gap in predicted scores against the share of math-intensive questions in an exam, holding the previous year's performance constant (at the mean). We observe that when exams contain very few math-intensive questions, girls slightly outperform boys, but as the share of such questions increases, the gender gap turns negative and grows steadily larger. At around 40% math-intensive content, girls score roughly 0.2 σ lower than boys with the same prior performance. This finding demonstrates that the gender gap is not merely a reflection of initial ability differences; rather, it emerges specifically in response to the increasing mathematical demands of the science exam.

5 Robustness checks

5.1 Alternate choice of fixed effects, and clustering of standard errors

We estimate a modified version of equation (2) by using (School $\times$ Year) interactive fixed effects in place of separate school and year fixed effects, and by clustering the standard errors at the School $\times$ Year level rather than the student level. The interaction term coefficient in Appendix Table A6 shows that the gender gap in math-intensive science questions remains consistent with the results obtained in Table 3.

5.2 Attrition

The ASSET exam database (2022–2024) includes schools and students who do not appear in all three years, either due to attrition or late entry. Table 7 shows the distribution of school and student presence across years. To ensure that such sample variation does not bias our estimates, we undertake a series of robustness checks.

First, Appendix Table A7 shows that average school-level performance in science does not correlate with school participation in the full three-year window. Second, Appendix Table A8 demonstrates that the relationship between student performance

$\text{MathIntensiveShare})_{ie} + \beta_3 \text{Girl}_i + \beta_4 \text{MathIntensiveShare}_e + \text{GradeFE}_g + \text{SchoolFE}_s + \text{YearFE}_t + \epsilon_{iegst}$.

and continued presence in the data does not differ systematically by gender. This evidence suggests that entry and exit from the dataset are not selectively associated with student ability in ways that would bias gender comparisons. To further address potential concerns of endogenous attrition, we implement Lee bounds using our exam-level specification with the binary math intensity measure (Lee, 2009).[11] Both the lower and upper bounds remain negative and statistically significant, confirming that our main results are robust even in the most conservative selection scenario.

We also check robustness to attrition by estimating equation (2) on the subsample of schools that are observed in all three years ("always-in" schools; see Appendix Table A9), as well as students who are observed in all three years ("always-in" students; see Appendix Table A10). The results are qualitatively the same as our main findings. Collectively, these checks confirm our results are not driven by selective test participation or data availability.

6 Conclusion

This paper studies the gender gaps in mathematics and science achievement across grades 3 to 10 using granular, question-level data from a large-scale standardized exam in India. In mathematics, we find significant gender gaps favoring boys across all grades from 3 to 10. The direction and magnitude of these differences from our estimates are consistent with those reported in other studies (Hyde et al., 2008; Pope and Sydnor, 2010). In science, we find that girls outperform boys in early grades, but start underperforming by grade 7, and do so till grade 10. To explain this, we investigate mathematics-intensive evaluation of science proficiency as a potential explanation for girls' science performance. Our findings suggest that the gender gap in science is driven primarily by girls' lower performance on math-intensive science questions. While the disadvantage for girls on math-intensive questions is consistent from grade 3 to 10, the share of such questions in science exams increases from an

[11]This specification is best suited to the assumptions underlying Lee bounds estimation.

average of 8% in grades 3 to 6 to an average of 20% in grades 8 to 10. We also show that the math intensity of science exams influences girls more than boys, conditional on previous year performance.

Our results should be interpreted with a few caveats. First, we cannot control for time-varying student attitudes, motivation, or household characteristics that might influence performance. Second, the gender gap in quantitative-science performance might have implications beyond test scores, and could influence science study in higher secondary grades and college, as well as scientific occupations. Absent tracking data, we cannot investigate these interesting outcome variables. Finally, just because we find evidence of gender gaps in the performance on quantitative science in this setting does not imply that the results will be the same in other settings. The dynamics might be very different in magnitude or direction in other institutional or social settings. For example, in social settings with different expectations from girls, initial mathematics training might be stronger which might alleviate the gender gap in science performance.

Nonetheless, our study contributes new evidence on both the timing and the drivers of gender gaps in science achievement. By highlighting the role of math-intensive evaluation, we show that gender differences in science might be primarily driven by quantitative skill demands rather than overall ability in science. These findings suggest that strengthening foundational math skills and confidence for girls, or reconsidering how math-intensive material is introduced in science assessments, could have significant positive spillovers in reducing gender gaps in science education. Our findings are also useful in explaining the gender-based sorting in higher education and future careers, with men more likely to pursue quantitative science, engineering and technology tracks, and women disproportionately choosing non-quantitative science (such as life sciences), social science and humanities. Finally, our results could inform the design of interventions to address gender gaps in science. One possibility is to teach and evaluate science in non-quantitative ways, including laboratory-based and experiential teaching, but future research could consider other options as well.

References

Akyol, P., Krishna, K., and Lychagin, S. (2024). Targeting the gender placement gap: Marks versus money. NBER Working Paper No. 33074.

Altonji, J., Blom, E., and Meghir, C. (2012). Heterogeneity in human capital investments: High school curriculum, college major, and careers. *Annual Review of Economics*, 4(1):185–223.

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

Balart, P., Oosterveen, M., and Webbink, D. (2018). Test scores, noncognitive skills and economic growth. *Economics of Education Review*, 63:134–153.

Banerjee, A., Bhattacharjee, S., Chattopadhyay, R., Duflo, E., Ganimian, A., Rajah, K., and Spelke, E. (2025). Children's arithmetic skills do not transfer between applied and academic mathematics. *Nature*, pages 1–9.

Bell, A., Chetty, R., Jaravel, X., Petkova, N., and Van Reenen, J. (2019). Who becomes an inventor in America? The importance of exposure to innovation. *Quarterly Journal of Economics*, 134(2):647–713.

Benbow, C. and Minor, L. (1986). Mathematically talented males and females and achievement in the high school sciences. *American Educational Research Journal*, 23(3):425–436.

Bharadwaj, P., De Giorgi, G., Hansen, D., and Neilson, C. A. (2016). The gender gap in mathematics: Evidence from chile. *Economic Development and Cultural Change*, 65(1):141–166.

Black, S., Muller, C., Spitz-Oener, A., He, Z., Hung, K., and Warren, J. (2021). The importance of STEM: High school knowledge, skills and occupations in an era of growing inequality. *Research Policy*, 50(7):104249.

Bonnot, V. and Croizet, J.-C. (2007). Stereotype internalization and women's math performance: The role of interference in working memory. *Journal of Experimental Social Psychology*, 43(6):857–866.

Brenøe, A. and Zölitz, U. (2020). Exposure to more female peers widens the gender gap in STEM participation. *Journal of Labor Economics*, 38(4):1009–1054.

Burkam, D., Lee, V., and Smerdon, B. (1997). Gender and science learning early in high school: Subject matter and laboratory experiences. *American Educational Research Journal*, 34(2):297–331.

Buser, T., Niederle, M., and Oosterbeek, H. (2014). Gender, competitiveness, and career choices. *Quarterly Journal of Economics*, 129(3):1409–1447.

Buser, T. and Yuan, H. (2019). Do women give up competing more easily? Evidence from the lab and the Dutch Math Olympiad. *American Economic Journal: Applied Economics*, 11(3):225–252.

Cai, X., Lu, Y., Pan, J., and Zhong, S. (2019). Gender gap under pressure: Evidence from China's National College entrance examination. *Review of Economics and Statistics*, 101(2):249–263.

Card, D. and Payne, A. (2021). High school choices and the gender gap in STEM. *Economic Inquiry*, 59(1):9–28.

Carrell, S., Page, M., and West, J. (2010). Sex and science: How professor gender perpetuates the gender gap. *Quarterly Journal of Economics*, 125(3):1101–1144.

Cimpian, J., Kim, T., and McDermott, Z. (2020). Understanding persistent gender gaps in STEM. *Science*, 368(6497):1317–1319.

Contini, D., Di Tommaso, M., and Mendolia, S. (2017). The gender gap in mathematics achievement: Evidence from Italian data. *Economics of Education Review*, 58:32–42.

Dasgupta, A. and Sharma, A. (2022). Preferences or expectations: Understanding the gender gap in major choice. *Oxford Economic Papers*, 74(4):1167–1194.

Davison, C. and Dustova, G. (2017). A quantitative assessment of student performance and examination format. *Journal of Instructional Pedagogies*, 18.

Delaney, J. and Devereux, P. (2019). Understanding gender differences in STEM: Evidence from college applications. *Economics of Education Review*, 72:219–238.

Delaney, J. and Devereux, P. (2021). High school rank in math and English and the gender gap in STEM. *Labour Economics*, 69:101969.

Dercon, S. and Singh, A. (2013). From nutrition to aspirations and self-efficacy: Gender bias over time among children in four countries. *World Development*, 45:31–50.

Ellison, G. and Swanson, A. (2010). The gender gap in secondary school mathematics at high achievement levels: Evidence from the American mathematics competitions. *Journal of Economic Perspectives*, 24(2):109–128.

Friedman-Sokuler, N. and Justman, M. (2020). Gender, culture and STEM: Counter-intuitive patterns in Arab society. *Economics of Education Review*, 74:101947.

Fryer, R. and Levitt, S. (2010). An empirical analysis of the gender gap in mathematics. *American Economic Journal: Applied Economics*, 2(2):210–240.

