Tracing the Origins of Gender Bias in Teacher Grading*

Konstantina Maragkou, *University of Cambridge*

Abstract

We analyse new administrative records to assess the role of general ability in explaining gender gaps in teacher-assigned grades across ten "university-preferred" STEM and non-STEM subject areas. The evidence comes from England, where A-level students apply to university using teacher predictions rather than exam results. We find that, conditional on exam grades, boys receive less favourable predictions from their teachers. However, this differential grading is substantially reduced when accounting for gender differences in general ability. In STEM, the gap is rather reversed, with a grade penalty identified against girls with similar general ability and achieved grades at A-level. Our findings provide evidence that teachers are not neutral to students' attributes captured in our measure of general ability, underscoring the serious implications of relying on predicted grades for university applications instead of exam results.

JEL Classifications: I21, I23, I24, J16, J71

Key Words: subjective assessments, gender gaps, general ability, STEM

*Many thanks to to the Economic and Social Research Council (ESRC) for funding this work through Administrative Data Research UK (ADR UK) Grant Number: ES/W012405/1. Special thanks to Nadir Zanini and everyone else involved in compiling the datasets analysed in this study. The dataset is accessed via the Office for National Statistics (ONS) Secure Research Service. The use of this data in this work does not imply the endorsement of the ONS in relation to the interpretation or analysis of the statistical data. This work uses research datasets which may not exactly reproduce National Statistics aggregates. The datasets are not publicly available due to their highly sensitive nature. However, they are available to accredited researchers who make a valid project application. For more information visit `https://www.ons.gov.uk/aboutus/whatwedo/statistics/requestingstatistics`

1 Introduction

Across most developed countries, data consistently show that girls outperform boys academically from early childhood through secondary school, and they continue to outpace them in university enrolment. This academic advantage is less pronounced in science, technology, engineering, and maths (STEM), with boys often having an advantage, reflecting a broader international pattern where girls, despite their strong educational performance, are underrepresented in STEM[1]. These educational gaps raise important policy concerns from both efficiency and equity perspectives (Cavaglia et al., 2020). From an efficiency standpoint, the underutilisation and misallocation of talent suggests a system that fails to utilise its full potential, potentially hindering economic growth. From an equity standpoint, the gender gap underscores inequality of opportunity, given the high financial returns to education, especially in certain STEM disciplines.

A significant body of research has examined the role of teachers in contributing to these educational gaps. Teachers have a profound influence on student success, they can inspire and motivate, but they can also negatively impact student outcomes if they hold low expectations for certain groups. Several studies have shown that teachers' subjective assessments of student abilities based on gender have significant influence on girls' and boys' educational trajectories and career choices, leading to long-term consequences if students are subjected to

[1]For example, the latest OECD report on gender highlights that while overall gender gaps in maths and science are relatively small, girls remain underrepresented in STEM-related fields once they leave school. This under-representation in STEM career paths is mirrored by broader gender disparities in the workforce, with graduate women earning only 76% of what their male counterparts make. This means that girls are less likely to translate their good school performance into fields of studies for higher education that offer better employment prospects. The report also highlights that this disparity is partly due to the lower self-confidence in girls' abilities to tackle maths and science problems, which is exacerbated by gender stereotypes that they encounter at home, in school, and across their communities, reinforcing traditional views that may discourage girls from pursuing STEM disciplines. The report is available here: `https://www.oecd-ilibrary.org/sites/34680dd5-en/index.html?itemId=/content/publication/34680dd5-en`

biased perceptions[2]. However, despite the significant implications of gender-based teaching practises, their underlying causes are still not fully understood.

In this paper we examine gender gaps in grades provided by teachers compared to results from externally marked exams across STEM and non-STEM fields of study. The aim of the study is to determine the direction and magnitude of these gaps, and explain the root causes driving them. The evidence comes from England, where students apply to university using grade predictions rather than their actual A-level exam results. The system promotes "aspirational predictions" to motivate students to aim higher and strive for better outcomes. However, any biases in these predictions can have far-reaching consequences. Predicted grades not only have direct implications on the university offers students receive, but they can also influence their perceptions, expectations of their abilities, learning goals, and the effort they put into their studies (Leckie and Maragkou, 2024).

The analysis is based on newly linked administrative data that include all university applicants, providing comprehensive details on their applications, exam results, and key socio-demographic characteristics. Our empirical strategy follows two main steps. First, we investigate whether there are systematic differences between predicted grades and exam results by student gender across fields of study. Second, we examine whether these differences can be explained by variations in boys' and girls' general ability, extending beyond subject-specific proficiency. To gain a deeper understanding of what drives these disparities, we investigate a range of factors potentially linked to general ability, as well as the predicted-achieved grade gap, including individual student characteristics and aspects of the application process.

[2]Bias or discrimination in grading is commonly assessed by identifying systematic differences between objective (blind) and subjective (non-blind) assessments across different groups. This approach was pioneered in economics by Blank (1991) and was first applied to the economics of education by Lavy (2008), followed by applications from various others e.g. Bonesrønning (2008); Cornwell et al. (2013); Burgess and Greaves (2013); Botelho et al. (2015). Numerous studies have linked teacher bias to having negative effects on students' academic progress, exam performance, university admission prospects, and their choice of university studies e.g. Falch and Naper (2013); Breda and Ly (2015); Lavy and Sand (2018); Lavy and Megalokonomou (2019); Terrier (2020).

We find substantial gender gaps in predicted grades, conditional on achieved grades. Consistent with previous research, these gaps favour girls and are evident across all levels of the achieved grade distribution. Similar to Lavy and Megalokonomou (2024), we observe more pronounced gender differences in non-STEM subjects, with less pronounced gaps in STEM. The results remain consistent across alternative specifications and robust against a range of potential issues, including measurement error in exam scores, statistical discrimination, and sample selection biases.

These findings lead to an important question: why do girls receive more favourable predictions compared to boys with the same subject-specific academic achievements? Teacher stereotyping is one possible explanation. However, while teacher predictions should ideally be objective and independent of external factors Burgess and Greaves (2013), variations in student attributes and behaviours could significantly impact teachers' ability to predict student outcomes. Understanding grading biases in relation to these factors is, therefore, crucial not only because of its impact on the accurate assessment of student ability, but also because it could explain grading biases related to other factors (Ferman and Fontes, 2022).

The second part of the paper examines if differences between boys' and girls' general ability can explain the predicted-achieved grade gaps by gender. We measure general ability through points from exams taken at age 16, two years before A-levels, with the best eight subjects capped. This measure is interdisciplinary as it considers performance across various domains, therefore encompassing a diverse range of cognitive and, arguably, non-cognitive skills[3]. We

[3] A recent nationally representative study by Smith et al. (2021) established a strong negative correlation between hyperactivity disorder and exam performance at age 16, with boys exhibiting significantly higher levels of hyperactivity compared to girls. Mendolia and Walker (2014) demonstrated that higher performance at age 16 exams correlates with an internal locus of control and higher self-esteem. Similarly, Gutman and Vorhaus (2012) show that better performance at age 16 exams is linked to lower levels of behavioural issues with hyperactivity and attention deficit emerging as the strongest predictors of low exam performance at age 16. Rothon et al. (2009) examined the impact of depressive symptoms and also found a negative association between depressive symptoms and exam performance at age 16 for boys, but not for girls.

demonstrate substantial gender gaps in general ability, with girls consistently scoring higher than boys, particularly girls studying STEM at A-level. These gaps persist when accounting for subject-specific proficiency, indicating a broader pattern of skill gaps based on gender. After adjusting for this skill differential, the gender gap in predicted grades against boys is substantially reduced in non-STEM. In STEM, the gap is rather reversed, in favour of boys.

Several studies on teacher gender bias employed individual fixed-effects and exploited within-student across-subject variation to mitigate concerns relating to unobserved student traits and behaviour[4]. However, this strategy does not allow to understand why differential predictions occur. Few studies attempted to proxy student behaviour, but they often lacked reliable measures[5]. Three papers had access to comprehensive measures of classroom behaviour. Ferman and Fontes (2022) show that classroom behaviour explains approximately two-thirds of the discrimination against boys. Similarly, Black and de New (2020) find that after adjusting for gender differences in hyperactivity, the teacher-test score gap in maths favours boys. But, these papers are not focused on gender and discuss these results very briefly. Cornwell et al. (2013) is the only paper to our knowledge that focuses on gender and considers the role of student behaviour using detailed measures. The paper demonstrates that teachers tend to hold more positive perceptions about the abilities of well-behaved students, and that the preferential treatment of girls diminishes when classroom behaviour is considered. However, Cornwell et al. (2013) do not establish a direct link between these variations in teacher perceptions and differences in grading.

The paper makes three important contributions to the literature on gender gaps in academic achievement and teacher grading biases. First, it connects the gap between predicted and actual grades, based on student gender, to differences in general ability between boys and

[4]See e.g. Burgess and Greaves (2013); Black and de New (2020); Breda and Ly (2015); Carlana (2019); Terrier (2020).

[5]For example, Terrier (2020) found that accounting for a student's disciplinary warnings obtained from the class council or school exclusions, did not impact the identified gender gap.

girls. The direct pathway through which our measure of general ability captures differential student capability is not straightforward to interpret. However, our comprehensive dataset allows us to examine several possible explanations. For instance, Murphy and Wyness (2020) suggest that varying rates of growth in achievement from the time of prediction to the time of the exam, or differences in motivation and effort in response to predicted grades, could be driving the observed prediction gap. Similarly, Leckie and Maragkou (2024) propose that students who are more ambitious, such as those applying to selective courses or universities, may exert more pressure on teachers for higher predictions. However, our study shows that gender-based gaps in teacher predictions cannot be explained by differences in ambition, differential responses to predicted grades, or varying rates of academic progress.

Girls' advantage in overall competence as a plausible explanation for differential predictions has not been discussed before. This consideration is important because predicted grades might be influenced not only by girls' and boys' subject-specific strengths but also by their broader capacity and aptitude. Breda and Napp (2019) and Goulas et al. (2022) demonstrate that girls who excel in maths are more likely than boys to achieve even better results in reading, and this relative advantage impacts girls' educational choices, contributing to their under-representation in maths-intensive fields. We also find substantial gender gaps in cross-subject skills: girls studying STEM at A-level are equally proficient in maths as boys, but they significantly outperform them in English. We show that this gap between subject-specific proficiency and overall competence is important in explaining gender-based differences in predicted grades.

Our second contribution is to provide evidence on predicted-achieved grade gaps in a high-stakes context, where both grades carry significant impact for student outcomes. Previous research, including Cornwell et al. (2013), has typically focused on low-stakes assessments, relying on internal measures such as class-based tests. By focusing on predicted grades the analysis benefits from a degree of outcome comparability that may have been lacking in previous studies in cases where teacher assessments measured different skills compared to external exams. In addition, and crucially, focusing on teacher predictions made for university applications has broader implications for national debates surrounding student assessment and university admissions criteria. Despite official guidelines indicating that predicted grades should not be influenced by student behaviour or background, our findings suggest that teachers' predictions are subtly influenced by such factors.

Finally, our third contribution is to extend the existing literature by broadening the analysis to encompass a wider array of "university-preferred" subject areas. These subjects, commonly classified as facilitating, have been previously associated with enrolment in more selective university programs (Dilnot, 2018). Importantly, assessing gender gaps across fields of study allows for a more comprehensive analysis, capturing the diversity of academic disciplines and their respective gender dynamics. Breda and Ly (2015) show that teacher evaluations exhibit bias favouring girls in more male-dominated subjects and boys in more female-dominated subjects. Our analysis also identifies the gender ratio as a potentially important factor. We observe that the conditional gap in predicted grades is more prominent in subjects predominantly dominated by girls, both in STEM and non-STEM. Evidently, the pattern of the conditional predicted-achieved grade gap across subjects aligns well with a gender-stereotyped model: girls tend to receive more favourable predictions in non-STEM, and boys tend to receive more favourable predictions in STEM. Understanding whether these residual gaps stem from gender stereotyped bias among teachers is crucial, but exploring this in depth exceeds the scope of this paper. We will return to answering this question using appropriate identification in our future research, as well as examining it's potential implications on university application outcomes.

The rest of this paper unfolds as follows. In the next section, we set the context by explaining the English schooling and higher education application systems. We also discuss the specifics of our dataset, outline our analytical sample, and present key descriptive statistics. Section

3 details our empirical framework. Section 4 presents our results and robustness checks. Finally, in Section 6, we draw our conclusions.

2 Setting and data

2.1 Institutional background

In England, the compulsory school curriculum is structured into five Key Stages. Progress is measured through standardised assessments at the end of Key Stage 2 (grade 6), when students are 11 years old, and at the end of Key Stage 4 (grade 11), at age 16. Key Stage 2 exams assess students in maths, science, and English, while Key Stage 4 exams typically cover 8 to 10 subjects and result in the award of a General Certificate of Secondary Education (GCSE) for each subject. The fifth and final Key Stage (grades 12-13) offers more diverse options. Students at this stage can choose to pursue vocational courses, typically in a college setting, or academic qualifications, commonly Advanced-level qualifications (A-levels), at sixth form schools or colleges. A-levels represent the predominant pathway to university and typically involve studying three subjects. The combination of subjects is entirely optional but for university applicants there is some general direction via a list of eight "university preferred" subjects classified as "facilitating". These include biology, chemistry, English (language and literature), geography, history, maths, modern and classical languages, and physics. Upon completing the final Key Stage at age 18, students can progress to higher education by enrolling in university courses.

The higher education admissions process is overseen by the University and College Admissions Service (UCAS), a centralised body which enables students to apply for up to five university courses in each application round. This process requires students to select courses almost a year in advance of entry and prior to receiving their exam results. The most selective universities typically require three A-levels at grades AAB, however there is considerable variability in course-specific requirements across institutions. Crucially for our study's context, students apply to university based on predicted grades provided by their school teachers for each A-level subject they study. Teachers must submit their grade predictions to UCAS by January of each academic year, approximately eight months before A-level exams take place[6]. They receive some general guidance on predicting student grades, but schools may also impose their own regulations. Predicted grades are described by UCAS as the grades a student's school or college believes they are likely to achieve in favourable circumstances. These predictions are expected to be: (i) in the best interest of applicants, (ii) aspirational yet achievable, (iii) determined by professional judgement, and (iv) data-driven.

University admissions tutors evaluate applications based on these predictions, and an accompanying personal statement. Following this selection process, universities offer places to students conditional on achieving specific grades. Once offers are received, students select a first-choice and backup course from their available options. Only after this decision-making process is complete do students sit their A-level exams. Upon receiving their grades, they can then enroll in their first-choice course or, if they fail to achieve the required grades, their backup choice. If neither option is viable, students can enter a process known as "clearing," where they can apply to courses that still have available places.

2.2 Dataset and analytical sample

We use a newly linked administrative dataset facilitated by the "Grading and Admissions Data for England" initiative. School-related information is sourced from the Department for Education's National Pupil Database, providing comprehensive records on student prior attainment, alongside basic socio-demographic characteristics. The data on A-level exam results are provided by the Office of Qualifications and Examinations Regulation (Ofqual),

[6]Certain institutions and courses require earlier submission of predicted grades, with deadlines in October. This earlier timeline is typical for the two highly selective universities, Oxford and Cambridge, along with specific courses in medicine, dentistry, and veterinary science. In our robustness checks, we tested whether excluding applicants with earlier predicted grade submissions would impact our results, and found that it had no significant qualitative effect.

while information on predicted grades and university applications is sourced from UCAS. The dataset pertains to the cohorts of students in England who applied to universities during the years 2017 to 2019 with one or more A-level results. The dataset excludes mature or international students, as well as those who applied via qualifications other than A-levels[7].

We conduct our investigation at the subject-level, focusing on the seven "facilitating" subjects mentioned above (excluding modern and classical languages)[8], along with three additional subjects listed as "usefull" and that are studied by many students (Dilnot, 2018). We classify five subjects as STEM-related: biology, chemistry, economics, maths and physics; and five as non-STEM: English literature, geography, history, psychology and sociology.

The data cover over a million student-subject observations (n=1,018,833). We exclude students with missing (or invalid) predicted or achieved grades (6%), as well as those for whom we lack records of prior attainment at ages 11 and 16 (24%)[9], and other relevant application and student characteristics (2%). Our total analytical sample is composed of 722,353 student-subject observations[10].

The majority of university applicants in our data are girls (57%)[11]. Table 1 shows the

[7]The dataset does not contain any information on individual teachers, thereby limiting our ability to explore the role of teachers' gender in the subjective assessment of students' ability as in previous studies like Dee (2005) and Breda and Ly (2015).

[8]Modern and classical languages are not considered because they cover a wide range of subjects that are generally not studied by many students.

[9]The reduction in our sample size resulting from the exclusion of students with missing prior attainment is primarily because these students have typically attended private schools, where public records of their exam results are unavailable. As part of our robustness checks, we run our baseline specification (which does not include controls for prior attainment) on the full sample with non-missing predicted and achieved grades to assess potential variations in our results, yet we do not find any significant differences.

[10]The total number of applicants across the three UCAS application rounds is 698,701 individuals. Among them, 70% (n=487,456) study at least one of the subjects analysed in our study and have non-missing predicted and achieved grades. We have non-missing measures of prior attainment (and other characteristics) for 74%, totalling 361,041 unique university applicants and 722,353 student-subject observations.

[11]This ratio is consistent with the ratio in the full sample of university applicants in the 2017-2019 application rounds.

share of boys and girls in each A-level subject (bottom of panel A and B respectively), suggesting notable differences in subject-taking by gender. Non-STEM subjects such as English literature, psychology, and sociology tend to be female-dominated, while subjects like geography and history have a more balanced ratio of boys and girls. Additionally, the maths-intensive STEM subjects, including economics, maths, and physics, tend to be dominated by boys. Within the STEM subjects, girls are more prevalent in biology and, to a somewhat lessen extend, chemistry. These trends highlight the gender trends previously discussed, pertaining to the the higher university enrolment of girls and the significant disparities in fields of study choices.

2.3 Predicted and achieved grades

We analyse predicted and achieved grades in the school-leaving examinations known as A-levels. A-levels are typically taken in three subjects, and constitute the dominant, but not exclusive, pathway to university. Predicted and achieved grades are scored between A* to E. To simplify the analysis, we convert grades to scores, assigning the following scores per grade: A*=6, A=5, B=4, C=3, D=2, E=1. The first part of each respective panel in Table 1 presents the mean values and standard deviations for predicted (PG) and achieved (AG) grades, and the difference between the two (PG - AG), by A-level subject. Panels A and B provide the descriptive statistics for boys and girls respectively, while Panel C presents the t-test results for mean gender differences.

Predicted grades are, on average, one grade higher compared to achieved grades. Focusing on the third panel, the gender gaps in raw achievement are generally small within STEM subjects. However, in non-STEM, girls consistently outperform boys by approximately one-fifth of a grade in both predicted and achieved grades. The difference between predicted and achieved grades (PG-AG) between boys and girls is more pronounced in STEM, ranging between 0.05 and 0.20 grades in favour of girls. In non-STEM the evidence is more mixed and of smaller magnitude, with some instances of predicted grades favouring boys[12].

The first part of each panel in Table 1 provide also the descriptive statistics for accurate prediction (PG = AG), under-prediction (PG < AG), and over-prediction (PG > AG). Across all subjects, only approximately one-third of the students receive accurate predictions, while the majority (around 60%) are over-predicted. Under-prediction is observed in less than 10% of the sample, highlighting again the optimistic nature of predicted grades.

Figure 1 illustrates the optimism in predicted grades through kernel-density estimates, contrasting the predicted (dashed line) and achieved (solid line) grades for boys (grey line) and girls (black line). Across all subjects, there is a notable difference between the distributions of predicted and achieved grades. Achieved grades exhibit a mostly normal distribution, while predicted grades consistently shift towards higher grades, indicating a pervasive optimism. Notably, in STEM, the distributions of both predicted and achieved grades are nearly identical across genders, as indicated by the significant overlap between the grey and black lines. However, in some non-STEM subjects, the distribution slightly skews towards higher scores for girls. Interestingly, this directional shift is consistent between predicted and achieved grades.

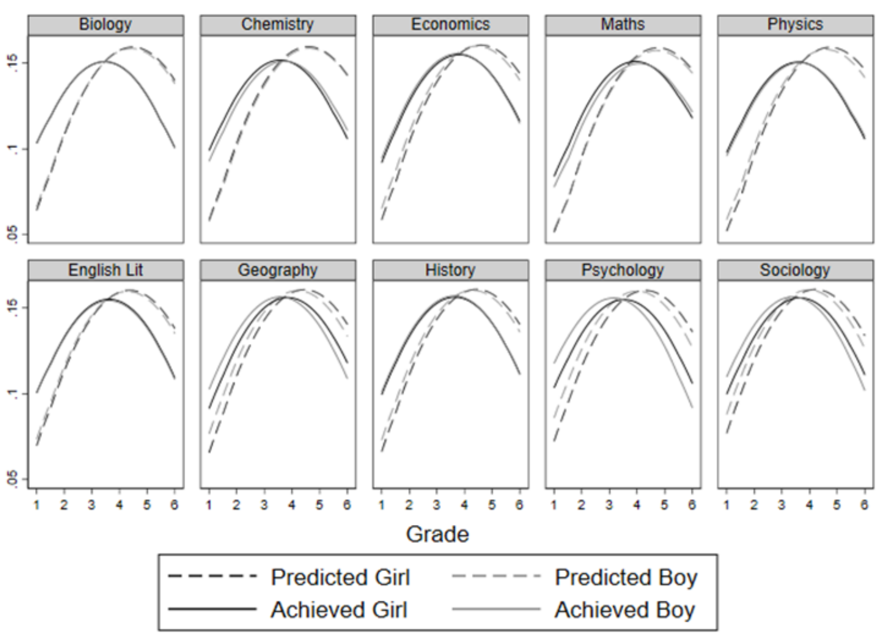
Figure 2 plots the students' predicted grades against their achieved grades. The 45° line represents the expected distribution of grades if every student's prediction matched their actual performance. Points above this line indicate over-prediction, while those below suggest under-prediction. The figure clearly shows that most students are over-predicted. Additionally, the plot shows that the primary factor driving this over-prediction is the existence of

[12]The difference between predicted and achieved grades (PG-AG) between boys and girls presented in Table 1 is equivalent to what would have been obtained through the difference-in-differences specification introduced in Lavy (2008). The estimates derived from this specification differ from those reported in our empirical analysis below. This difference arises from variations in the distribution of achieved grades by gender, which our specification accounts for by isolating variation in predicted grades for students who ultimately achieve the same grade. For a more detailed discussion of the differences between the two approaches, refer to Delaney and Devereux (2023).

ceiling and floor effects within the grading system. Students who ultimately achieve lower grades, show a higher tendency to be over-predicted. For example, those who achieved the lowest possible grade (E) are over-predicted by an average of three grades. On the other hand, students who achieve at the highest level (grade A*) are generally under-predicted, with an average difference of about half a grade. This trend is especially notable in non-STEM subjects.

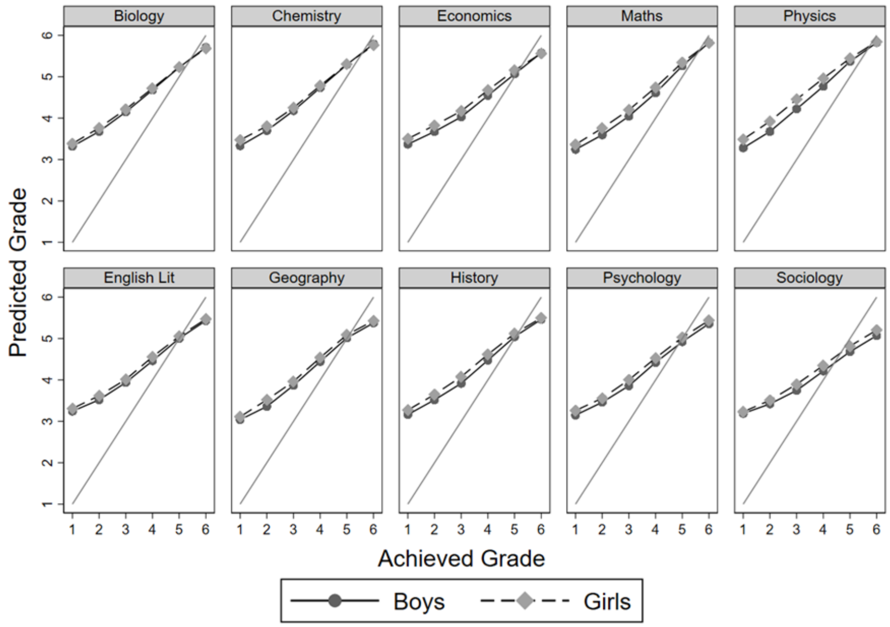
Despite these inherent grading effects, our analysis concentrates on examining gender disparities in predicted grades among students who ultimately achieve similar academic outcomes. By focusing on students who attain the same final grade, we can effectively mitigate the influence of these mechanical grading effects, ensuring that our analysis and findings remain centred on the core question of gender-based differences in the predicted-achieved grade gap, and the potential role of general ability in explaining these gaps.

Figure 1: Predicted and achieved grades distribution by gender



Notes: The histograms display the distribution of the predicted and achieved grades for boys and girls across A-level subjects. The following points per grade are used in the calculation: A*=6,A=5, B=4, C=3, D=2, E=1. The sample comprises university applicants between 2017-2019. *Source: GRADE 2017-2019*

Figure 2: Accuracy of predicted grades by gender and subject



Notes: The figure displays the mean predicted grade at each point of the achieved grade distribution for boys and girls across A-level subjects. The following points per grade are used in the calculation: A*=6,A=5, B=4, C=3, D=2, E=1. The 45° line represents the expected distribution of grades if every student's prediction matched their actual performance. The sample comprises university applicants between 2017-2019. *Source: GRADE 2017-2019*

2.4 General ability

We measure general ability based on points collected from the grade 11 exams capped at each student's best eight subjects and standardised by cohort. Students have the flexibility to select the number and types of subjects they study, in addition to some compulsory ones[13]. Generally, they choose the optional subjects that align with their interests and strengths, aiming to maximise their grades. This score, therefore, reflects broader academic capabilities, even though it allows for some degree of specialisation. Further, the capped score allows for a more robust level of comparability than the total or mean score, as it adjusts for differences in the number of exams (and resits) taken by students (Adamecz

[13]The only compulsory subjects are maths, English, and Science (single, double, or triple).

et al., 2024)[14]. Additionally, the standardisation process places each student's score within the context of the broader applicant pool (and not only those included in our analytical sample), offering a more accurate measure of their standing[15].

The descriptive statistics for our measure of general ability are presented in the second part of each panel in Table 1 for each A-level subject. Focusing on the third panel, showing mean differences between boys and girls, the gender gaps in the grade 11 point score are stark. Girls consistently score higher than boys, with differences ranging from 19% to 31% of a standard deviation in STEM, and from 10% to 22% of a standard deviation in non-STEM. The consistent trend of girls outperforming boys suggests a broader advantage in academic skills and aptitudes among female students.

Figure 3 illustrates the association between gender and general ability across A-level subjects, conditional on subject-specific skills as measured by exam performance at ages 18, 16 and 11[16]. The left panel conditions only on the subject-specific skills, while the right panel adds controls for other socio-demographic characteristics and school-by-cohort fixed effects. In STEM, the conditional gender gaps remain comparable to the raw gaps documented in Table 1, ranging between 20% and 30% of a standard deviation. In contrast, the gaps in non-STEM are notably reduced, ranging between 2% to 7% of a standard deviation. The evidence in Figure 3 demonstrates that girls outperform boys in their overall competence beyond subject-specific performance, especially girls who study STEM at A-level.