Gemici, A. and Wiswall, M. (2014). Evolution of gender differences in post-secondary human capital investments: College majors. *International Economic Review*, 55(1):23–56.

Goulas, S., Griselda, S., and Megalokonomou, R. (2024). Comparative advantage and gender gap in STEM. *Journal of Human Resources*, 59(6):1937–1980.

Griselda, S. (2024). Gender gap in standardized tests: What are we measuring? *Journal of Economic Behavior & Organization*, 221:191–229.

Hyde, J., Lindberg, S., Linn, M., Ellis, A., and Williams, C. (2008). Gender similarities characterize math performance. *Science*, 321(5888):494–495.

Jain, T., Mukhopadhyay, A., Prakash, N., and Rakesh, R. (2022). Science education and labor market outcomes in a developing economy. *Economic Inquiry*, 60(2):741–763.

Kahn, S. and Ginther, D. (2017). Women and Science, Technology, Engineering, and Mathematics (STEM): Are differences in education and careers due to stereotypes, interests, or family? In *The Oxford Handbook of Women and the Economy*, pages 767–798. Oxford University Press.

Klein, D. (2018). Implementing a general framework for assessing interrater agreement in Stata. *The Stata Journal*, 18(4):871–901.

Kuuppelomäki, T., Kortelainen, M., Suhonen, T., and Virtanen, H. (2025). Labor market returns to elite STEM education. *Journal of Human Capital*.

Lee, D. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *Review of Economic Studies*, 76(3):1071–1102.

Martinot, P., Colnet, B., Breda, T., Sultan, J., Touitou, L., Huguet, P., Spelke, E., Dehaene-Lambertz, G., Bressoux, P., and Dehaene, S. (2025). Rapid emergence of a maths gender gap in first grade. *Nature*, pages 1–10.

Murphy, K., Shleifer, A., and Vishny, R. (1991). The allocation of talent: Implications for growth. *Quarterly Journal of Economics*, 106(2):503–530.

Niederle, M. and Vesterlund, L. (2010). Explaining the gender gap in math test scores: The role of competition. *Journal of Economic Perspectives*, 24(2):129–144.

Nollenberger, N., Rodriguez-Planas, N., and Sevilla, A. (2016). The math gender gap: The role of culture. *American Economic Review*, 106(5):257–261.

Nosek, B., Smyth, F., Sriram, N., Lindner, N., Devos, T., Ayala, A., Bar-Anan, Y., Bergh, R., Cai, H., Gonsalkorale, K., et al. (2009). National differences in gender–science stereotypes predict national sex differences in science and math achievement. *Proceedings of the National Academy of Sciences*, 106(26):10593–10597.

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

Park, H., Behrman, J., and Choi, J. (2018). Do single-sex schools enhance students' STEM (science, technology, engineering, and mathematics) outcomes? *Economics of Education Review*, 62:35–47.

Pope, D. and Sydnor, J. (2010). Geographic variation in the gender differences in test scores. *Journal of Economic Perspectives*, 24(2):95–108.

Porter, C. and Serra, D. (2020). Gender differences in the choice of major: The importance of female role models. *American Economic Journal: Applied Economics*, 12(3):226–254.

Reardon, S., Kalogrides, D., Fahle, E., Podolsky, A., and Zárate, R. (2018). The relationship between test item format and gender achievement gaps on math and ELA tests in fourth and eighth grades. *Educational Researcher*, 47(5):284–294.

Reed, D. K., Petscher, Y., and Truckenmiller, A. J. (2017). The contribution of general reading ability to science achievement. *Reading Research Quarterly*, 52(2):253–266.

Roychowdhury, P. (2021). (Em) powered by science? Estimating the causal effect of high school major choice on labor market earnings in India. *Economics of Education Review*, 82:102118.

Royer, J., Tronsky, L., Chan, Y., Jackson, S., and Marchant III, H. (1999). Math-fact retrieval as the cognitive mechanism underlying gender differences in math test performance. *Contemporary Educational Psychology*, 24(3):181–266.

Sahoo, S. and Klasen, S. (2018). Gender segregation in education and its implications for labour market outcomes: Evidence from India. IZA Discussion Papers No. 11660.

Sharma, S. and Sengupta, N. (2024). Debunking myths associated with affirmative action: Using evidence from the Supernumerary Seats Scheme. Working Paper.

Shi, Y. (2018). The puzzle of missing female engineers: Academic preparation, ability beliefs, and preferences. *Economics of Education Review*, 64:129–143.

Stoet, G. and Geary, D. (2013). Sex differences in mathematics and reading achievement are inversely related: Within-and across-nation assessment of 10 years of PISA data. *PloS one*, 8(3):e57988.

Wei, W., Lu, H., Zhao, H., Chen, C., Dong, Q., and Zhou, X. (2012). Gender differences in children's arithmetic performance are accounted for by gender differences in language abilities. *Psychological science*, 23(3):320–330.

World Bank (2012). World development report 2012: Gender equality and development. Technical report, World Bank.

World Bank (2018). World development report 2018: Learning to realize education's promise. Technical report, World Bank.

Zafar, B. (2013). College major choice and the gender gap. *Journal of Human Resources*, 48(3):545–595.

Tables

Table 1: Summary statistics

Grade	No. of student-exam obs	No. of student-question obs	No. of questions per exam	% of girls in sample	% of boys in sample	Prob. of correct answer: Girls	Boys	Diff.
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
				Math				
3	78,409	23,52,270	30	46%	54%	0.50	0.51	-0.01
4	77,494	30,99,760	40	46%	54%	0.46	0.48	-0.01
5	84,180	33,67,200	40	46%	54%	0.45	0.47	-0.02
6	77,667	31,06,680	40	46%	54%	0.40	0.42	-0.02
7	73,581	29,43,240	40	46%	54%	0.41	0.43	-0.02
8	72,772	29,10,880	40	46%	54%	0.39	0.42	-0.03
9	46,370	18,48,128	40	46%	54%	0.36	0.39	-0.03
10	21,352	8,50,221	40	48%	52%	0.35	0.38	-0.03
				Science				
3	75,494	22,64,820	30	46%	54%	0.48	0.47	0.01
4	73,381	25,68,335	35	46%	54%	0.50	0.49	0.01
5	81,004	28,35,140	35	46%	54%	0.47	0.46	0.01
6	76,185	34,18,167	45	46%	54%	0.44	0.43	0.01
7	72,291	32,53,095	45	46%	54%	0.42	0.42	-0.00
8	70,802	31,86,090	45	46%	54%	0.40	0.41	-0.01
9	44,703	20,11,635	45	47%	53%	0.39	0.40	-0.01
10	20,336	9,12,790	45	48%	52%	0.38	0.40	-0.02
				English				
3	78,471	27,46,485	35	46%	54%	0.53	0.50	0.03
4	76,087	38,04,350	50	46%	54%	0.54	0.50	0.04
5	83,516	41,75,800	50	46%	54%	0.53	0.49	0.04
6	77,608	46,56,480	60	46%	54%	0.53	0.49	0.04
7	73,574	44,14,440	60	46%	54%	0.53	0.48	0.05
8	72,595	43,55,700	60	46%	54%	0.53	0.48	0.04
9	47,224	33,05,680	70	46%	54%	0.49	0.45	0.04
10	21,427	14,99,890	70	48%	52%	0.52	0.48	0.04

Data source: ASSET Exam database (2022-24). Notes: This table shows the summary statistics of the sample used in this study. Columns (2) and (3) contain information on the number of exam-level and question-level observations, respectively, for each grade mentioned in the first column. Column (4) shows the number of questions per exam. Columns (5) and (6) show the share of girls and boys in the sample. Columns (7), (8), and (9) show the probability of a correct answer for girls, boys, and the difference between them.

Table 2: Share of math-intensive questions

Grade	**No. of Qs**	**Share of math-intensive questions**				
		AI	**KW**	**Ind 1**	**Ind 2**	**Avg.**
(1)	(2)	(3)	(4)	(5)	(6)	(7)
3	180	12.8%	6.1%	12.8%	8.3%	10.0%
4	210	9.5%	3.8%	8.1%	4.8%	6.5%
5	210	11.0%	5.2%	7.6%	7.1%	7.7%
6	270	14.1%	10.4%	11.5%	8.9%	11.2%
7	270	12.2%	14.8%	15.9%	9.6%	13.1%
8	270	18.9%	19.6%	21.5%	14.4%	18.6%
9	270	17.8%	23.0%	21.1%	15.2%	19.3%
10	270	26.3%	25.2%	18.5%	23.0%	23.2%
Total	1950	15.7%	14.4%	15.1%	11.9%	14.3%

Source: 1950 unique science questions from the ASSET exam database (2022-24). We classify each question into math-intensive or not using the different approaches described in section 2.3.