Collectively, this descriptive evidence suggests that our measure of general ability extends beyond subject-specific skills, potentially serving as a proxy for more broadly defined mea-

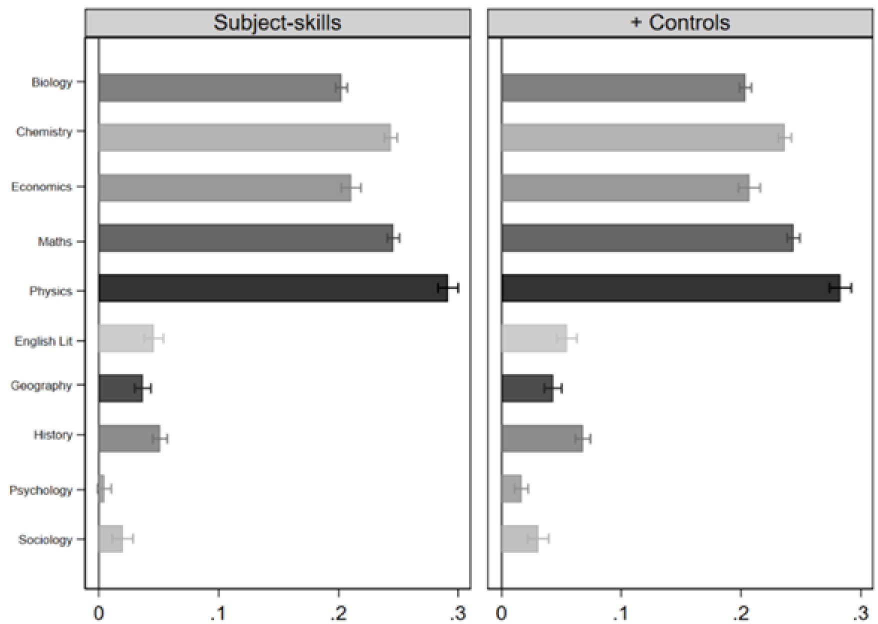
[14]While we use the capped score for our main analysis our results remain qualitatively similar when we consider the mean or total score.

[15]The GCSE qualifications underwent reform in 2018, resulting in a change in the measurement scale. The standardisation process is also providing a level of comparability across different cohorts.

[16]Exam performance at age 18 is based on the achieved A-level grade in the corresponding subject. At ages 16 and 11, we consider exam scores from grade 11 and grade 16, respectively, in maths for STEM subjects and English for non-STEM.

sures of student traits. While our data lack information on noncognitive skills to directly connect this measure to student behaviour, previous studies established a strong association with traits like hyperactivity, self-esteem and locus of control (Smith et al., 2021; Mendolia and Walker, 2014).

Figure 3: Conditional gender gaps in general ability



Notes: The figure displays coefficient estimates for gender from a regression of the capped grade 11 point score, conditional on gender, subject-specific skills at ages 11 and 16 (standardised), and achieved A-level grades. The right panel of the figure includes additional controls for parental occupation, ethnicity, and school fixed effects. The dependent variable is our standardised measure of general ability, based on the capped GCSE point score. The positive coefficient values indicate the advantage for girls in general ability, over and above subject-specific skills and the additional controls. The sample comprises university applicants between 2017-2019. *Source: GRADE 2017-2019*

2.5 Student and application characteristics

The second part of each respective panel in Table 1 provides also the descriptive statistics for grade 11 and grade 6 standardised exam scores in maths and English[17], as well as other student and application characteristics. All measures of prior attainment are standardised by cohort.

Boys outperform girls in their prior performance in maths across most A-level subjects, but not in the maths-intensive ones. The gender differences in maths attainment are, however, relatively small. The biggest gap in grade 6 maths performance is observed in English literature where boys outperform girls by about 21% of a standard deviation. By grade 11, the gender differences become even smaller, ranging from approximately 2% of a standard deviation in biology and geography to about 9% in English literature.

Conversely, girls outperform boys in English prior attainment by a considerable margin, ranging from one-third to one-half of a standard deviation, across both grades. The gender gaps in the lagged teacher assessment from grade 6[18] follows a similar pattern, with higher average scores for boys in maths but almost twice as high scores for girls in English. These observations highlight a relative advantage for girls in English compared to maths, a mechanism often linked to their under-representation in maths-intensive fields (Breda and Napp, 2019).

Focusing on application-specific characteristics next, we consider two indicators of selective applications: whether students have applied to a high-tariff institution[19] and whether they

[17]We consider the total marks achieved in KS2 exams (grade 6) and the grade score achieved in full GCSE for maths and English.

[18]The grade 6 teacher assessment is based on the National Curriculum level assigned for maths and English by the primary school teacher (at the end of Key Stage 2). For comparability purposes we standardised it by cohort.

[19]Among the 407 institutions in our dataset, 38 are classified as high-tariff by UCAS. These include: Bath, Glasgow School of Art, Edinburgh, St Andrews, Newcastle, Cambridge, Cardiff, Southampton, Hull York Medical School, Warwick, Oxford, Durham, Royal Veterinary College, University College London, Aberdeen, London School of Economics and Political Science, Bristol, Glasgow, Courtauld Institute of Art, King's College London, Impe-

have applied to a STEM course[20], considering the higher entry requirements of these programs. As expected, students studying STEM at A-level are more likely to apply to more selective institutions and to STEM degree courses[21]. Overall, gender differences in application characteristics are relatively small. Finally, Table 1 shows that boys tend to marginally report higher levels of learning difficulties, while girls report higher levels of other disabilities. However, these gender differences are again minimal.

In Appendix Table A.1, we also provide the descriptive statistics for parental occupation and ethnicity by gender across A-level subjects. As expected, there are no notable patterns across these characteristics by student gender. However, it is worth noting that the majority of students in our sample have parents in professional and managerial occupations, ranging from about 40% in Sociology to over 60% in Physics. Additionally, there are some notable patterns in A-level subject taking by ethnic background. The two humanities subjects are predominantly taken by white students (approximately 85%), compared to about 75% in other non-STEM subjects and around 65% in STEM subjects.

rial College London, Birmingham, Manchester, Heriot-Watt, Sheffield, St George's University of London, University of London Institute in Paris, Exeter, Lancaster, York, Starthclyde, Liverpool, Dundee, Leeds, Nottingham, University of London, Susex, Brighton and Sussex Medical School.

[20]We classify courses as STEM based on HESA JACS 3.0: Principal subject codes. JACS 1-9 are classified as STEM, and JACS A-J are classified as non-STEM. The subject codes are available here: `https://www.hesa.ac.uk/support/documentation/jacs/jacs3-principal`

[21]The only exception is economics which is associated with a low probability of applying to a STEM course (however a high probability of applying to a selective course). Most likely this is because the economics course falls under the (non-STEM) social studies category according to HESA JACS 3.0 principal subject codes for degree courses. Consequently, many students studying economics at A-level that have applied for an economics or economics-related course appear in the non-STEM courses.

Table 1: Descriptive statistics of achievement, student characteristics and application characteristics

	(1) Biology	(2) Chemistry	(3) Econ.	(4) Maths	(5) Physics	(6) Eng. Lit.	(7) Geogr.	(8) History	(9) Psychol.	(10) Sociol.
Panel A: Boys										
Part 1: Predicted and achieved grades										
PG	4.428	4.608	4.453	4.730	4.594	4.286	4.211	4.294	4.008	3.973
	(1.044)	(1.036)	(0.962)	(1.109)	(1.079)	(0.979)	(0.950)	(0.937)	(0.970)	(0.925)
AG	3.465	3.725	3.754	4.086	3.643	3.616	3.577	3.627	3.162	3.400
	(1.412)	(1.406)	(1.203)	(1.461)	(1.441)	(1.265)	(1.160)	(1.128)	(1.191)	(1.158)
PG-AG	0.963	0.884	0.699	0.644	0.951	0.671	0.634	0.667	0.846	0.572
	(1.051)	(1.008)	(1.008)	(1.010)	(1.012)	(1.017)	(0.915)	(0.934)	(1.022)	(1.056)
PG=AG	0.285	0.322	0.337	0.405	0.299	0.330	0.346	0.342	0.282	0.333
	(0.452)	(0.467)	(0.473)	(0.491)	(0.458)	(0.470)	(0.476)	(0.474)	(0.450)	(0.471)
PG>AG	0.654	0.624	0.566	0.510	0.653	0.560	0.562	0.568	0.635	0.525
	(0.476)	(0.484)	(0.496)	(0.500)	(0.476)	(0.496)	(0.496)	(0.495)	(0.481)	(0.499)
PG<AG	0.061	0.054	0.098	0.086	0.048	0.111	0.093	0.090	0.083	0.142
	(0.239)	(0.226)	(0.297)	(0.280)	(0.214)	(0.314)	(0.290)	(0.286)	(0.275)	(0.349)
Part 2: Prior achievement, and student and application characteristics										
G11 point score	0.630	0.779	0.450	0.738	0.764	0.310	0.349	0.322	0.143	-0.171
	(0.600)	(0.590)	(0.626)	(0.585)	(0.579)	(0.682)	(0.611)	(0.659)	(0.581)	(0.567)
G11 Maths	0.642	0.939	0.567	1.082	1.079	-0.050	0.222	0.088	0.027	-0.394
	(0.732)	(0.654)	(0.739)	(0.528)	(0.572)	(0.856)	(0.792)	(0.851)	(0.730)	(0.724)
G11 English	0.231	0.296	0.138	0.248	0.244	0.443	0.062	0.205	-0.034	-0.242
	(0.878)	(0.908)	(0.855)	(0.907)	(0.910)	(0.903)	(0.860)	(0.903)	(0.817)	(0.806)
G6 Maths	0.478	0.630	0.512	0.743	0.751	0.115	0.298	0.200	0.171	-0.188
	(0.699)	(0.642)	(0.678)	(0.546)	(0.549)	(0.885)	(0.776)	(0.844)	(0.803)	(0.923)
G6 English	0.175	0.222	0.125	0.238	0.257	0.301	0.074	0.197	-0.000	-0.249
	(0.861)	(0.876)	(0.866)	(0.873)	(0.859)	(0.849)	(0.857)	(0.861)	(0.862)	(0.909)
G6 Maths TA	0.436	0.575	0.474	0.672	0.681	0.125	0.279	0.206	0.156	-0.153
	(0.804)	(0.778)	(0.797)	(0.714)	(0.728)	(0.923)	(0.853)	(0.900)	(0.866)	(0.934)
G6 English TA	0.175	0.216	0.142	0.243	0.252	0.240	0.102	0.166	0.018	-0.206

	(0.891)	(0.883)	(0.898)	(0.881)	(0.874)	(0.869)	(0.904)	(0.895)	(0.919)	(0.955)
Applied high tariff	0.817	0.892	0.780	0.866	0.873	0.706	0.684	0.724	0.614	0.528
	(0.387)	(0.311)	(0.414)	(0.341)	(0.333)	(0.455)	(0.465)	(0.447)	(0.487)	(0.499)
Applied STEM	0.840	0.900	0.245	0.745	0.884	0.177	0.499	0.227	0.550	0.275
	(0.366)	(0.299)	(0.430)	(0.436)	(0.321)	(0.382)	(0.500)	(0.419)	(0.498)	(0.447)
Learning difficulty	0.024	0.023	0.020	0.024	0.030	0.020	0.029	0.028	0.023	0.029
	(0.155)	(0.151)	(0.138)	(0.153)	(0.171)	(0.140)	(0.169)	(0.164)	(0.150)	(0.169)
Other disability	0.039	0.040	0.028	0.040	0.047	0.059	0.035	0.049	0.047	0.049
	(0.194)	(0.196)	(0.166)	(0.196)	(0.211)	(0.235)	(0.184)	(0.217)	(0.211)	(0.215)
N	34235	37211	27272	65356	41551	13498	23586	31143	22261	11280
% in subject	36	46	69	60	79	21	46	41	23	20
Panel B: Girls										
Part 1: Predicted and achieved grades										
PG	4.473	4.587	4.594	4.752	4.739	4.366	4.444	4.436	4.300	4.206
	(0.998)	(0.981)	(0.928)	(1.022)	(1.006)	(0.940)	(0.917)	(0.906)	(0.946)	(0.910)
AG	3.459	3.573	3.803	3.939	3.591	3.622	3.838	3.652	3.532	3.647
	(1.420)	(1.379)	(1.234)	(1.413)	(1.427)	(1.229)	(1.180)	(1.166)	(1.239)	(1.188)
PG-AG	1.014	1.013	0.791	0.813	1.148	0.745	0.606	0.784	0.768	0.559
	(1.049)	(1.021)	(1.018)	(0.990)	(1.029)	(0.988)	(0.905)	(0.935)	(0.995)	(1.021)
PG=AG	0.275	0.281	0.312	0.350	0.249	0.312	0.361	0.312	0.312	0.341
	(0.447)	(0.450)	(0.463)	(0.477)	(0.433)	(0.463)	(0.480)	(0.463)	(0.463)	(0.474)
PG>AG	0.671	0.677	0.603	0.594	0.721	0.598	0.545	0.618	0.601	0.520
	(0.470)	(0.467)	(0.489)	(0.491)	(0.448)	(0.490)	(0.498)	(0.486)	(0.490)	(0.500)
PG<AG	0.054	0.041	0.085	0.057	0.030	0.091	0.094	0.070	0.087	0.139
	(0.225)	(0.199)	(0.279)	(0.232)	(0.169)	(0.287)	(0.292)	(0.255)	(0.281)	(0.346)
Part 2: Prior achievement, and student and application characteristics										
G11 point score	0.818	0.968	0.667	0.981	1.075	0.414	0.568	0.515	0.329	-0.010
	(0.554)	(0.532)	(0.597)	(0.534)	(0.504)	(0.674)	(0.597)	(0.638)	(0.596)	(0.600)
G11 Maths	0.618	0.872	0.584	1.128	1.156	-0.136	0.196	0.035	-0.009	-0.453
	(0.724)	(0.656)	(0.740)	(0.500)	(0.537)	(0.862)	(0.801)	(0.849)	(0.762)	(0.750)
G11 English	0.587	0.681	0.459	0.670	0.767	0.560	0.408	0.529	0.258	-0.001
	(0.835)	(0.845)	(0.849)	(0.857)	(0.836)	(0.874)	(0.838)	(0.866)	(0.822)	(0.810)

G6 Maths	0.370	0.505	0.405	0.697	0.764	-0.096	0.146	0.029	-0.000	-0.373
	(0.726)	(0.683)	(0.717)	(0.553)	(0.527)	(0.922)	(0.804)	(0.881)	(0.839)	(0.941)
G6 English	0.468	0.517	0.363	0.545	0.642	0.416	0.344	0.447	0.232	-0.021
	(0.798)	(0.805)	(0.845)	(0.802)	(0.761)	(0.843)	(0.810)	(0.818)	(0.838)	(0.887)
G6 Maths TA	0.325	0.449	0.372	0.616	0.665	-0.068	0.130	0.042	-0.002	-0.324
	(0.812)	(0.783)	(0.810)	(0.696)	(0.685)	(0.926)	(0.865)	(0.897)	(0.880)	(0.912)
G6 English TA	0.396	0.430	0.330	0.474	0.540	0.328	0.319	0.355	0.205	-0.024
	(0.790)	(0.785)	(0.837)	(0.759)	(0.712)	(0.833)	(0.818)	(0.815)	(0.868)	(0.924)
Applied high tariff	0.824	0.893	0.807	0.873	0.917	0.697	0.726	0.756	0.641	0.557
	(0.381)	(0.309)	(0.395)	(0.333)	(0.276)	(0.459)	(0.446)	(0.429)	(0.480)	(0.497)
Applied STEM	0.852	0.912	0.230	0.710	0.877	0.226	0.515	0.225	0.584	0.335
	(0.355)	(0.283)	(0.421)	(0.454)	(0.328)	(0.418)	(0.500)	(0.418)	(0.493)	(0.472)
Learning difficulty	0.021	0.018	0.022	0.022	0.031	0.018	0.031	0.021	0.021	0.023
	(0.144)	(0.134)	(0.146)	(0.146)	(0.173)	(0.133)	(0.173)	(0.145)	(0.145)	(0.151)
Other disability	0.052	0.048	0.036	0.050	0.063	0.069	0.047	0.063	0.066	0.063
	(0.222)	(0.213)	(0.186)	(0.219)	(0.243)	(0.254)	(0.213)	(0.243)	(0.248)	(0.243)
N	61821	43247	12232	43118	11171	50597	27941	44078	76415	44340
% in subject	64	54	31	40	21	79	54	59	77	80
Panel C: Difference (Boys - Girls)										
Part 1: Predicted and achieved grades										
PG	-0.044***	0.022**	-0.140***	-0.023***	-0.145***	-0.080***	-0.233***	-0.142***	-0.292***	-0.233***
AG	0.006	0.151***	-0.048***	0.146***	0.052***	-0.006	-0.261***	-0.024**	-0.370***	-0.247***
PG-AG	-0.050***	-0.130***	-0.092***	-0.169***	-0.197***	-0.074***	0.028***	-0.118***	0.079***	0.014
PG=AG	0.010**	0.041***	0.025***	0.055***	0.049***	0.018***	-0.015***	0.030***	-0.030***	-0.008
PG>AG	-0.017***	-0.053***	-0.038***	-0.084***	-0.068***	-0.038***	0.016***	-0.050***	0.034***	0.005
PG<AG	0.007***	0.013***	0.013***	0.029***	0.019***	0.020***	-0.001	0.020***	-0.004	0.003
Part 2: Prior achievement, and student and application characteristics										
G11 point score	-0.188***	-0.189***	-0.217***	-0.243***	-0.311***	-0.104***	-0.218***	-0.193***	-0.186***	-0.161***
G11 Maths	0.024***	0.067***	-0.017*	-0.046***	-0.078***	0.087***	0.026***	0.052***	0.036***	0.059***
G11 English	-0.356***	-0.385***	-0.321***	-0.422***	-0.523***	-0.117***	-0.345***	-0.324***	-0.292***	-0.242***
G6 Maths	0.108***	0.125***	0.106***	0.046***	-0.014*	0.211***	0.152***	0.170***	0.172***	0.185***
G6 English	-0.293***	-0.295***	-0.238***	-0.307***	-0.385***	-0.115***	-0.269***	-0.250**	-0.232***	-0.227***

G6 Maths TA	0.110***	0.126***	0.102***	0.056***	0.016*	0.194***	0.150***	0.164**	0.158***	0.171***
G6 English TA	-0.222***	-0.214***	-0.188***	-0.232***	-0.288***	-0.087***	-0.217***	-0.189***	-0.188***	-0.182***
Applied high tariff	-0.008**	-0.001	-0.027***	-0.007**	-0.044***	0.009*	-0.042***	-0.033***	-0.027***	-0.029***
Applied STEM	-0.011***	-0.012***	0.015***	0.036***	0.006	-0.049***	-0.015***	0.002	-0.035***	-0.060***
Learning difficulty	0.003***	0.005***	-0.002	0.002*	-0.001	0.002	-0.001	0.006***	0.002	0.006***
Other disability	-0.013***	-0.008***	-0.008***	-0.010***	-0.017***	-0.011***	-0.012***	-0.014***	-0.019***	-0.014***
N	96056	80458	39504	108474	52722	64095	51527	75221	98676	55620

Notes: Columns 1–10 report mean values and standard deviations (in parentheses) for each A-level subject, respectively. Panel A provides descriptive statistics for boys, Panel B for girls, and Panel C presents the results of t-tests for mean differences across genders. Significance levels are denoted as: *** p<0.01, ** p<0.05, * p<0.10. The first part of each panel describes predicted and achieved grades, where PG represents predicted grades and AG represents achieved grades. Grades are converted into scores using the following points per grade: A* = 6, A = 5, B = 4, C = 3, D = 2, E = 1.The second part of each panel provides information on prior achievement, student characteristics, and application characteristics. The grade 11 point score is based on GCSE exam scores, capped at each student's best eight subjects. Grade 11 scores in maths and English reflect the grades achieved in full GCSE exams, excluding equivalents. Grade 6 scores are total marks achieved in KS2 maths and English exams, with grade 6 TA representing the teacher-assessed National Curriculum level for maths and English. All prior attainment measures from grade 6 and grade 11 are standardised for university applicants by year, accounting also for the GCSE reforms in 2017 that introduced a different grading scale across years. High tariff universities are 38 institutions classified by UCAS as selective based on their entry requirements (see section 2.5 for the list of these institutions). STEM courses are classified based on UCAS Principal subject codes. Learning difficulties and other disabilities are self-reported by students. The '% in subject' indicates the share of boys (Panel A) and girls (Panel B) in each subject. The sample comprises university applicants between 2017-2019. *Source: GRADE 2017-2019*

3 Methods

The objective of this paper is twofold. First, it aims to determine if there are consistent gaps in teacher predictions based on the gender of the student, over and above the student's true ability. Second, it seeks to analyse the underlying causes of any identified gaps.

While a student's true underlying ability in a subject is unobserved, we begin our analysis by assuming that the achieved grade in the externally marked A-level exam, denoted as AG, effectively mirrors the genuine proficiency, denoted as P, of student i in subject s:

$$AG_{i,s} = P_{i,s} + \mu_{i,s} \tag{1}$$

The error term $\mu_{i,s}$ captures any testing noise, such as instances where some students perform below their potential due to off-days during testing, while others may exceed expectations due to fortuitous circumstances. These occurrences are likely stochastic in nature, consistent with the notion of classical measurement error. Measurement error could be severe in student data and failing to use methods that effectively address such errors could lead to invalid inferences (Cunha et al., 2021). We revisit this discussion and consider the potential impact of measurement error in our robustness checks later on.

In contrast, the predicted grade, denoted as PG, is determined through the teacher's subjective judgement of the student's ability, and therefore, it may also be susceptible to influence from external factors. We are particularly interested on the impact of gender, denoted as $G_{i,s}$ and any potential variations in the coefficient β across subjects:

$$PG_{i,s} = P_{i,s} + \beta G_{i,s} + \theta_{i,s} \tag{2}$$

Our central focus revolves around the idea that the interaction between the student and the teacher shapes the relationship between the student's proficiency in the subject, $P_{i,s}$, and their predicted grade, $PG_{i,s}$. Essentially, we embody the proposition that students with similar subject-specific abilities should, on average, receive comparable predicted grades from their teacher. By combining equations (1) and (2) we can effectively eliminate the unobserved subject-specific ability component and any other factors (e.g., family and school characteristics) that influence predicted grades, provided they impact achieved grades in a similar way.

To address the possibility of school policies influencing how teachers predict students with specific characteristics we incorporate school-by-cohort fixed effects, η_s, in our baseline specification. This helps mitigate concerns about non-random student selection to schools. We estimate our model separately for each subject and, therefore, the inclusion of school-by-cohort fixed effects implies that our identification strategy relies on variations in teacher predictions within A-level classes[22]. To account for potential gender-based distinctions in observable characteristics previously linked to teacher bias, we introduce a vector of controls for parental occupation and ethnic background, '$X'_{i,s}$. These variables aim to counteract any gender differences in socio-demographic characteristics which could potentially influence teacher predictions.

Combining equations (1) and (2) and allowing for these additional components, we estimate the following baseline specification:

[22]Here, "class" refers to the intersection of school, cohort and subject -for instance, the maths A-level class of the 2017 cohort at Hills Road Sixth Form College. While teacher fixed effects would have been ideal, our dataset lacks any information pertaining to individual teachers. To address concerns about non-random student assignment to teachers within classes, in one of our robustness checks we limit the sample to small classes where each cohort is potentially taught by only one teacher, therefore ensuring uniform teaching across all students. Since there is no upper limit on the number of students comprising an A-level class, we provide estimates based on a sample restricted to classes of 20 students or fewer, but we also explored variations with larger and smaller class sizes. Importantly, our findings exhibit qualitative consistency across these different sample sizes and we are therefore confident that non-random selection to teachers is not a concern in our models.

$$PG_{i,s} = \rho AG_{i,s} + \beta G_{i,s} + \delta` X'_{i,s} + \eta_s + \epsilon_{i,s} \tag{3}$$

where $\epsilon_{i,s} = \mu_{i,s} - \theta_{i,s}$.

We estimate our model sequentially by progressively increasing the number of controls. We use Ordinary Least Squares (OLS) regression and cluster the standard errors at the school-by-cohort level. The predicted and achieved grades range from 1 to 6, where 1 is equivalent to the lowest grade, E, and 6 is equivalent to the highest grade, A*. We do not standardise the scores because the original scale is consistent across predicted and achieved grades, enabling clear and meaningful interpretation of the results.

The coefficient β captures discrepancies in predicted grades attributed to student gender. We validate our measurement of β through comprehensive robustness checks. These involve exploring alternative specifications, including employing linear probability models, where the dependent variables assesses the probability of under-prediction ($AG_{i,s} > PG_{i,s}$) or over-prediction ($AG_{i,s} < PG_{i,s}$), similar to Burgess and Greaves (2013) and Shi and Zhu (2023). Additionally, we investigate potential nonlinear relationships between achieved and predicted grades, utilise more refined measures of ability derived from the total marks achieved in the external exam rather than the broad grade categories, address potential measurement error in achieved grades using an instrumental variable approach, consider the possibility of statistical discrimination driving our results, and assess the stability of our findings across different samples.

Next, we introduce a measure of general ability to evaluate its predictive validity on teacher grades and its potential impact on the predicted-achieved grade gap based on student gender. As outlined in section 2.4, our measure of general ability is derived from the capped points obtained in eight subject-based exams, providing a comprehensive assessment of student performance across various academic areas. To gain deeper insights into our findings, we explore various hypotheses related to general ability by including additional control variables that may correlate with both general ability and the predicted-achieved grade gap. These variables include application characteristics, subject-specific skills, and proxies for student behaviour. Our conclusions are drawn from the combined evidence accumulated through these analyses.

4 Results

4.1 Predicted-achieved grade gaps

Table 2 presents the association between gender and predicted grades across A-level subjects. Column 1 controls only for the grade achieved, showing that boys are significantly less likely to receive a higher prediction compared to equally achieving girls. This gap ranges between 0.04 and 0.17 grade points in STEM, and between 0.08 and 0.13 grade points in non-STEM. Given that 1 grade point represents the attainment of a full qualification at the lowest grade, and students typically take qualifications in three subjects, these gaps are considerable. Notably, the average gender gap is smaller in STEM (0.09 grade points) compared to non-STEM (0.13 grade points). Introducing socio-demographic controls in column 2 and school-by-cohort fixed effects in column 3 produces nearly identical estimates. This outcome is expected, as we wouldn't anticipate gender differences along these characteristics[23]. Also, introducing these additional controls explains very little additional variation in predicted grades, as indicated from the R-squared reported across specifications.

In the next section, we show our results are stable to several alternative specifications and robustness tests. Appendix Table A.2 indicates that the identified patterns by student gender are also consistent across all levels of the achieved grade distribution, however with some

[23] This stands in contrast, for instance, to studies on ethnic gaps, where the disparities diminish considerably once controlling for socio-economic background, indicating a strong correlation between these measures.

more pronounced gaps among the lowest-achieving students.

The coefficients from column 3 of Table 2 for the other socio-demographic characteristics (parental occupation and ethnicity) are provided in the Appendix Table A.3. In line with the direction of the effects reported in previous studies, we find that Black students and those from lower socio-economic backgrounds tend to receive lower predictions from their teachers across most subjects. Conversely, students from Asian backgrounds tend to receive higher predictions in STEM and lower predictions in non-STEM. The biggest gaps in favour of White students and students from higher socio-economic backgrounds are observed in English literature.

4.2 Robustness checks

In this section, we test the robustness of our findings across four key dimensions: an alternative definition of our dependent variable, the choice of our specification and sensitivity to measurement error, the potential influence of statistical discrimination, and, finally, variations in our sample composition. All robustness checks are run on our preferred specification (specification 3 from Table 2). Taken together, our underlying findings on gender remain essentially unchanged.