Table 3: Gender gap in math-intensive science questions

	Dependent variable: question correct				
VARIABLES	**AI**	**KW**	**Ind1**	**Ind2**	**Avg**
Math-intensive	-0.016***	-0.020***	-0.036***	-0.016***	-0.040***
	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)
Girl x Math-intensive	-0.021***	-0.005***	-0.020***	-0.023***	-0.028***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Mean of dependent variable	0.438	0.438	0.438	0.438	0.438
	(0.496)	(0.496)	(0.496)	(0.496)	(0.496)
Difficulty level	✓	✓	✓	✓	✓
Qno	✓	✓	✓	✓	✓
Language complexity	✓	✓	✓	✓	✓
Student FE	✓	✓	✓	✓	✓
Grade FE	✓	✓	✓	✓	✓
School FE	✓	✓	✓	✓	✓
Year FE	✓	✓	✓	✓	✓
Skill tested FE	✓	✓	✓	✓	✓
Constant	0.681***	0.680***	0.683***	0.680***	0.682***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	20,450,072	20,450,072	20,450,072	20,450,072	20,450,072
Adjusted R-squared	0.099	0.099	0.100	0.099	0.100

Source: Asset exam database (2022-24). Each column shows the estimates of Equation 2 for each classification method separately. Robust standard errors in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$

Table 4: Gender gap in math-intensive science questions by grade (Avg classification)

	Dependent variable: question correct								
VARIABLES	**Overall**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Math-intensive Avg	-0.040***	-0.115***	0.029***	-0.035***	0.021***	-0.040***	-0.026***	-0.036***	-0.054***
	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)	(0.002)
Girl x Math-intensive Avg	-0.028***	-0.040***	-0.032***	-0.058***	-0.031***	-0.012***	-0.023***	-0.022***	-0.019***
	(0.001)	(0.003)	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)
Math-intensive Avg proportion	14%	10%	7%	8%	11%	13%	19%	19%	23%
Mean of dependent variable	0.438	0.472	0.493	0.469	0.436	0.419	0.407	0.393	0.391
	(0.496)	(0.499)	(0.500)	(0.499)	(0.496)	(0.493)	(0.491)	(0.488)	(0.488)
Difficulty level	✓	✓	✓	✓	✓	✓	✓	✓	✓
Qno	✓	✓	✓	✓	✓	✓	✓	✓	✓
Language complexity	✓	✓	✓	✓	✓	✓	✓	✓	✓
Student FE	✓	✓	✓	✓	✓	✓	✓	✓	✓
Grade FE	✓	×	×	×	×	×	×	×	×
School FE	✓	✓	✓	✓	✓	✓	✓	✓	✓
Year FE	✓	✓	✓	✓	✓	✓	✓	✓	✓
Skill tested FE	✓	✓	✓	✓	✓	✓	✓	✓	✓
Constant	0.682***	0.801***	0.789***	0.719***	0.564***	0.637***	0.657***	0.610***	0.650***
	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)
Observations	20,450,072	2,264,820	2,568,335	2,835,140	3,418,167	3,253,095	3,186,090	2,011,635	912,790
Adjusted R-squared	0.100	0.145	0.138	0.112	0.101	0.090	0.092	0.076	0.084

Source: Asset exam database (2022-24). The "Overall" column shows the estimates of Equation 2 for the Avg measure of classification. The subsequent columns show the results estimated for each grade separately. Robust standard errors in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$

Table 5: Question distribution by branch

Grade	Science	Physics	Chemistry	Biology
	Number of questions			
3	180	80	14	86
4	210	89	33	88
5	210	107	15	88
6	270	134	15	121
7	270	143	32	95
8	270	135	39	96
9	270	130	38	102
10	270	124	44	102
Total	1950	942	230	778
	Share of math-intensive questions			
3	10.0%	13.8%	12.5%	6.1%
4	6.5%	11.0%	1.5%	4.0%
5	7.7%	12.9%	5.0%	2.0%
6	11.2%	13.1%	6.7%	9.7%
7	13.1%	16.4%	13.3%	8.2%
8	18.6%	27.6%	11.5%	8.9%
9	19.3%	24.0%	24.3%	11.3%
10	23.2%	29.2%	21.6%	16.7%
Total	14.3%	19.1%	13.7%	8.6%

Source: 1950 science questions from the ASSET exam database (2022-24). The upper panel shows the number of questions in the science subject and the distribution of physics, chemistry, and biology therein. Similarly, the lower panel shows the share of math-intensive questions for the science subject overall and each branch. For brevity and ease of readability, we only show this as per the Avg measure of classification.

Table 6: Gender gap in science exams

	Dependent variable: percentage in exam					
VARIABLES	**AI**	**KW**	**Ind1**	**Ind2**	**Avg**	**Binary**
Math-intensive proportion	-0.299***	-0.683***	-0.471***	-0.513***	-0.863***	-0.071***
	(0.009)	(0.007)	(0.005)	(0.009)	(0.010)	(0.001)
Girl x Math-intensive proportion	-0.022*	-0.031***	-0.022***	-0.053***	-0.048***	-0.003**
	(0.012)	(0.010)	(0.008)	(0.013)	(0.013)	(0.002)
Mean of dependent variable	0.443	0.443	0.443	0.443	0.443	0.443
	(0.171)	(0.171)	(0.171)	(0.171)	(0.171)	(0.171)
Student FE	✓	✓	✓	✓	✓	✓
School FE	✓	✓	✓	✓	✓	✓
Year FE	✓	✓	✓	✓	✓	✓
Constant	0.495***	0.536***	0.520***	0.504***	0.559***	0.493***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)
Observations	275,671	275,671	275,671	275,671	275,671	275,671
Adjusted R-squared	0.604	0.633	0.631	0.613	0.634	0.615

Source: Asset exam database (2022-24). Each column shows the estimates of equation (3) for each classification method separately, as discussed in section 4.2. Robust standard errors in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$

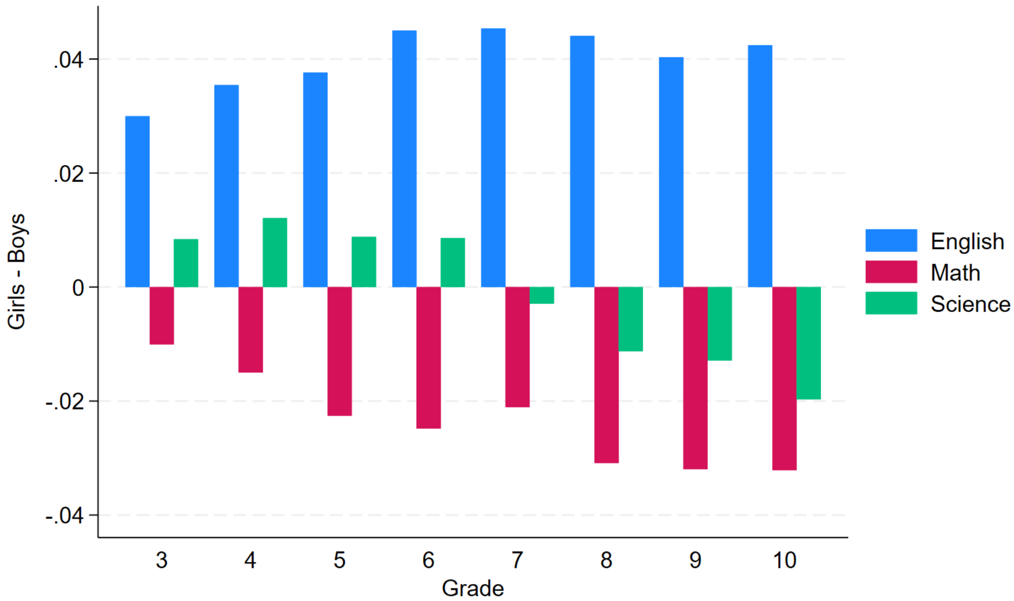
Table 7: Attrition in sample

Particulars	**Schools**	**Students**
Observed in all three years	135	31,360
Observed in 2022 and 2023	94	33,514
Observed in 2023 and 2024	52	37,225
Observed in 2022 and 2024	5	3,163
Observed only in 2022	128	74,552
Observed only in 2023	71	70,732
Observed only in 2024	106	1,02,891
Total	591	3,53,437

Source: ASSET exam database (2022-24). This table shows the number of schools and students appearing in the sample for each year and across year combinations.