4.2.1 Probability of under- and over- prediction

We employ linear probability models to evaluate gender gaps in the likelihood of under- and over-prediction. In these models, the dependent variable equals one if the predicted grade is lower than the achieved grade (PG $<$ AG) in the under-prediction model, and equals one if the predicted grade is higher than the achieved grade (PG $>$ AG) in the over-prediction model. We follow the approach of Shi and Zhu (2023) by excluding students who attained grade E in the under-prediction model, and students who achieved a grade A* in the over-prediction model, as these students, mechanically, cannot be under- or over- predicted. We present our results in Appendix Table A.4. To facilitate interpretation, we also provide mean values for the dependent variables. Consistent with our main results, the findings show that girls are less likely to be under-predicted and more likely to be over-predicted. Specifically, girls are 15% to 41% less likely to be under-predicted in STEM, and 24% to 33% less likely to be under-predicted in non-STEM. Similarly, the probability of over-prediction ranges from 3% to 8% in STEM, and from 8% to 11% in non-STEM. These findings indicate substantial gender gaps in the probability of under- and over- prediction in favour of girls. Moreover, these results support our observation that the gaps are more pronounced in STEM and less evident in the non-STEM subjects.

4.2.2 Specification and measurement error

We address the potential nonlinearity between predicted and achieved grades by incorporating a second-order polynomial. In addition, we demonstrate the robustness of our results when achieved grades enter our specification more flexibly as a fixed effect instead of a linear function. Next, we investigate the possibility that student score distributions within each of the six broad achievement grades vary by gender. If girls' mean scores are situated at higher points within grades, this could potentially explain why teachers assign them higher predictions. To check if this might be the case we control for the total marks achieved in the A-level exam instead of the assigned grade[24]. We present the results for each of these analyses in Appendix Table A.5, columns 1 to 3, respectively. Collectively, these alternative specifications yield estimates that align closely with our main results.

Next, we address the possibility that exam scores may measure underlying ability with error and that our findings on gender gaps might be attributable to this error Cunha et al. (2021).

[24]The availability of total A-level marks is a unique feature of our dataset due to the inclusion of exam results directly from exam boards. This level of detailed information has not been accessible to previous studies that used A-level results. Due to a reform in the marking system, in 2017 and, to a lesser extent in 2018, which transitioned from modular A-levels to linear A-levels, some subjects had a mix of students taking both versions. To ensure a uniform marking system, we standardised the total marks by subject, cohort and marking type.

This could be the case under the assumptions of random error in exam grades and gender differences in underlying skill distributions (Hanushek and Rivkin, 2009). In such a scenario, if boys exhibit lower academic achievement compared to girls, then boys who seemingly perform well on exams may be more likely to have benefited from positive random error (e.g. a higher degree of luck with the exam material), and more so than girls. Consequently, even in the absence of any bias, teachers may be less inclined to predict boys as high achievers compared to girls if their predictions align more closely with students' true abilities.

To address the possibility of measurement error explaining our results, we follow the literature and utilise lagged exam results (from grade 11) and twice-lagged exam results (from grade 6) as instruments for achieved grades at A-level, considering prior performance in maths for STEM and prior performance in English for non-STEM subjects. Relying on lagged and twice-lagged scores as instruments[25] helps mitigate concerns about contemporaneous shocks. Moreover, instrumenting student ability with measures that predate predicted grades can reduce bias from reverse causality, if teacher predictions affect exam performance and, consequently, the grades achieved (Murphy and Wyness, 2020; Leckie and Maragkou, 2024). Essentially, this IV approach allows us to account for measurement error under the assumption of uncorrelated errors over time and to rule out the possibility that our results are driven by differential reactions to predicted grades reflected in the grades achieved.

We present the results in column 4 when using lagged exams as IV, and in column 5 when using twice-lagged exams as IV[26]. Under these specifications, teachers are even more likely to favour girls in some STEM subjects. On the other hand, in some non-STEM subjects the gap is sometimes reduced, although not eliminated. Collectively, we conclude that measurement error in exam scores and reverse causality do not explain the gender gaps in predicted grades within STEM, but contribute to reducing the observed gaps in certain non-STEM subjects

[25]Instead, for example, on A-level scores from other subjects.

[26]Indicative of the cumulative nature of student performance, previous scores exhibit a strong correlation with current ones, as implied by the first-stage coefficients and corresponding F-statistics. The results of the first stage are available upon request.

(specifically in geography, psychology and sociology).

Statistical error has been recognised as a particularly important factor in explaining why students from low socio-economic backgrounds, or certain ethnic minorities, appear as falling behind their equally achieving peers from higher socio-economic or White backgrounds (Hanushek and Rivkin, 2009; Jerrim and Vignoles, 2013). It is possible that for these same reasons these groups of students appear as receiving lower predictions from their teachers in our baseline specification. To explore this hypothesis, we present our preferred IV estimates (from column 5) for socio-economic background and ethnicity in Appendix Table A.6. Remarkably, our findings indicate that the observed gap in predicted grades against Black students and students from lower socio-economic backgrounds is often reversed when measurement error is corrected. This observation aligns with similar findings by van Huizen et al. (2024).

4.2.3 Statistical discrimination

The observed gender gaps in teacher predictions could be due to statistical discrimination driven by the superior performance of girls in A-level exams at the class level. If teachers anticipate higher performance from girls based on past exam results, they might assign different predictions to boys and girls, even if their actual performance is equivalent. This scenario would suggest that the estimated differences in predictions reflect statistical stereotyping rather than explicit bias on the part of teachers (Lavy, 2008; Burgess and Greaves, 2013).

Our analysis does not support this hypothesis since boys consistently receive lower predictions across all subjects, regardless of whether they perform equally or even outperform girls on average. However, we formally test this possibility by adopting a methodology similar to that of Lavy (2008). We focus on university applicants from the 2019 round and estimate separate models for applicants in classes where girls have historically outperformed boys based on average performance in the previous two years, and classes where boys have historically outperformed girls. We present the results in the Appendix Table A7. The results are consistent across classes where girls outperform boys, and vice versa. Thus, our analysis does not find evidence to support the notion of statistical discrimination as an explanation for our results.

4.2.4 Sample selection

In our final set of robustness checks we evaluate the stability of our estimates across variations in the analytical sample. First, we compare the results from the analytical sample to those in the full sample of university applicants, conditional only on having non-missing predicted and achieved grades. Next, we address the possibility of non-random assignment to teachers within A-level classes by restricting our sample to small classes in an attempt to ensure that all students are taught by the same teacher[27]. Moreover, we consider the possibility that the relationship between predicted and achieved grades might be different for students with disability, particularly those with learning difficulties. As a robustness check we omit disabled students from the analysis (8% of the sample). Subsequently, we exclude students at the tails of the predicted-achieved grade difference distribution i.e. those who are under- or over- predicted by 3 or more grades (5% of the sample), and present the results in column 4. Then, we restrict our sample to only "typical applicants" being those who studied for 3 A-level subjects (80% of applicants). Finally, we drop students who applied to institutions that require predicted grades earlier than the typical January's deadline (10%)[28]. We present the results in the Appendix Table A8 columns 1 to 5 respectively. Taken together, the evidence from these additional analyses suggests that our main results are robust to alternative sample definitions.

[27]We present results when defining single-teacher classes as those with no more than 20 students but, since there is no cap in the A-level student-teacher ratio, we have experimented with alternative definitions as well.

[28]Specifically, the two most selective universities of Oxford and Cambridge require predicted grades by October of each academic year. Some courses in medicine, dentistry, and veterinary medicine also have the same early deadline. However as this varies by institution we do not exclude applicants from these courses.

4.3 Understanding the predicted-achieved grade gaps by gender

4.3.1 The role of general ability

Having established the presence and direction of the predicted-achieved grade gaps by student gender, we now shift our focus to understanding their underlying causes. We are particularly interested in the role of general ability. As evidenced in Table 1 the average grade 11 capped score for girls is more than 20% of a standard deviation higher compared to the score for boys in STEM, and about 17% of a standard deviation higher in non-STEM. Further, Figure 3 demonstrated that this gender gap in overall competence gap remains substantial when accounting for differences in subject-specific skills. To examine the link between general ability and the predicted-achieved grade gap we re-estimate our baseline specification augmented with the capped grade 11 point score.

We present the results in column 4 of Table 2. Controlling for general ability almost entirely diminishes the gender gap in three of the five non-STEM subjects. However, the gap in the two mostly girl-dominated subjects of psychology and sociology, persists significantly. Conversely, in STEM, controlling for general ability leads to a gender gap reversal. Girls with similar general ability and achieved grades receive lower predicted grades compared to boys. The only exception is physics where the gender gap becomes statistically insignificant. Interestingly, the trend in the magnitude of the gender gap in STEM is similar to that observed in the non-STEM subjects, but in the opposite direction. Specifically, the largest gender gaps are seen in the girl-dominated subjects of biology and chemistry. These gaps exceed the raw gaps observed in specification 1. Evidently, the grade 11 capped score accounts for a significant portion of the variation in predicted grades, as indicated by the considerable increase in the R-squared, particularly in non-STEM subjects.

These findings underscore the significant impact of grading biases favouring general ability beyond subject-specific proficiency, highlighting girls' comparative advantage in overall competence. The predictive power of general ability on predicted grades shows the value of more comprehensive assessments. To help us explain the predicted-achieved grade gap by student gender it is important to understand the sources of differences in general ability. We do this in the next section by introducing additional statistical controls for various mechanisms potentially associated with general ability and teacher predictions. These include factors such as differential ambitiousness, student behaviour, differential progress, and cross-disciplinary skills.

4.3.2 Unpacking general ability

In this section, we investigate further the factors influencing the difference between predicted and achieved grades based on student gender by incorporating additional controls into our baseline model. This allows us to explore other potential reasons for the observed gender gap in predicted grades, and to gain a deeper understanding of how our measure of general ability relates to it. We present our findings in Table 3.

First, we explore the potential impact of differential ambitiousness, proxied by whether students have applied to at least one high-tariff institution and whether they have applied to a STEM course, given the higher requirements associated with these courses. Next, we investigate two potential proxies for student behaviour: a lagged measure of teacher-assessed grade from grade 6, and disability indicators (learning-related or other disabilities). Lavy (2008) suggests that if students' behaviour remains consistent over time, then a teacher evaluation from a previous grade may capture this, thereby controlling for variations in behaviour. Similarly, Burgess and Greaves (2013) suggest that having learning-related disabilities could be correlated with poorer behaviour, or that being labelled as having a learning difficulty may reinforce teachers' low expectations. We report the results on gender with the inclusion of these additional controls in columns 1 to 4, respectively. Overall, we do not find evidence that differences in application characteristics between boys and girls, or differences in student behaviour along the dimensions captured in our variables, explain the predicted-achieved grade gap by student gender.

It is possible that the gender gap in predicted grades captures differential progress. Since teacher predictions are made several months prior to the actual exams (typically 9 months in advance), it's plausible that at the time of prediction, the grades accurately mirror the students' abilities at that point. Consequently, any unanticipated progress made by certain groups between the prediction and the exam may account for the differential predictions observed (Murphy and Wyness, 2020). To check for this possibility we control for past exam results from grade 11 (age 16) and grade 6 (age 11) in the relevant subject (maths in STEM and English in non-STEM). The grade 6 score could be capturing subject-specific cognitive skills or subject performance at baseline, while the grade 11 score could capture everything that happened between ages 11 and 16. The achieved grade at A-level, then, will capture progress from age 16 to age 18. Introducing these measures in column 5 does little to change our results in STEM A-level subjects. In non-STEM, considering prior attainment in English diminishes the predicted-achieved grade gap against boys, although not entirely. Overall, the findings suggest that differential progress does not offer an explanation for the gender gaps observed in our main results for STEM, while it provides some explanation for non-STEM.

Similar to the notion of our measure of general ability, performance in other subjects might serve as a proxy for latent capabilities considered by teachers when making predictions about student grades. To check for this, we augment the specification presented in column 5 by incorporating additional controls for lagged cross-subject skills from grade 11 and grade 6 (i.e., English in STEM and maths in non-STEM). We present the results in column 6. The findings show that the gap in biology and chemistry is now reversed, in favour of boys, aligning with our results on general ability. The gaps in the other three STEM subjects remain in favour of girls but are substantially reduced in magnitude. This is probably because performance in English correlates only partly with our more comprehensive measure of ability based on eight subjects. The gap in non-STEM subjects remains in favour of girls, particularly in psychology and sociology, also in line with our previous results.

Lastly, in column 7 we incorporate all the previously mentioned controls along with the grade 11 capped score. The results closely mirror those reported in column 4 of Table 2. Moreover, the inclusion of all the additional controls alongside the grade 11 score explains very little (if any) additional variation in predicted grades compared to the inclusion of the grade 11 score alone. We present the coefficients on the full-specification in the Appendix Table A9. Consistent with expectations, the results show that applying to a high-tariff university correlates with higher predicted grades. Similarly, applying to a STEM course is linked to higher predictions in STEM subjects but lower predictions in non-STEM subjects, with economics being the exception, for the reasons previously discussed. Notably, students reporting learning difficulties tend to receive lower predictions from their teachers despite achieving the same results. Moreover, both subject-specific and cross-subject skills at grade 11 exhibit a positive association with predicted grades. However, cross-subject skills at grade 6 display a mostly negative correlation with predicted grades.

Our findings demonstrate that the factors influencing predicted grades extend beyond subject-specific competencies and factors like varying levels of ambition and rates of progress. Previous research has linked gender disparities in academic performance to traits like hyperactivity and depressive symptoms (Gutman and Vorhaus, 2012; Rothon et al., 2009). These findings are consistent with evidence that noncognitive skills foster cognitive abilities, as measured by achievement tests (Cunha and Heckman, 2008), and that personality traits might hold more significance in teacher-assigned grades compared to performance on standardised tests (Borghans et al., 2016). Regardless of the specific mechanisms through which our measure of general ability intersects with predicted grades, our results indicate that teachers don't maintain impartiality towards such behaviours when predicting student grades for university applications. This raises significant concerns regarding the current university application system, which relies on subjective assessments of students' abilities. Based on this evidence, we draw our conclusions in the following section.

Table 2: Predicted-achieved grade gaps by student gender

		(1)	(2)	(3)	(4)
Biology	Girl	0.047***	0.051***	0.063***	-0.058***
(n=96056)		(0.007)	(0.007)	(0.006)	(0.006)
	R^2	0.45	0.46	0.48	0.54
Chemistry	Girl	0.053***	0.055***	0.052***	-0.078***
(n=80458)		(0.007)	(0.007)	(0.006)	(0.006)
	R^2	0.47	0.47	0.49	0.55
Economics	Girl	0.118***	0.115***	0.108***	-0.032***
(n=39504)		(0.011)	(0.011)	(0.009)	(0.009)
	R^2	0.35	0.35	0.34	0.46
Maths	Girl	0.101***	0.099***	0.104***	-0.028***
(n=108474)		(0.006)	(0.006)	(0.005)	(0.005)
	R^2	0.52	0.52	0.54	0.58
Physics	Girl	0.172***	0.170***	0.163***	-0.011
(n=52722)		(0.009)	(0.009)	(0.009)	(0.009)
	R^2	0.50	0.51	0.53	0.58
English Lit	Girl	0.077***	0.082***	0.102***	0.025***
(n=64095)		(0.009)	(0.009)	(0.008)	(0.007)
	R^2	0.38	0.38	0.36	0.51
Geography	Girl	0.099***	0.100***	0.106***	0.013**
(n=51527)		(0.008)	(0.008)	(0.007)	(0.006)
	R^2	0.43	0.43	0.42	0.54
History	Girl	0.130***	0.133***	0.150***	0.018***
(n=75221)		(0.007)	(0.007)	(0.006)	(0.006)
	R^2	0.38	0.38	0.36	0.50
Psychology	Girl	0.119***	0.120***	0.137***	0.086***
(n=98676)		(0.007)	(0.007)	(0.006)	(0.006)
	R^2	0.37	0.38	0.40	0.48
Sociology	Girl	0.130***	0.133***	0.142***	0.087***
(n=55620)		(0.010)	(0.010)	(0.009)	(0.008)
	R^2	0.30	0.30	0.31	0.41
Achieved grade		✓	✓	✓	✓
Socio-demogr. controls			✓	✓	✓
School fixed effects				✓	✓
G11 point score					✓

Notes: The dependent variable is the predicted grade. All regressions control for the achieved grade. Predicted and achieved grades are scored between E to A* and transformed to the following scores per grade: A*=6, A=5, B=4, C=3, D=2, E=1. Column 1 shows the row gender gap in predicted grades conditional on achieved grade only. Column 2 adds controls for socio-demographic characteristics (parental occupation and ethnicity). Column 3 adds school fixed effects. Column 4 adds the standardised point score from age 16 exams capped at each student's best eight subjects. The standard errors are clustered at the school-by-cohort level and reported in parentheses. Significance levels are denoted as: *** p<0.01, ** p<0.05, * p<0.10. The sample comprises university applicants between 2017-2019. *Source: GRADE 2017-2019*

Table 3: Mechanisms of the predicted-achieved grade gaps by gender

		(1) High tarrif	(2) STEM	(3) Disab.	(4) G6 TA	(5) Sub. skill	(6) Cross-sub.	(7) All	(8) All+G11
Biology	Girl	0.056***	0.061***	0.062***	0.072***	0.067***	-0.011*	-0.000	-0.046***
(n=96056)		(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.005)	(0.005)
	R^2	0.51	0.48	0.48	0.48	0.51	0.50	0.55	0.57
Chemistry	Girl	0.045***	0.050***	0.051***	0.062***	0.062***	-0.033***	-0.010	-0.053***
(n=80458)		(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
	R^2	0.52	0.49	0.49	0.50	0.52	0.51	0.56	0.57
Economics	Girl	0.096***	0.109***	0.109***	0.121***	0.111***	0.024**	0.035***	-0.020**
(n=39504)		(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
	R^2	0.40	0.34	0.34	0.35	0.40	0.38	0.47	0.50
Maths	Girl	0.095***	0.107***	0.104***	0.109***	0.071***	0.029***	0.030***	-0.001
(n=108474)		(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
	R^2	0.56	0.54	0.54	0.54	0.57	0.55	0.60	0.61
Physics	Girl	0.138***	0.163***	0.163***	0.165***	0.131***	0.071***	0.043***	-0.015*
(n=52722)		(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
	R^2	0.55	0.53	0.53	0.53	0.55	0.55	0.59	0.60
English Lit	Girl	0.104***	0.104***	0.101***	0.086***	0.049***	0.125***	0.073***	0.028***
(n=64095)		(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.007)	(0.007)
	R^2	0.41	0.36	0.36	0.38	0.45	0.42	0.51	0.54
Geography	Girl	0.104***	0.105***	0.106***	0.084***	0.031***	0.137***	0.063***	0.013**
(n=51527)		(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.006)
	R^2	0.47	0.43	0.43	0.44	0.47	0.47	0.53	0.56
History	Girl	0.133***	0.150***	0.149***	0.120***	0.043***	0.163***	0.059***	0.001
(n=75221)		(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
	R^2	0.42	0.36	0.36	0.38	0.43	0.42	0.49	0.53
Psychology	Girl	0.143***	0.135***	0.136***	0.123***	0.089***	0.171***	0.119***	0.087***
(n=98676)		(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
	R^2	0.44	0.40	0.40	0.41	0.43	0.43	0.48	0.51
Sociology	Girl	0.142***	0.142***	0.141***	0.128***	0.096***	0.171***	0.119***	0.085***
(n=55620)		(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.008)	(0.008)
	R^2	0.36	0.31	0.31	0.33	0.36	0.35	0.41	0.44

See next page for table notes.

Notes: The dependent variable is the predicted grade. Predicted and achieved grades are scored between E to A* and transformed to the following scores per grade: A*=6, A=5, B=4, C=3, D=2, E=1. All regressions control for the achieved grade, socio-demographic characteristics (parental occupation and ethnicity) and school fixed effects. Column 1 controls for applying to high tariff institution, column 2 for applying to a STEM course, column 3 for disability (learning or other) and column 4 for grade 6 subject-specific teacher assessment (maths for STEM and English for non-STEM). Column 5 controls for subject-specific skills (scores from grades 6 and 11 in the relevant subject), while column 6 for both subject-specific and cross-subject skills (English for STEM and maths for non-STEM). Column 7 includes all controls from columns 1 to 6 and column 8 includes all controls plus the grade 11 point score. The standard errors are clustered at the school-by-cohort level and reported in parentheses. Significance levels are denoted as: *** $p<0.01$, ** $p<0.05$, * $p<0.10$. The sample comprises university applicants between 2017-2019. *Source: GRADE 2017-2019*

5 Conclusions

This paper utilises comprehensive administrative records to provide evidence for the existence of significant gender-based gaps in teacher-predicted grades for university applications. It reveals that boys are predicted less favourably compared to equally achieving girls in the same class. These gaps are evident throughout the entire achieved grade distribution and across various subject areas. The gap between predicted and achieved grades across genders can be traced to differences in general ability between boys and girls, with girls demonstrating significantly higher levels of overall competence. When controlling for these skill differentials, the gender gap in predicted grades against boys is substantially reduced in non-STEM subjects. In STEM, the gap shifts in favour of boys, indicating a complex pattern of gender-based prediction bias. In future work, the focus will be on identifying the root causes of these residual biases, especially with regard to the identified gender trends in STEM compared to non-STEM, and their potential implications for university application outcomes.

The implications of these results are potentially serious. Teacher-predicted grades have a direct influence on university admissions and can ultimately impact students' career opportunities. The results suggest that teachers consider factors beyond subject-specific knowledge when making predictions, indicating that broader attributes or societal biases may influence their judgement. Consequently, the current educational system unintentionally perpetuates gender-based disparities, creating unequal opportunities for students. The paper emphasises the need for further research to fully understand the complex mechanisms contributing to the gender gaps in teacher predictions. Additionally, it calls for a comprehensive re-evaluation of university admissions practices to ensure fairness. One proposed solution is to re-frame the university admissions process, shifting away from reliance on teacher predictions. A new approach could involve creating a system where students have more control over their university choices by allowing them to adjust their applications after receiving their actual exam results. In this model, universities would make final offers based on those results rather than on teacher-predicted grades. This approach would increase fairness by reducing subjective judgements, allowing students' actual achievements to drive admissions decisions.

References

Adamecz, A., M. Henderson, and N. Shure (2024). Intergenerational educational mobility–the role of non-cognitive skills. *Education Economics 32*(1), 59–78.

Black, N. and S. C. de New (2020). Short, heavy and underrated? teacher assessment biases by children's body size. *Oxford Bulletin of Economics and Statistics 82*(5), 961–987.

Blank, R. M. (1991). The effects of double-blind versus single-blind reviewing: Experimental evidence from the american economic review. *The American Economic Review*, 1041–1067.

Bonesrønning, H. (2008). The effect of grading practices on gender differences in academic performance. *Bulletin of Economic Research 60*(3), 245–264.

Borghans, L., B. H. Golsteyn, J. J. Heckman, and J. E. Humphries (2016). What grades and achievement tests measure. *Proceedings of the national Academy of Sciences 113*(47), 13354–13359.

Botelho, F., R. A. Madeira, and M. A. Rangel (2015). Racial discrimination in grading: Evidence from brazil. *American Economic Journal: Applied Economics 7*(4), 37–52.

Breda, T. and S. T. Ly (2015). Professors in core science fields are not always biased against women: Evidence from france. *American Economic Journal: Applied Economics 7*(4), 53–75.

Breda, T. and C. Napp (2019). Girls' comparative advantage in reading can largely explain the gender gap in math-related fields. *Proceedings of the National Academy of Sciences 116*(31), 15435–15440.

Burgess, S. and E. Greaves (2013). Test scores, subjective assessment, and stereotyping of ethnic minorities. *Journal of Labor Economics 31*(3), 535–576.

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

Cavaglia, C., S. Machin, S. McNally, and J. Ruiz-Valenzuela (2020). Gender, achievement, and subject choice in english education. *Oxford Review of Economic Policy 36*(4), 816–835.

Cornwell, C., D. B. Mustard, and J. Van Parys (2013). Noncognitive skills and the gender disparities in test scores and teacher assessments: Evidence from primary school. *Journal of Human resources 48*(1), 236–264.

Cunha, F. and J. J. Heckman (2008). Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation. *Journal of human resources 43*(4), 738–782.

Cunha, F., E. Nielsen, and B. Williams (2021). The econometrics of early childhood human capital and investments. *Annual Review of Economics 13*, 487–513.

Dee, T. S. (2005). A teacher like me: Does race, ethnicity, or gender matter? *American Economic Review 95*(2), 158–165.

Delaney, J. and P. J. Devereux (2023). Gender differences in teacher judgement of comparative advantage. *Available at SSRN 4664225*.

Dilnot, C. (2018). The relationship between a-level subject choice and league table score of university attended: the 'facilitating', the 'less suitable', and the counter-intuitive. *Oxford Review of Education 44*(1), 118–137.

Falch, T. and L. R. Naper (2013). Educational evaluation schemes and gender gaps in student achievement. *Economics of Education Review 36*, 12–25.

Ferman, B. and L. F. Fontes (2022). Assessing knowledge or classroom behavior? evidence of teachers' grading bias. *Journal of Public Economics 216*, 104773.

Goulas, S., S. Griselda, and R. Megalokonomou (2022). Comparative advantage and gender gap in stem. *Journal of Human Resources*, 0320–10781R2.

Gutman, L. and J. Vorhaus (2012). The impact of pupil behaviour and wellbeing on educational outcomes.

Hanushek, E. A. and S. G. Rivkin (2009). Harming the best: How schools affect the black-white achievement gap. *Journal of policy analysis and management 28*(3), 366–393.

Jerrim, J. and A. Vignoles (2013). Social mobility, regression to the mean and the cognitive development of high ability children from disadvantaged homes. *Journal of the Royal Statistical Society Series A: Statistics in Society 176*(4), 887–906.

Lavy, V. (2008). Do gender stereotypes reduce girls' or boys' human capital outcomes? evidence from a natural experiment. *Journal of Public Economics 92*(10-11), 2083–2105.

Lavy, V. and R. Megalokonomou (2019). Persistency in teachers' grading bias and effects on longer-term outcomes: University admissions exams and choice of field of study. Technical report, National Bureau of Economic Research.

Lavy, V. and R. Megalokonomou (2024). The short-and the long-run impact of gender-biased teachers. *American Economic Journal: Applied Economics 16*(2), 176–218.

Lavy, V. and E. Sand (2018). On the origins of gender gaps in human capital: Short-and long-term consequences of teachers' biases. *Journal of Public Economics 167*, 263–279.

Leckie, G. and K. Maragkou (2024). Student sociodemographic and school type differences in teacher-predicted vs. achieved grades for university admission. *Higher Education*, 1–27.

Mendolia, S. and I. Walker (2014). The effect of personality traits on subject choice and performance in high school: Evidence from an english cohort. *Economics of Education Review 43*, 47–65.

Murphy, R. and G. Wyness (2020). Minority report: the impact of predicted grades on university admissions of disadvantaged groups. *Education Economics 28*(4), 333–350.

Rothon, C., J. Head, C. Clark, E. Klineberg, V. Cattell, and S. Stansfeld (2009). The impact of psychological distress on the educational achievement of adolescents at the end of compulsory education. *Social psychiatry and psychiatric epidemiology 44*, 421–427.

Shi, Y. and M. Zhu (2023). "Model minorities" in the classroom? Positive evaluation bias towards Asian students and its consequences. *Journal of Public Economics 220*, 104838.

Smith, N. R., L. Marshall, M. Albakri, M. Smuk, A. Hagell, and S. Stansfeld (2021). Adolescent mental health difficulties and educational attainment: findings from the uk household longitudinal study. *BMJ open 11*(7), e046792.

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

van Huizen, T., M. Jacobs, and M. Oosterveen (2024). Teacher bias or measurement error? *arXiv preprint arXiv:2401.04200*.