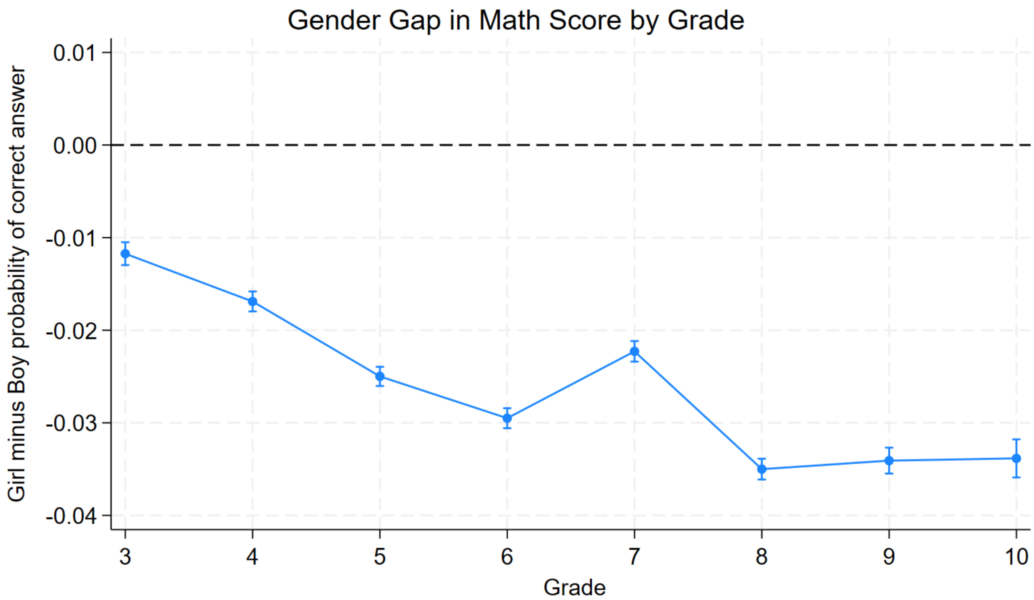
Figures

Figure 1: Summary of scores by subject, grade and gender



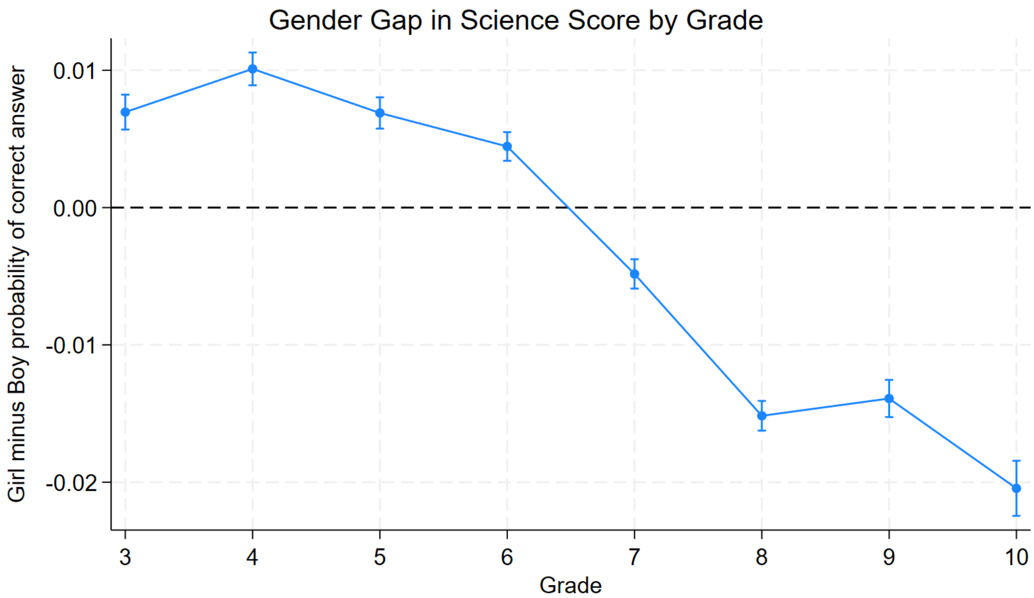
Data source: ASSET Exam database (2022-24). Notes: The figure shows the difference in the probability of answering a question correctly (girls minus boys) by grade for science, mathematics, and English. Boys outperform girls in all grades in mathematics, and girls outperform boys in all grades in English. Finally, in science, girls perform better than boys in grades 3 to 6, but this pattern is reversed in grades 7 to 10.

Figure 2: Gender gap in mathematics



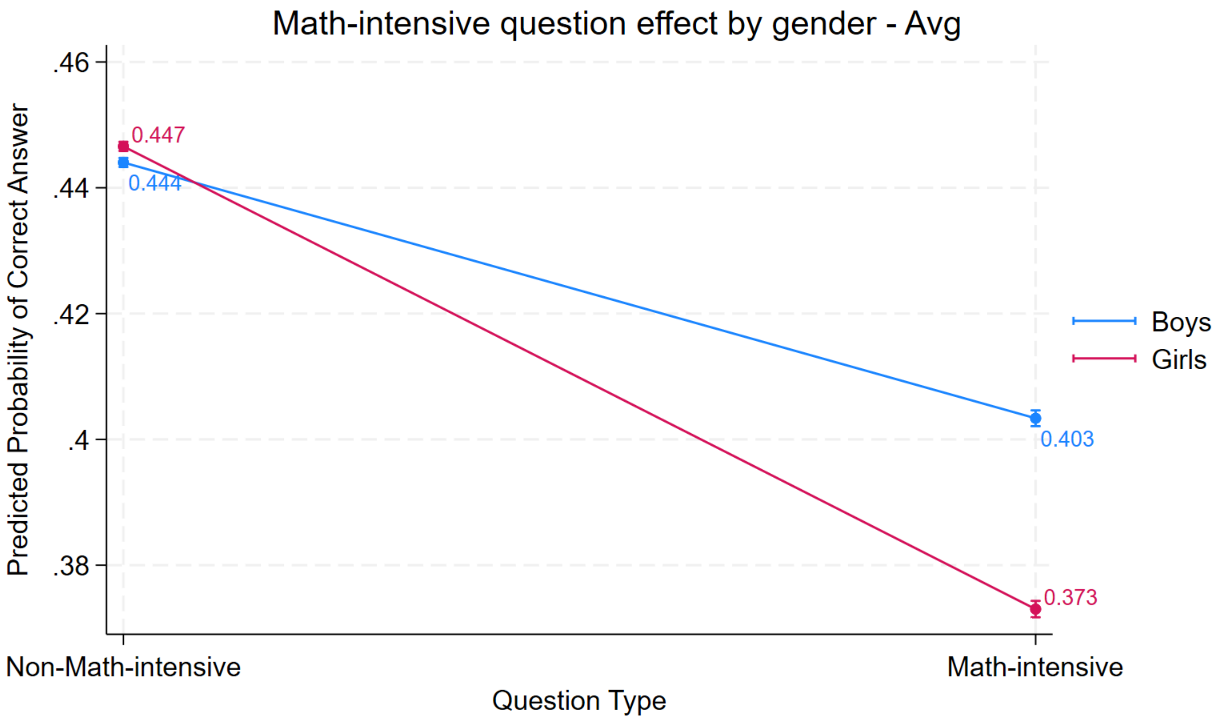
Data source: ASSET Exam database (2022-24). Notes: The figure depicts the difference in the probability of a correct answer between girls and boys in mathematics.

Figure 3: Gender gap in science



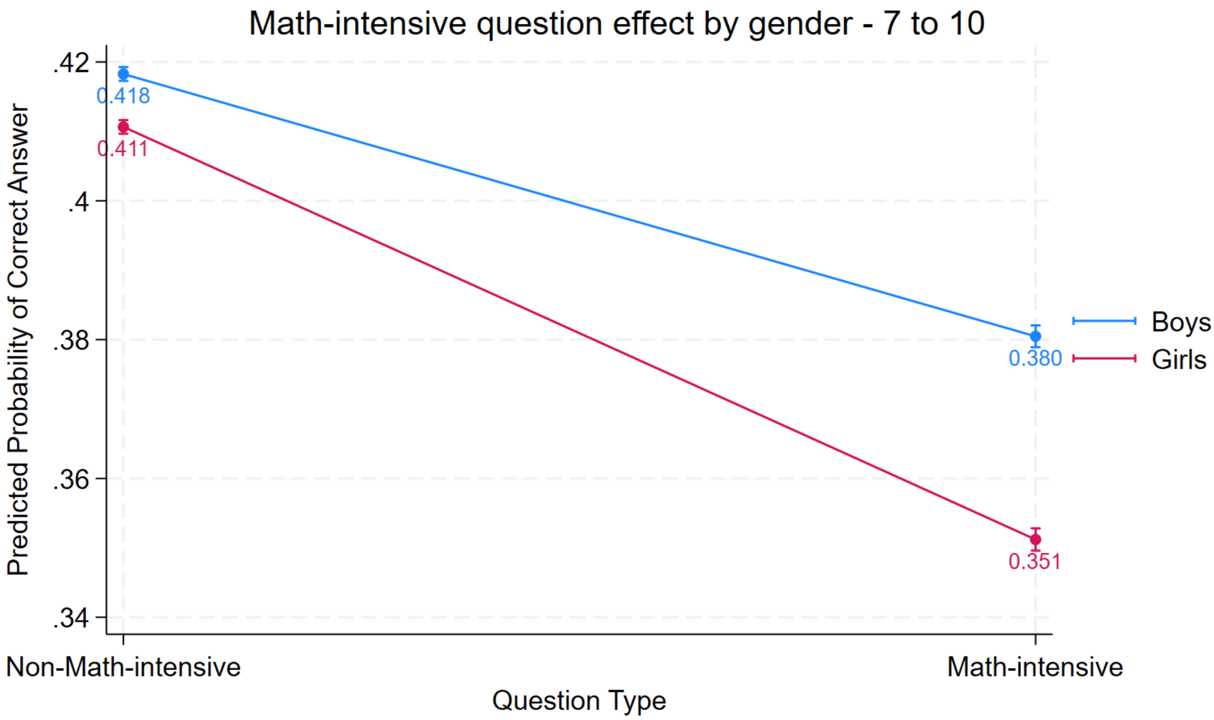
Data source: ASSET Exam database (2022-24). Notes: The figure depicts the difference in the probability of a correct answer between girls and boys in science.