Table A.1: Descriptive statistics of other socio-demographic characteristics

	(1) Biology	(2) Chemistry	(3) Econ.	(4) Maths	(5) Physics	(6) Eng. Lit.	(7) Geogr.	(8) History	(9) Psychol.	(10) Sociol.
Panel A: Boys										
Part 1: Parental occupation										
Manag. & profess.	0.539	0.541	0.551	0.571	0.615	0.519	0.591	0.536	0.485	0.422
	(0.498)	(0.498)	(0.497)	(0.495)	(0.487)	(0.500)	(0.492)	(0.499)	(0.500)	(0.494)
Intermediate	0.114	0.109	0.109	0.109	0.105	0.127	0.118	0.124	0.126	0.130
	(0.318)	(0.311)	(0.311)	(0.311)	(0.306)	(0.333)	(0.322)	(0.330)	(0.332)	(0.336)
Small employers	0.077	0.078	0.075	0.069	0.055	0.069	0.066	0.066	0.076	0.086
	(0.266)	(0.268)	(0.263)	(0.253)	(0.228)	(0.253)	(0.248)	(0.248)	(0.265)	(0.281)
Lower supervisory	0.056	0.054	0.040	0.053	0.050	0.053	0.052	0.052	0.060	0.064
	(0.229)	(0.226)	(0.196)	(0.224)	(0.218)	(0.224)	(0.222)	(0.221)	(0.237)	(0.244)
Semi-routine	0.073	0.075	0.074	0.066	0.056	0.079	0.055	0.073	0.087	0.104
	(0.260)	(0.263)	(0.261)	(0.249)	(0.230)	(0.270)	(0.229)	(0.259)	(0.282)	(0.305)
Routine	0.052	0.050	0.047	0.046	0.039	0.055	0.045	0.054	0.065	0.079
	(0.221)	(0.218)	(0.213)	(0.209)	(0.194)	(0.228)	(0.206)	(0.225)	(0.246)	(0.269)
Unknown	0.090	0.094	0.104	0.086	0.080	0.099	0.074	0.096	0.101	0.115
	(0.286)	(0.292)	(0.305)	(0.281)	(0.271)	(0.298)	(0.261)	(0.295)	(0.302)	(0.320)
Part 2: Ethnic group										
White	0.675	0.594	0.558	0.664	0.730	0.746	0.848	0.781	0.733	0.682
	(0.469)	(0.491)	(0.497)	(0.472)	(0.444)	(0.435)	(0.359)	(0.414)	(0.442)	(0.466)
Black	0.061	0.075	0.100	0.054	0.044	0.062	0.026	0.053	0.063	0.092
	(0.239)	(0.263)	(0.300)	(0.226)	(0.204)	(0.240)	(0.160)	(0.224)	(0.242)	(0.288)
Indian	0.066	0.087	0.111	0.079	0.061	0.033	0.030	0.030	0.042	0.037
	(0.248)	(0.282)	(0.314)	(0.270)	(0.239)	(0.179)	(0.172)	(0.170)	(0.200)	(0.190)
Pakistani	0.060	0.076	0.045	0.042	0.024	0.037	0.016	0.030	0.049	0.058
	(0.238)	(0.264)	(0.208)	(0.200)	(0.152)	(0.190)	(0.127)	(0.172)	(0.215)	(0.235)
Bangladeshi	0.027	0.031	0.035	0.023	0.011	0.023	0.012	0.019	0.024	0.036
	(0.161)	(0.175)	(0.185)	(0.150)	(0.106)	(0.150)	(0.107)	(0.135)	(0.152)	(0.187)
Chinese	0.009	0.014	0.015	0.017	0.019	0.004	0.006	0.004	0.004	0.003

	(0.096)	(0.116)	(0.122)	(0.129)	(0.135)	(0.062)	(0.077)	(0.062)	(0.064)	(0.058)
Other	0.103	0.124	0.135	0.121	0.113	0.095	0.061	0.083	0.086	0.091
	(0.303)	(0.330)	(0.342)	(0.326)	(0.316)	(0.293)	(0.240)	(0.276)	(0.280)	(0.287)
N	34235	37211	27272	65356	41551	13498	23586	31143	22261	11280
Panel B: Girls										
Part 1: Parental occupation										
Manag. & profess.	0.550	0.554	0.563	0.572	0.591	0.550	0.608	0.575	0.506	0.458
	(0.498)	(0.497)	(0.496)	(0.495)	(0.492)	(0.497)	(0.488)	(0.494)	(0.500)	(0.498)
Intermediate	0.111	0.107	0.110	0.108	0.107	0.119	0.117	0.118	0.123	0.128
	(0.315)	(0.309)	(0.313)	(0.310)	(0.309)	(0.324)	(0.321)	(0.323)	(0.328)	(0.334)
Small employers	0.080	0.079	0.079	0.072	0.066	0.069	0.064	0.063	0.075	0.084
	(0.271)	(0.269)	(0.270)	(0.259)	(0.248)	(0.254)	(0.246)	(0.243)	(0.263)	(0.278)
Lower supervisory	0.052	0.052	0.045	0.053	0.054	0.048	0.050	0.046	0.056	0.057
	(0.222)	(0.223)	(0.208)	(0.223)	(0.225)	(0.213)	(0.218)	(0.209)	(0.231)	(0.232)
Semi-routine	0.068	0.065	0.062	0.061	0.056	0.069	0.052	0.063	0.080	0.095
	(0.251)	(0.247)	(0.241)	(0.240)	(0.230)	(0.253)	(0.221)	(0.243)	(0.271)	(0.294)
Routine	0.048	0.049	0.043	0.045	0.041	0.046	0.036	0.045	0.058	0.065
	(0.214)	(0.216)	(0.203)	(0.207)	(0.199)	(0.209)	(0.187)	(0.207)	(0.234)	(0.246)
Unknown	0.091	0.094	0.097	0.089	0.086	0.099	0.073	0.090	0.102	0.112
	(0.287)	(0.292)	(0.296)	(0.284)	(0.280)	(0.298)	(0.260)	(0.286)	(0.303)	(0.316)
Part 2: Ethnic group										
White	0.651	0.618	0.629	0.689	0.745	0.770	0.850	0.823	0.724	0.672
	(0.477)	(0.486)	(0.483)	(0.463)	(0.436)	(0.421)	(0.357)	(0.382)	(0.447)	(0.469)
Black	0.052	0.052	0.075	0.044	0.034	0.058	0.026	0.038	0.066	0.100
	(0.221)	(0.222)	(0.264)	(0.205)	(0.182)	(0.234)	(0.158)	(0.191)	(0.249)	(0.300)
Indian	0.073	0.087	0.088	0.073	0.060	0.026	0.031	0.027	0.043	0.039
	(0.261)	(0.282)	(0.284)	(0.261)	(0.237)	(0.160)	(0.174)	(0.162)	(0.204)	(0.194)
Pakistani	0.069	0.073	0.047	0.042	0.029	0.029	0.016	0.023	0.045	0.052
	(0.253)	(0.260)	(0.212)	(0.202)	(0.169)	(0.168)	(0.125)	(0.150)	(0.207)	(0.222)
Bangladeshi	0.027	0.030	0.033	0.024	0.015	0.017	0.010	0.013	0.023	0.031
	(0.161)	(0.170)	(0.180)	(0.154)	(0.122)	(0.128)	(0.102)	(0.112)	(0.151)	(0.172)
Chinese	0.011	0.014	0.010	0.014	0.014	0.005	0.004	0.003	0.005	0.004

	(0.102)	(0.116)	(0.099)	(0.118)	(0.117)	(0.068)	(0.066)	(0.057)	(0.071)	(0.061)
Other	0.118	0.127	0.117	0.113	0.103	0.095	0.062	0.073	0.093	0.102
	(0.323)	(0.333)	(0.321)	(0.317)	(0.304)	(0.293)	(0.241)	(0.261)	(0.291)	(0.302)
N	61821	43247	12232	43118	11171	50597	27941	44078	76415	44340
Panel C: Difference (Boys - Girls)										
Part 1: Parental occupation										
Manag. & profess.	0.011**	0.013***	0.012*	0.001	-0.025***	0.031***	0.017***	0.039***	0.021***	0.036***
Intermediate	-0.003	-0.002	0.001	-0.001	0.002	-0.007*	-0.001	-0.006*	-0.004	-0.002
Small employers	0.003	0.001	0.004	0.003*	0.011***	0.001	-0.001	-0.003	-0.001	-0.002
Lower supervisory	-0.003*	-0.001	0.005*	-0.000	0.004	-0.005*	-0.002	-0.006***	-0.003	-0.007**
Semi-routine	-0.005**	-0.009***	-0.012***	-0.005***	-0.000	-0.011***	-0.004	-0.009***	-0.007***	-0.008**
Routine	-0.003*	-0.001	-0.004	-0.001	0.002	-0.009***	-0.008***	-0.009***	-0.007***	-0.014***
Unknown	0.001	-0.000	-0.007*	0.002	0.006*	0.000	-0.001	-0.006**	0.000	-0.003
Part 2: Ethnic group										
White	-0.024***	0.025***	0.071***	0.025***	0.015**	0.024***	0.003	0.042***	-0.009**	-0.010*
Black	-0.009***	-0.023***	-0.025***	-0.010***	-0.009***	-0.003	-0.001	-0.015***	0.004*	0.009**
Indian	0.008***	0.000	-0.023***	-0.006***	-0.001	-0.007***	0.001	-0.003*	0.002	0.002
Pakistani	0.009***	-0.003	0.002	0.001	0.006***	-0.008***	-0.000	-0.007***	-0.004*	-0.006**
Bangladeshi	-0.000	-0.002	-0.002	0.001	0.004***	-0.006***	-0.001	-0.006***	-0.000	-0.006**
Chinese	0.001	0.000	-0.005***	-0.003***	-0.005**	0.001	-0.002*	-0.001	0.001	0.000
Other	0.016***	0.003	-0.018***	-0.008***	-0.010**	0.000	0.001	-0.010***	0.007**	0.011***
N	96056	80458	39504	108474	52722	64095	51527	75221	98676	55620

Notes: Columns 1–10 report mean values and standard deviations (in parentheses) for each A-level subject, respectively. Panel A provides descriptive statistics for boys, Panel B for girls, and Panel C presents the results of t-tests for mean differences across genders. Significance levels are denoted as: *** p<0.01, ** p<0.05, * p<0.10. The first part of each panel describes self-reported measures of parental occupation. The second part describes self-reported ethnic groups. The sample comprises university applicants between 2017-2019.

Table A.2: Predicted-achieved grade gaps by student gender across the achieved grade distribution

		(1)	(2)	(3)	(4)	(5)	(6)
Achieved Grade		E	D	C	B	A	A*
Biology	Girl	0.126***	0.095***	0.070***	0.055***	0.026**	-0.000
		(0.024)	(0.015)	(0.013)	(0.012)	(0.012)	(0.016)
	N	8867	17803	22396	21482	18178	7330
Chemistry	Girl	0.147***	0.117***	0.069***	0.047***	-0.001	-0.039**
		(0.030)	(0.019)	(0.015)	(0.012)	(0.011)	(0.016)
	N	6317	12107	16988	19955	18737	6354
Economics	Girl	0.108	0.120***	0.158***	0.122***	0.050***	-0.056
		(0.073)	(0.034)	(0.023)	(0.018)	(0.018)	(0.040)
	N	1453	4679	9440	12347	9122	2463
Maths	Girl	0.148***	0.169***	0.151***	0.112***	0.075***	-0.011
		(0.030)	(0.019)	(0.015)	(0.012)	(0.009)	(0.008)
	N	6412	12265	18491	23948	29474	17884
Physics	Girl	0.177***	0.225***	0.221***	0.171***	0.064***	0.026
		(0.048)	(0.032)	(0.023)	(0.020)	(0.018)	(0.020)
	N	4470	8298	11012	12398	11485	5059
English Lit	Girl	0.152**	0.117***	0.125***	0.112***	0.046**	0.063*
		(0.065)	(0.025)	(0.016)	(0.016)	(0.023)	(0.032)
	N	2364	9491	18156	19153	9967	4964
Geography	Girl	0.104	0.135***	0.109***	0.118***	0.067***	0.133***
		(0.065)	(0.024)	(0.016)	(0.014)	(0.017)	(0.042)
	N	1632	6257	13569	16513	10642	2914
History	Girl	0.082	0.139***	0.172***	0.164***	0.089***	0.058*
		(0.053)	(0.019)	(0.012)	(0.011)	(0.015)	(0.030)
	N	2416	9669	21010	25037	13541	3548
Psychology	Girl	0.162***	0.099***	0.168***	0.136***	0.135***	0.084**
		(0.030)	(0.016)	(0.012)	(0.012)	(0.019)	(0.039)
	N	5997	16242	28157	29016	14315	4949
Sociology	Girl	0.107	0.098***	0.171***	0.150***	0.149***	0.119**
		(0.067)	(0.025)	(0.017)	(0.016)	(0.027)	(0.053)
	N	2148	7669	15724	18180	8695	3204

Notes: The dependent variable is the predicted grade. Predicted and achieved grades are scored between E to A* and transformed to the following scores per grade: A*=6, A=5, B=4, C=3, D=2, E=1. All regressions control for socio-demographic characteristics (parental occupation and ethnicity) and school-by-cohort fixed effects. The standard errors are clustered at the school-by-cohort level and reported in parentheses. Significance levels are denoted as: *** $p<0.01$, ** $p<0.05$, * $p<0.10$. The sample comprises university applicants between 2017-2019. *Source: GRADE 2017-2019*