Figure 4: Gender gap in science questions - all grades



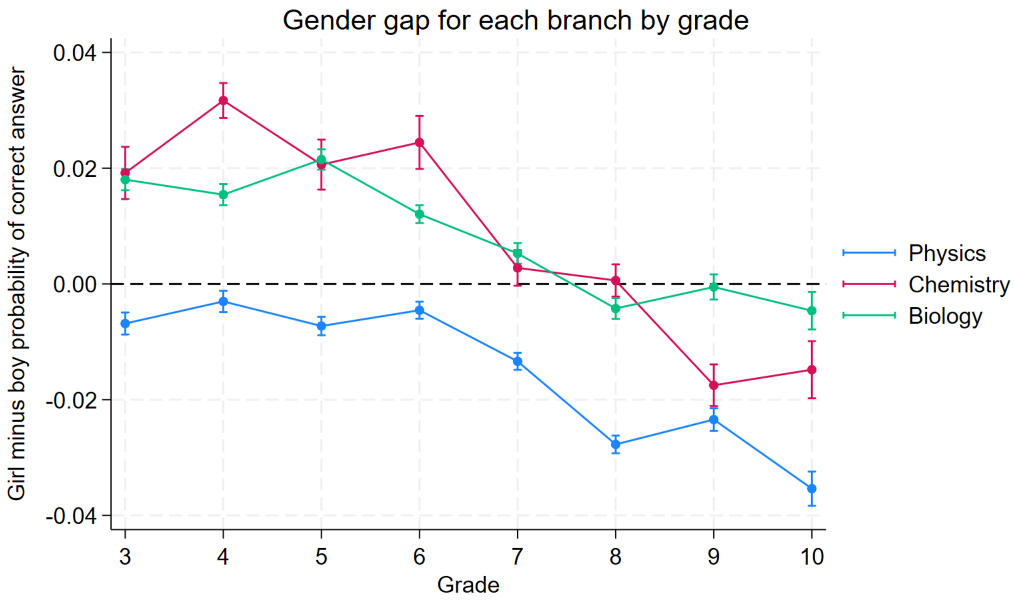
Data source: ASSET Exam database (2022-24) and the question classifications by the authors. Notes: The figure shows the probability of answering a math-intensive and a non-math-intensive science question correctly for boys and girls separately. On average, the probability of answering a non-math-intensive science question is around 44.5 percent for both boys and girls. However, for math-intensive science questions, boys are much more likely to get it right compared to girls. This difference is statistically significant as the confidence intervals do not overlap.

Figure 5: Gender gap in science questions - grades 7 to 10



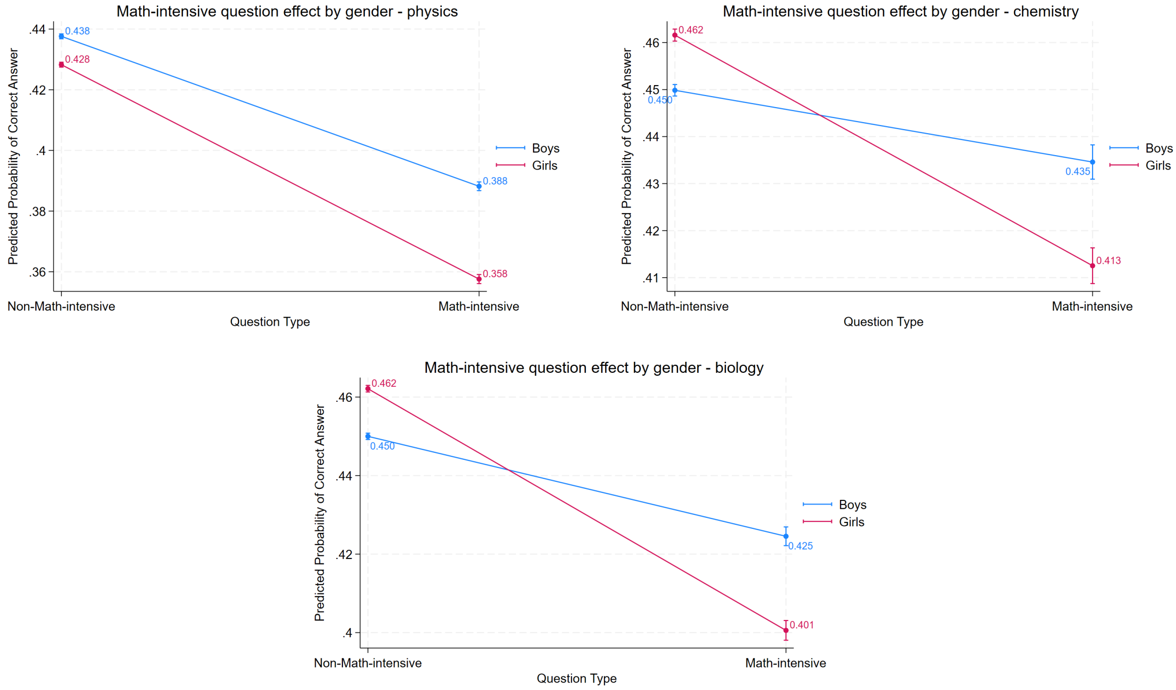
Data source: ASSET Exam database (2022-24) and the question classifications by the authors. Notes: The figure shows the probability of answering a math-intensive and a non-math-intensive science question correctly for boys and girls separately in grades 7 to 10. On average, the probability of answering a non-math-intensive science question is around 41 percent for both boys and girls. However, for math-intensive science questions, boys are much more likely to get it right compared to girls. This difference is statistically significant as the confidence intervals do not overlap.

Figure 6: Gender gap in science branches



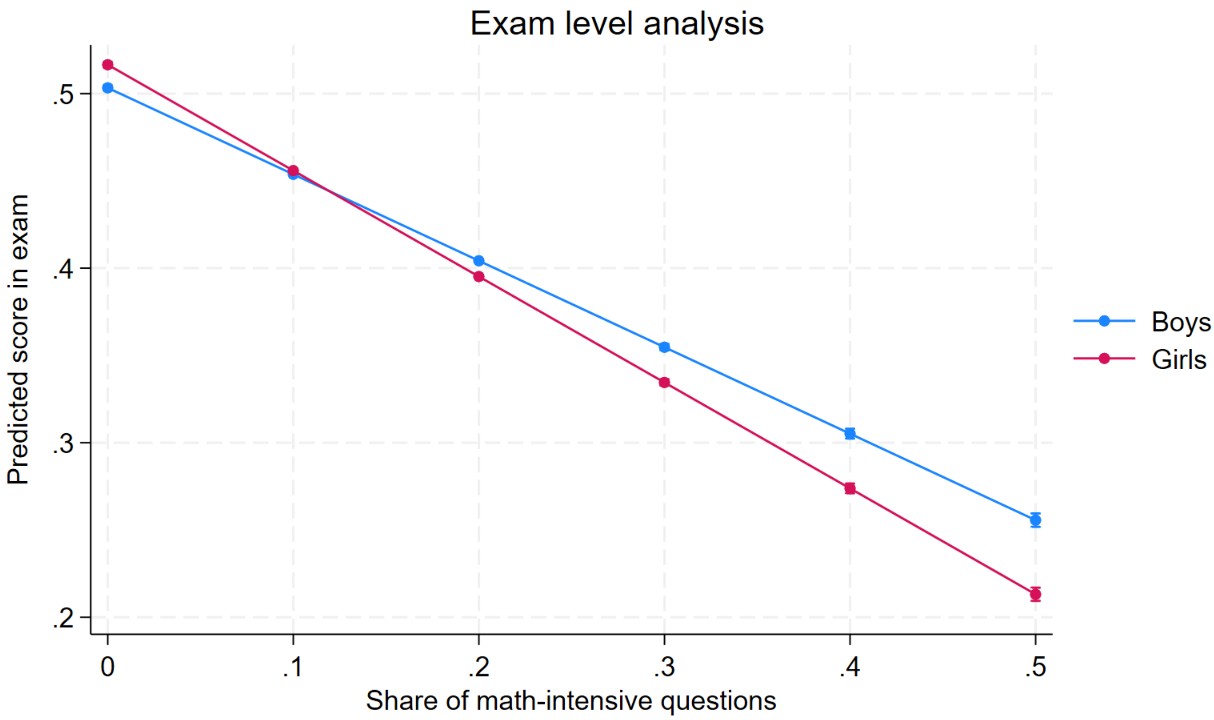
Data source: ASSET Exam database (2022-24). Notes: The figure depicts the difference in the probability of a correct answer between girls and boys in physics, chemistry, and biology.

Figure 7: Gender gap in science questions - across branches



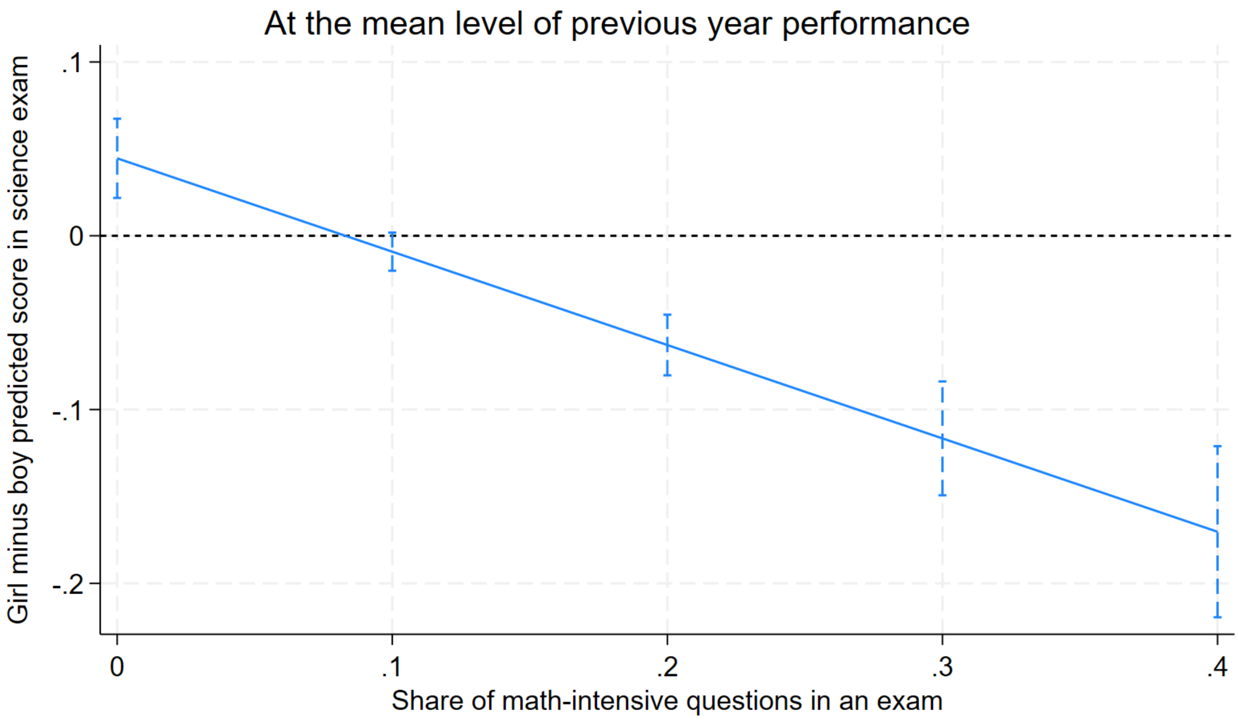
Data source: ASSET Exam database (2022-24) and the question classifications by the authors. Notes: The figure shows the probability of answering a math-intensive and a non-math-intensive science question correctly for boys and girls separately across the three branches of science (i.e., physics, chemistry, and biology). For non-math-intensive questions, girls do better than boys (except in physics, where the disadvantage is 1%). However, for math-intensive questions, boys are much more likely to get it right compared to girls in all three branches (a disadvantage of around 2.2% to 3%).

Figure 8: Gender gap in science exams



Data source: ASSET Exam database (2022-24) and the question classifications by the authors. Notes: The figure shows the relation of the proportion of math-intensive questions (using the average measure) in an exam with the score in that exam. When the share of math-intensive questions is lower than 10 percent, girls outperform boys, however, after the 20 percent threshold, boys outperform girls. Overall, as the share of math-intensive science questions increases, the scores of both boys and girls go down, but the decline is greater for girls than boys.

Figure 9: Value-added type conditional gap in science exams



Data source: ASSET Exam database (2022-24) and the question classifications by the authors. Notes: The figure shows the relation of the proportion of math-intensive questions (using the average measure) in an exam with the gender gap in the score in that exam, conditional on the previous year's performance being the same (at the mean). When the share of math-intensive questions is lower than 10 percent, girls outperform boys, however, after the 10 percent threshold, boys outperform girls. This is for only those students who had the same level of performance in the previous year's science exam. The difference in scores on the y-axis is expressed in standard deviations.

Appendix Tables

Table A1: Summary Statistics - Additional question-level variables

Grade	Mean (SD) Difficulty level	Q no.	Language complexity	Two most frequently tested skills
		Math		
3	1.92	15.50	2.02	Basic shapes and geometry
	(0.65)	(8.68)	(0.75)	Applications in daily life
4	2.02	20.50	2.20	Basic shapes and geometry
	(0.63)	(11.57)	(0.82)	Applications in daily life
5	2.01	20.50	2.39	Applications in daily life
	(0.64)	(11.57)	(0.88)	Number sense
6	2.04	20.50	2.49	Applications in daily life
	(0.62)	(11.57)	(0.94)	Arithmetic operations
7	2.03	20.50	2.92	Fractions, decimals and ratios
	(0.64)	(11.57)	(1.11)	Problem solving
8	2.05	20.50	3.07	Integers and rational numbers
	(0.63)	(11.57)	(1.09)	Fractions, decimals and ratios
9	2.08	20.50	2.73	Mensuration
	(0.61)	(11.57)	(1.02)	Algebra
10	2.12	20.50	2.89	Rational and irrational numbers
	(0.68)	(11.57)	(0.94)	Problem solving
		Science		
3	1.94	15.50	2.70	Data interpretation
	(0.65)	(8.68)	(0.78)	Scientific instruments & processes
4	1.98	18.00	3.11	Definition or description
	(0.65)	(10.12)	(0.78)	Classification or comparison
5	2.02	18.00	3.44	Hypothesis formulation
	(0.67)	(10.12)	(0.85)	Application of information
6	1.90	23.00	3.54	Application of information
	(0.62)	(13.01)	(0.79)	Classification or comparison
7	1.96	23.00	3.70	Application of information
	(0.64)	(13.01)	(0.98)	Identifying trends and properties
8	1.98	23.00	3.85	Application of information
	(0.68)	(13.01)	(0.87)	Scientific instruments & processes
9	1.95	23.00	3.94	Recollection of concepts
	(0.65)	(13.01)	(1.07)	Application of information
10	2.01	23.00	3.88	Application of information
	(0.67)	(13.01)	(0.89)	Scientific instruments & processes

Source: ASSET exam database (2022-24). Notes: This table shows the descriptive statistics of the question-level control variables for math and science by grade. Difficulty level is increasing on a scale of 1-3, and language complexity is increasing on a scale of 1-7. The last column shows the two most frequent skills tested.

Table A2: Frequency of math-intensive keywords in science questions by grade

Keyword/Grade	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Total**
Graphs/charts/plots	6	1	0	8	10	23	27	25	100
Units (current, force, mass, etc.)	4	3	7	9	14	10	14	18	79
Calculate/compute/ evaluate/determine	0	0	2	3	2	2	6	4	19
Estimation and approximation	1	2	0	3	4	1	3	3	17
Temperature in units	0	0	0	2	4	1	5	5	17
Ratio/proportion	0	0	0	0	0	3	4	2	9
Time - numbers	0	0	1	0	2	4	0	1	8
Averages (mean)	0	0	0	3	1	3	0	0	7
Conversion units	0	0	0	0	1	2	0	2	5
Percentage	0	0	0	0	1	2	2	0	5
Balance setup	0	0	0	0	0	0	0	4	4
Degrees and angles	0	1	0	0	1	1	0	1	4
At least/at most - numbers	0	0	1	0	0	0	1	1	3
Table with numbers	0	1	0	0	0	1	0	1	3
Work done phrase	0	0	0	0	0	0	0	1	1
Total	11	8	11	28	40	53	62	68	281

Source: ASSET exam database (2022-24). This table shows the number of questions where the respective mathematical/quantitative keyword appears. This includes 1950 science questions across all grades.

Table A3: Examples of math-intensive and non-math-intensive questions by grade

Grade	Non math-intensive	Math-intensive
3	Venus flytrap is a plant that can eat small insects. Which of the following is DEFINITELY true about a Venus flytrap plant?	One filled cup of cooked eggs gives 13 g of protein. The table below shows how many cups of different foods are needed to give the same amount of protein as one cup of cooked eggs. Which food item has the MOST amount of protein?
3	Sneha opened a window at 8 P.M. and saw the rising moon in front of her as shown below. Which direction was she facing?	Harsh collected data about the time required to cook some common food items by boiling and by frying on the same stove. Based on the table, which of the following is CORRECT?
4	I am sweet to taste.- I have many seeds.- I come in different colours.- I am usually found in a bunch.- Humans also eat my stem and flower.Who am I?	A two-pan balance is perfectly balanced when no weights are placed on its pans. An object of unknown weight is placed on pan Q. Fig. 1 shows the balance when a 4 kg weight is placed on pan P. Fig. 2 shows the balance when a 2 kg weight is placed on pan P. What is LIKELY to be the weight of the object placed on pan Q?
4	Paul was flying from one place to another by an aeroplane. He could see the Sun rising through the window on his left. Which of the following correctly shows the direction in which the plane is likely to be flying?	The diagram below shows the position of the Moon at its different phases in a complete lunar cycle. It takes about 30 days from one full moon to another. The time between which of the following phases will be CLOSEST to 15 days?
5	Wasim made three groups of animals. He saw an insect which he could place in ALL the three groups.Which of the following groups could he have made?	Suraj was travelling to Singapore from Bangalore. His flight from Bangalore took off at 3:00 AM Indian time and landed in Singapore 4 hours later. If Indian time is behind of time in Singapore by 2.5 hours, then what was the time in Singapore when his flight landed?
5	Anita knew that a harder material can scratch a softer material. She found out the following:- Marble stone or a knife blade CANNOT scratch window glass.- A knife blade can scratch gold and marble stone.Which among the following is the HARDEST?	The time taken by different planets to go once around the Sun is given below. Mercury: 88 days, Venus: 225 days, Earth: 1 year, Mars: 687 days, Jupiter: 12 years, Saturn: 30 years, Uranus: 84 years, Neptune: 165 years. How many planets would have completed AT LEAST one round around the Sun between March 1, 2018 to April 1, 2019?
6	Acrylic is a transparent plastic that is used instead of glass in some applications.Which of the following properties of acrylic would be a DISADVANTAGE if it is used to make lenses for spectacles?	Shown below is a chart showing the variation in the average weight, height and lifespan of different breeds of dogs. Which of these is TRUE, based on the chart?
6	Ravi grouped some animals into two groups based on the position of their eyes on the face. How else can he classify these animals into the SAME groups?	Abida wanted to check which threads were able to carry more weight - silk or polyester.She arranged the following set-up. The silk thread consisted of 10 strands of silk and the polyester thread consisted of 50 strands of polyester. Each individual strand of silk and polyester were of the same thickness. She first attached a 5 g weight to each of the two threads and then went on increasing this by 3 g till one of the threads broke.Would she have been able to correctly identify the thread that can carry more weight?
7	The diagram of the human lungs shows that the left lung is smaller than the right lung. The space marked by X is to accommodate an organ of the human body. Which organ could it be?	Shown below are some conversions for volume. 1 centilitre = 10 millilitres. 1 gallon = 379 centilitres. Which of the following volumes will be NEAREST to 1 litre?
7	The figure below shows how the compass needle is deflected when a bar X is brought near a magnetic compass. Which of the following could be true about the bar X? P) It is a bar magnet with its north pole facing the compass. Q) It is a bar magnet with its south pole facing the compass. R) It is a plain iron bar.	Sushil was filling a 30 cm tall vase with water from a tap. He measured and noted down the level of water filled every 20 seconds. Which of the following lines in the graph correctly indicates the possible data collected by Sushil?

8	The diagram below shows how an amoeba feeds by capturing a food particle. Which of these components of blood is most likely to perform its function in a similar way?	The pointer on the disc rotates by 15 degrees every half hour. How long will it take for the pointer to complete ONE complete rotation around the disc?
8	The potato peeler shown in Figure 1 has sharp edges on both sides of the centre strip. Figure 2 shows how this peeler is used to peel a potato. What could be the likely reason for both sides of the centre strip being sharp?	Which of the following line graphs could be representing the same data as shown in the chart for the year 2007?
9	The level indicator is connected to the boiler through two valves. Which of the valves should be open to get a correct reading on the level indicator?	A part of a ruler having markings in inches and centimetres is shown below. Aman measures the length of his pencil as X centimetres.Which of the following can he do to find out its length in inches? P) multiply X by a number which is less than 1. Q) multiply X by a number which is greater than 1. R) divide X by a number which is less than 1. S) divide X by a number which is greater than 1
9	A pencil is placed in front of a white screen. Two coloured lights giving out a beam of light, which spreads on the entire screen, are placed adjacent to each other as shown below. A pencil is placed between the sources of light and the screen as shown below. When both the lights are switched on in a dark room. two distinct shadows P and Q are observed on the screen. What is likely to be the colour of the shadows?	The graph below shows the percentage change in the length with temperature (from the length at 0 degC) of a rod made of English china clay. At which of the following temperatures, would a rod made of this clay be the SMALLEST in length?
10	Commensalism is a type of relationship between two organisms in which one organism benefits while the other is unaffected.Which of the following is an example of commensalism?	The graph below shows the global extraction of natural resources from ecosystems and mines between the period 1980 to 2005. Based on the graph, which of the following is TRUE?
10	Nanda mixes 50 g each of two miscible liquids, Liquid X (boiling point 56 degC) and Liquid Y (boiling point 110 degC) in a flask. He wants to find if a chemical change has occurred on mixing the two liquids. Which of the following should he do to confirm if mixing the two liquids causes a chemical change?	Rocima 101 is a biocidal agent. It is used to kill or control the spread of harmful microbes. Below is a graph representing the effect of 6% and 8% of this biocide on the growth of bacteria and mould when applied once, twice and three times. Which of the following is TRUE based on this?

Source: ASSET exam database (2022-24). This table provides examples of science questions classified as non-math-intensive (Avg measure = 0) and math-intensive (Avg measure = 1) across grades. These examples reflect cases where all classification methods agreed on the math-intensity status of the question.

Table A4: Interrater agreement

Particulars	**Coef.**	**Std.Err.**	**[95%Conf.**	**Interval]**
Percent Agreement	0.88	0.01	0.87	0.89
Brennan and Prediger	0.76	0.01	0.74	0.78
Cohen/Conger's Kappa	0.52	0.02	0.48	0.55
Scott/Fleiss' Kappa	0.52	0.02	0.48	0.55
Gwet's AC	0.84	0.01	0.83	0.86
Krippendorff's Alpha	0.52	0.02	0.48	0.55

Data source: The science questions are sourced from the ASSET exam database (2022-24) and the classification variables are created as part of the analysis in this study. Notes: This table shows the level of agreement or consistency between the four classification methods (AI, KW, Ind 1, and Ind 2) using Klein (2018)'s reliability tests.

Table A5: Gender gap in math-intensive science questions - coed school subsample

	Dependent variable: question correct				
VARIABLES	**AI**	**KW**	**Ind1**	**Ind2**	**Avg**
Math-intensive	-0.016***	-0.020***	-0.037***	-0.016***	-0.040***
	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)
Girl x Math-intensive	-0.021***	-0.006***	-0.020***	-0.023***	-0.027***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Mean of dependent variable	0.439	0.439	0.439	0.439	0.439
	(0.496)	(0.496)	(0.496)	(0.496)	(0.496)
Difficulty level	✓	✓	✓	✓	✓
Qno	✓	✓	✓	✓	✓
Language complexity	✓	✓	✓	✓	✓
Student FE	✓	✓	✓	✓	✓
Grade FE	✓	✓	✓	✓	✓
School FE	✓	✓	✓	✓	✓
Year FE	✓	✓	✓	✓	✓
Skill tested FE	✓	✓	✓	✓	✓
Constant	0.682***	0.681***	0.684***	0.681***	0.683***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	19,967,100	19,967,100	19,967,100	19,967,100	19,967,100
Adjusted R-squared	0.100	0.100	0.100	0.100	0.100

Source: Asset exam database (2022-24). Each column shows the estimates of Equation 2 for co-ed schools. Robust standard errors in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$

Table A6: Main result - robustness with school × year clustering and FE

	Dependent variable: question correct				
VARIABLES	**AI**	**KW**	**Ind1**	**Ind2**	**Avg**
Math-intensive	-0.015***	-0.018***	-0.034***	-0.014***	-0.037***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
Girl x Math-intensive	-0.021***	-0.005***	-0.020***	-0.023***	-0.027***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Mean of dependent variable	0.443	0.443	0.443	0.443	0.443
	(0.171)	(0.171)	(0.171)	(0.171)	(0.171)
Difficulty level	✓	✓	✓	✓	✓
Qno	✓	✓	✓	✓	✓
Language complexity	✓	✓	✓	✓	✓
Student FE	✓	✓	✓	✓	✓
Grade FE	✓	✓	✓	✓	✓
School × Year FE	✓	✓	✓	✓	✓
Skill tested FE	✓	✓	✓	✓	✓
Constant	0.677***	0.676***	0.679***	0.676***	0.678***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Observations	20,450,072	20,450,072	20,450,072	20,450,072	20,450,072
Adjusted R-squared	0.101	0.101	0.101	0.101	0.101

Source: Asset exam database (2022-24). This is the same question-level specification discussed in section 3.2 with School × Year fixed effects, and clustering of standard errors at the School × Year level. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table A7: School-level attrition

Variables	Dependent variable: 'always-in' school
Score percentage science	0.076
	(0.048)
Co-ed school	0.238
	(0.166)
Pincode FE	✓
Board FE	✓
Constant	-0.037
	(0.259)
Observations	275
Adjusted R-squared	0.291

Source: Asset exam database (2022-24). This table shows the relationship between school-level performance in science and the probability of being observed in the sample for all three years. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table A8: Student-level attrition

Variables	**Dependent variable: 'always-in' student**
Girl	0.013***
	(0.003)
Score percentage science	0.062***
	(0.010)
Girl x score percentage	-0.009
	(0.006)
School FE	✓
Constant	0.059***
	(0.004)
Observations	354,756
Adjusted R-squared	0.222

Source: Asset exam database (2022-24). This table shows the association of the student's gender and performance in science with the probability of being observed in the sample for all three years. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table A9: Gender gap in math-intensive science questions - 'always-in' school sub-sample

	Dependent variable: question correct				
VARIABLES	**AI**	**KW**	**Ind1**	**Ind2**	**Avg**
Math-intensive	-0.014***	-0.020***	-0.031***	-0.012***	-0.035***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Girl x Math-intensive	-0.021***	-0.006***	-0.021***	-0.022***	-0.028***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Mean of dependent variable	0.449	0.449	0.449	0.449	0.449
	(0.497)	(0.497)	(0.497)	(0.497)	(0.497)
Difficulty level	✓	✓	✓	✓	✓
Qno	✓	✓	✓	✓	✓
Language complexity	✓	✓	✓	✓	✓
Student FE	✓	✓	✓	✓	✓
Grade FE	✓	✓	✓	✓	✓
School FE	✓	✓	✓	✓	✓
Year FE	✓	✓	✓	✓	✓
Skill tested FE	✓	✓	✓	✓	✓
Constant	0.697***	0.696***	0.699***	0.696***	0.698***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	9,905,537	9,905,537	9,905,537	9,905,537	9,905,537
Adjusted R-squared	0.100	0.100	0.101	0.100	0.101

Source: Asset exam database (2022-24). Each column shows the estimates of Equation 2 for those schools that are observed in the sample in all three years. Robust standard errors in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$

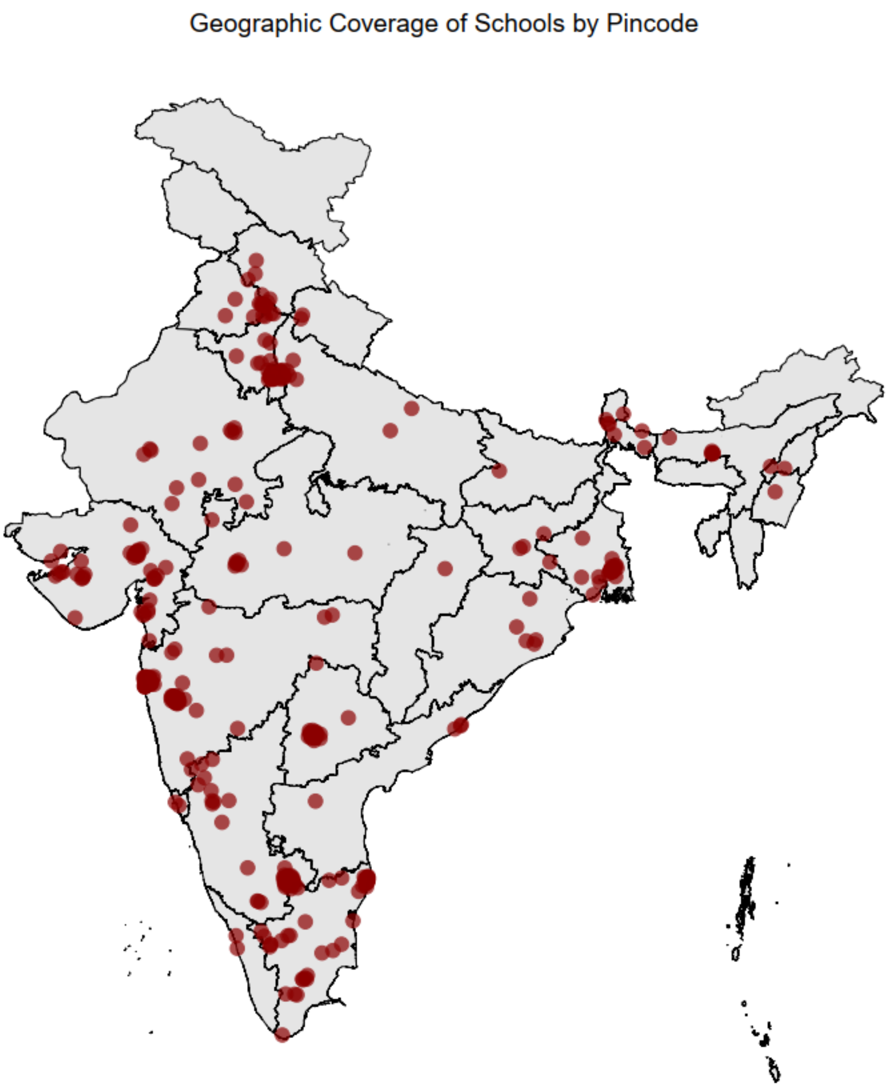
Table A10: Gender gap in math-intensive science questions - 'always-in' student sub-sample

	Dependent variable: question correct				
VARIABLES	**AI**	**KW**	**Ind1**	**Ind2**	**Avg**
Math-intensive	-0.008***	-0.023***	-0.023***	-0.005***	-0.028***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Girl x Math-intensive	-0.022***	-0.004***	-0.020***	-0.023***	-0.027***
	(0.002)	(0.002)	(0.001)	(0.002)	(0.002)
Mean of dependent variable	0.460	0.460	0.460	0.460	0.460
	(0.498)	(0.498)	(0.498)	(0.498)	(0.498)
Difficulty level	✓	✓	✓	✓	✓
Qno	✓	✓	✓	✓	✓
Language complexity	✓	✓	✓	✓	✓
Student FE	✓	✓	✓	✓	✓
Grade FE	✓	✓	✓	✓	✓
School FE	✓	✓	✓	✓	✓
Year FE	✓	✓	✓	✓	✓
Skill tested FE	✓	✓	✓	✓	✓
Constant	0.717***	0.717***	0.718***	0.716***	0.718***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	3,855,646	3,855,646	3,855,646	3,855,646	3,855,646
Adjusted R-squared	0.097	0.097	0.097	0.096	0.097

Source: Asset exam database (2022-24). Each column shows the estimates of Equation 2 for those students who are observed in the sample in all three years. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Appendix Figures

Figure A1: School coverage map



Source: The shape file for the base map is sourced from the Survey of India official website. Information on school pin codes is obtained from the ASSET exam database (2022-24). Notes: The figure shows the geographical coverage of 591 schools that took the ASSET exam during 2022-24. The coverage is fairly spread throughout the country across 147 cities in 26 states.

Figure A2: GPT-4o: example prompts and responses

Hi, I will be giving you a set of science questions meant for the school students of grade 3-10 and your task is: Whether the question is mathematics based or not (requiring the student to apply math knowledge for the particular grade the student is in). Please use a binary classification where 1 indicates the question requires mathematical reasoning and 0 if it does not. Additionally you have to give me the language complexity of the question based on the grade of the student. You must give it on a scale from 1 to 10, where 1 denotes extremely simple language and 10 denotes very complex language, for the target grade. Additionally, I want you to classify each question into one of the following disciplinary categories based on its core concept: Physics, Chemistry, Biology

Sure, please go ahead and share the science questions. For each one, I will return:

- **Math-based**: 1 or 0
- **Language complexity**: Score from 1 to 10 (relative to the given grade)
- **Disciplinary category**: Physics / Chemistry / Biology

✅ Criteria for Classifying as Math-Based (1):

A question is labeled math-based if **mathematical reasoning or computation appropriate to the student's grade** is required to answer it. This includes:

1. **Numerical Calculations**
   - Arithmetic: addition, subtraction, multiplication, division
   - Ratios, proportions, averages
   - Unit conversions (e.g., cm to m, g to kg)
   - Percentages (for older grades)
2. **Quantitative Problem-Solving**
   - Word problems involving measurable quantities
   - Speed = distance/time, force = mass × acceleration, etc.
   - Using formulas with substitution and computation
3. **Graph or Data Interpretation**
   - Reading or interpreting tables, graphs, charts
   - Estimating or comparing values numerically
4. **Geometrical Reasoning**
   - Shapes, area, volume in science contexts (e.g., calculating volume of water)
5. **Scientific Measurement**
   - Questions involving precision, significant figures, or units in a computational context
6. **Proportional or Algebraic Thinking**
   - Growth rates, chemical ratios, balance of equations (when numerical)

❌ Criteria for Not Math-Based (0):

A question is not math-based if it:

- Requires only conceptual understanding or factual recall
- Involves reasoning without numerical or formal quantitative manipulation
- Involves qualitative comparisons (e.g., heavier/lighter, hotter/colder)
- Uses numbers for context only (not for computation)

✅ Classification for Grade 7 Question:

Question (Cleaned):
Roger pours 500 mL of muddy water into each of two identical glass beakers. He keeps one beaker on the kitchen table and the other one inside the refrigerator. He checks the appearance of the water in both the beakers every 5 minutes.
What is he most likely trying to find out?

- **Math-based:** 0
  This question does not require mathematical reasoning. It focuses on observation and inference about physical processes.
- **Language complexity (Grade 7):** 3 / 10
  The language is simple and direct. The sentence structure and vocabulary are easily understandable for Grade 7.
- **Disciplinary category:** Physics
  The question relates to sedimentation and the effect of temperature on the settling of particles—both physical processes.

✅ Classification for Grade 10 Question:

Question (Cleaned):
A pulley is used to pull a 10 kg block to the top in 2 minutes. An identical pulley is used to pull a 20 kg block to the top in 4 minutes.
What is true about the work done and the power used to pull the two blocks up to the top?

- **Math-based:** 1
  Reasoning: The question requires understanding and comparing work and power, which involves calculations using formulas (work = force × distance; power = work/time). This involves mathematical reasoning appropriate for Grade 10.
- **Language complexity (Grade 10):** 5
  Reasoning: The question uses scientific terminology (work, power, pulley) and comparative reasoning requiring multi-step thinking.
- **Disciplinary category:** Physics
  Reasoning: The concepts of work and power are fundamental physics topics related to mechanics and energy.

Source: OpenAI chat window for the classification exercise. The first and second image shows the initial set of prompts to begin this task, where we understand how the model would classify the science questions. The third and fourth images show examples of the process followed by the model for a sample question meant for grade 7 and grade 10 respectively.