Table A.3: Predicted-achieved grade gaps specification 3

	(1) Biology	(2) Chemistry	(3) Economics	(4) Maths	(5) Physics	(6) English Lit	(7) Geography	(8) History	(9) Psychology	(10) Sociology
Girl	0.063***	0.052***	0.108***	0.104***	0.163***	0.102***	0.106***	0.150***	0.137***	0.142***
	(0.006)	(0.006)	(0.009)	(0.005)	(0.009)	(0.008)	(0.007)	(0.006)	(0.006)	(0.009)
Intermediate	-0.015*	-0.021**	-0.040***	-0.023***	-0.002	-0.042***	-0.014	-0.053***	-0.016**	-0.016
	(0.008)	(0.008)	(0.012)	(0.007)	(0.011)	(0.009)	(0.010)	(0.008)	(0.007)	(0.010)
Small employers	-0.023**	-0.032***	-0.037**	-0.018*	-0.014	-0.076***	-0.037***	-0.064***	-0.021**	-0.048***
	(0.009)	(0.010)	(0.015)	(0.009)	(0.014)	(0.012)	(0.013)	(0.011)	(0.009)	(0.012)
Lower supervisory	-0.024**	-0.034***	-0.023	-0.042***	-0.029*	-0.084***	-0.023	-0.050***	-0.024**	-0.024*
	(0.010)	(0.011)	(0.019)	(0.010)	(0.015)	(0.014)	(0.014)	(0.012)	(0.010)	(0.013)
Semi-routine	-0.021**	-0.048***	-0.053***	-0.026***	-0.015	-0.088***	-0.050***	-0.075***	-0.018**	-0.027**
	(0.009)	(0.010)	(0.016)	(0.010)	(0.014)	(0.011)	(0.015)	(0.011)	(0.009)	(0.011)
Routine	-0.034***	-0.037***	-0.032	-0.028**	-0.031*	-0.095***	-0.081***	-0.067***	-0.017*	-0.030**
	(0.011)	(0.012)	(0.020)	(0.011)	(0.018)	(0.014)	(0.017)	(0.012)	(0.010)	(0.012)
Unknown	-0.043***	-0.023***	-0.059***	-0.040***	-0.053***	-0.059***	-0.036***	-0.062***	-0.031***	-0.048***
	(0.009)	(0.009)	(0.014)	(0.008)	(0.012)	(0.010)	(0.012)	(0.009)	(0.008)	(0.011)
Black	0.027**	-0.075***	-0.050***	-0.038***	-0.023	-0.112***	-0.067***	-0.098***	-0.013	-0.036**
	(0.012)	(0.013)	(0.017)	(0.013)	(0.020)	(0.014)	(0.024)	(0.014)	(0.012)	(0.014)
Indian	0.132***	0.047***	0.047***	0.051***	0.055***	-0.069***	0.014	0.016	0.047***	-0.002
	(0.011)	(0.011)	(0.016)	(0.011)	(0.016)	(0.019)	(0.021)	(0.017)	(0.014)	(0.019)
Pakistani	0.105***	0.007	-0.067***	0.023*	0.050**	-0.151***	0.003	-0.078***	-0.005	-0.040**
	(0.013)	(0.013)	(0.022)	(0.014)	(0.024)	(0.019)	(0.028)	(0.019)	(0.014)	(0.018)
Bangladeshi	0.123***	0.005	-0.075***	0.050***	0.082**	-0.127***	-0.044	-0.033	-0.012	-0.026
	(0.018)	(0.016)	(0.027)	(0.019)	(0.032)	(0.024)	(0.037)	(0.024)	(0.019)	(0.021)
Chinese	0.135***	0.037*	0.059*	0.031*	0.079***	-0.001	0.050	0.021	0.067**	0.067
	(0.023)	(0.021)	(0.031)	(0.017)	(0.025)	(0.043)	(0.044)	(0.046)	(0.033)	(0.053)
Other	0.106***	0.034***	-0.010	0.052***	0.055***	-0.038***	0.005	-0.013	0.025***	-0.002
	(0.009)	(0.009)	(0.014)	(0.008)	(0.012)	(0.011)	(0.014)	(0.011)	(0.009)	(0.012)
Achieved grade	0.501***	0.514***	0.462***	0.546***	0.544***	0.449***	0.509***	0.471***	0.482***	0.422***
	(0.003)	(0.004)	(0.004)	(0.003)	(0.004)	(0.003)	(0.004)	(0.003)	(0.003)	(0.004)
Constant	2.663***	2.705***	2.745***	2.499***	2.612***	2.696***	2.404***	2.608***	2.475***	2.551***
	(0.011)	(0.014)	(0.017)	(0.013)	(0.014)	(0.014)	(0.015)	(0.012)	(0.012)	(0.015)
N	96058	80459	39504	108477	52723	64096	51528	75221	98676	55620
R^2	0.48	0.49	0.34	0.54	0.53	0.36	0.42	0.36	0.40	0.31

Notes: The dependent variable is the predicted grade. Predicted and achieved grades are scored between E to A* and transformed to the following scores per grade: A*=6, A=5, B=4, C=3, D=2, E=1. Reference groups: boy, professional and managerial occupation, White ethnic background. The standard errors are clustered at the school-by-cohort level and reported in parentheses. Significance levels are denoted as: *** p<0.01, ** p<0.05, * p<0.10. The sample comprises university applicants between 2017-2019. *Source: GRADE 2017-2019*

Table A.4: Probability of under- and over- prediction

	(1) Biol.	(2) Chem.	(3) Econ.	(4) Maths	(5) Physics	(6) Engl. Lit	(7) Geogr.	(8) Hist.	(9) Psychol.	(10) Sociol.
Model 1: Probability of under-prediction										
Girl	-0.009***	-0.008***	-0.015***	-0.024***	-0.020***	-0.024***	-0.024***	-0.027***	-0.030***	-0.040***
	(0.002)	(0.002)	(0.004)	(0.002)	(0.002)	(0.003)	(0.003)	(0.002)	(0.002)	(0.004)
N	87191	74142	38051	102065	48253	61732	49896	72805	92679	53472
R^2	0.06	0.03	0.07	0.03	0.03	0.14	0.11	0.10	0.11	0.17
Mean dep. Var	0.062	0.051	0.097	0.079	0.048	0.098	0.096	0.081	0.091	0.146
Model 2: Probability of over-prediction										
Girl	0.019***	0.018***	0.046***	0.053***	0.055***	0.054***	0.049***	0.068***	0.056***	0.063***
	(0.003)	(0.003)	(0.005)	(0.003)	(0.005)	(0.005)	(0.004)	(0.004)	(0.004)	(0.005)
N	88728	74105	37041	90593	47664	59132	48614	71673	93727	52416
R^2	0.21	0.18	0.20	0.16	0.14	0.22	0.22	0.21	0.21	0.26
Mean dep. Var	0.720	0.709	0.616	0.650	0.738	0.639	0.586	0.627	0.640	0.552

Notes: The depended variable is the probability of under-prediction (model 1) and over-prediction (model 2). All regressions control for the achieved grade, socio-demographic controls and school-by-grade fixed effects. Predicted and achieved grades are scored between E to A* and transformed to the following scores per grade: A*=6, A=5, B=4, C=3, D=2, E=1. The standard errors are clustered at the school-by-cohort level and reported in parentheses. Significance levels are denoted as: *** p<0.01, ** p<0.05, * p<0.10. The sample comprises university applicants between 2017-2019. Applicants who achieved grade E are excluded in the under-prediction model and applicants who achieved grade A* are excluded in the over-prediction model. *Source: GRADE 2017-2019*

Table A.5: Robustness checks on measurement

		(1) Sec.order pol.	(2) Grade FE	(3) Total marks	(4) IV-G11	(5) IV-G6
Biology	Girl	0.062***	0.062***	0.067***	0.067***	0.067***
(n=96058)		(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
	R^2	$R^2$0.48	0.48	0.50	0.46	0.46
Chemistry	Girl	0.053***	0.052***	0.055***	0.082***	0.089***
(n=80459)		(0.006)	(0.006)	(0.006)	(0.006)	(0.007)
	R^2	0.49	0.49	0.51	0.47	0.47
Economics	Girl	0.107***	0.107***	0.107***	0.099***	0.098***
(n=39504)		(0.009)	(0.009)	(0.009)	(0.012)	(0.012)
	R^2	0.34	0.34	0.36	0.35	0.35
Maths	Girl	0.108***	0.107***	0.109***	0.141***	0.149***
(n=108477)		(0.005)	(0.005)	(0.005)	(0.006)	(0.006)
	R^2	0.54	0.54	0.56	0.52	0.52
Physics	Girl	0.163***	0.162***	0.167***	0.180***	0.181***
(n=52723)		(0.009)	(0.009)	(0.009)	(0.010)	(0.010)
	R^2	0.53	0.53	0.55	0.50	0.50
English Lit	Girl	0.102***	0.102***	0.104***	0.095***	0.095***
(n=64096)		(0.008)	(0.008)	(0.008)	(0.010)	(0.009)
	R^2	0.36	0.36	0.37	0.38	0.38
Geography	Girl	0.106***	0.104***	0.095***	0.024***	0.015*
(n=51528)		(0.007)	(0.007)	(0.007)	(0.008)	(0.009)
	R^2	0.42	0.43	0.46	0.43	0.43
History	Girl	0.149***	0.148***	0.151***	0.133***	0.132***
(n=75221)		(0.006)	(0.006)	(0.006)	(0.007)	(0.007)
	R^2	0.36	0.37	0.40	0.38	0.38
Psychology	Girl	0.136***	0.135***	0.123***	0.025***	0.016**
(n=98676)		(0.006)	(0.006)	(0.006)	(0.007)	(0.008)
	R^2	0.40	0.40	0.42		
Sociology	Girl	0.142***	0.141***	0.138***	0.051***	0.047***
(n=55620)		(0.009)	(0.009)	(0.009)	(0.010)	(0.011)
	R^2	0.32	0.32	0.33	0.37	0.37

Notes: The depended variable is the is the predicted grade. All regressions control for the achieved grade, socio-demographic controls and school-by-grade fixed effects. Predicted and achieved grades are scored between E to A* and transformed to the following scores per grade: A*=6, A=5, B=4, C=3, D=2, E=1. Column 1 includes the square of the achieved grade. Column 2 controls for the achieved grade more flexibly as a fixed effect. Column 3 controls for the total achieved marks instead of the achieved grade. Columns 4 and 5 instrument achieved grade using lagged and twice-lagged scores from grade 11 and grade 6 in the relevant subject (maths for STEM and English for non-STEM). The standard errors are clustered at the school-by-cohort level and reported in parentheses. Significance levels are denoted as: *** p<0.01, ** p<0.05, * p<0.10. The sample comprises university applicants between 2017-2019. *Source: GRADE 2017-2019*

Table A.6: Full specification using twice lagged skills as IV for achieved grades

	(1) Biol.	(2) Chem.	(3) Econ.	(4) Maths	(5) Physics	(6) Engl. Lit	(7) Geogr.	(8) Hist.	(9) Psychol.	(10) Sociol.
Achieved grade	0.682***	0.760***	0.995***	0.821***	0.727***	0.822***	0.878***	0.900***	0.819***	0.826***
	(0.007)	(0.009)	(0.024)	(0.009)	(0.010)	(0.009)	(0.012)	(0.010)	(0.010)	(0.014)
Girl	0.067***	0.089***	0.098***	0.149***	0.181***	0.095***	0.015*	0.132***	0.016**	0.047***
	(0.006)	(0.007)	(0.012)	(0.006)	(0.010)	(0.009)	(0.009)	(0.007)	(0.008)	(0.011)
Intermediate	0.017**	0.020**	-0.009	0.017**	0.026**	0.004	0.032***	-0.005	0.013	0.017
	(0.008)	(0.009)	(0.016)	(0.008)	(0.011)	(0.010)	(0.012)	(0.010)	(0.008)	(0.012)
Small employers	0.013	0.017	0.046**	0.039***	0.023	0.006	0.026*	0.014	0.024**	-0.005
	(0.009)	(0.011)	(0.020)	(0.011)	(0.015)	(0.014)	(0.015)	(0.013)	(0.010)	(0.014)
Lower supervisory	0.020*	0.033***	0.055**	0.035***	0.023	-0.012	0.058***	0.044***	0.028**	0.021
	(0.011)	(0.012)	(0.024)	(0.012)	(0.016)	(0.015)	(0.018)	(0.015)	(0.011)	(0.016)
Semi-routine	0.028***	0.019*	0.016	0.055***	0.039**	-0.009	0.035**	0.006	0.042***	0.023*
	(0.010)	(0.012)	(0.020)	(0.011)	(0.016)	(0.013)	(0.018)	(0.013)	(0.010)	(0.014)
Routine	0.014	0.029**	0.056**	0.042***	0.016	-0.007	0.019	0.017	0.051***	0.031**
	(0.012)	(0.014)	(0.025)	(0.013)	(0.019)	(0.016)	(0.020)	(0.015)	(0.011)	(0.015)
Unknown	-0.000	0.044***	0.013	0.023**	-0.014	0.006	0.025*	0.001	0.034***	0.008
	(0.009)	(0.010)	(0.018)	(0.010)	(0.013)	(0.012)	(0.015)	(0.011)	(0.010)	(0.013)
Black	0.096***	0.019	0.170***	0.085***	0.086***	0.036**	0.147***	0.045**	0.074***	0.045***
	(0.013)	(0.014)	(0.023)	(0.015)	(0.022)	(0.017)	(0.029)	(0.018)	(0.013)	(0.017)
Indian	0.142***	0.080***	0.116***	0.077***	0.095***	0.003	0.074***	0.049**	0.068***	0.036*
	(0.012)	(0.012)	(0.021)	(0.012)	(0.017)	(0.022)	(0.024)	(0.020)	(0.015)	(0.022)
Pakistani	0.157***	0.105***	0.098***	0.123***	0.134***	-0.016	0.093***	0.034	0.053***	0.024
	(0.014)	(0.015)	(0.027)	(0.016)	(0.026)	(0.022)	(0.033)	(0.023)	(0.015)	(0.021)
Bangladeshi	0.157***	0.069***	0.084**	0.136***	0.158***	-0.015	0.026	0.063**	0.023	0.014
	(0.019)	(0.018)	(0.035)	(0.021)	(0.034)	(0.028)	(0.044)	(0.027)	(0.022)	(0.026)
Chinese	0.097***	-0.017	0.015	-0.017	0.068**	0.019	0.035	0.042	0.028	-0.000
	(0.024)	(0.023)	(0.041)	(0.020)	(0.026)	(0.049)	(0.051)	(0.053)	(0.037)	(0.060)
Other	0.125***	0.076***	0.105***	0.093***	0.093***	0.011	0.058***	0.032**	0.082***	0.046***
	(0.009)	(0.010)	(0.018)	(0.009)	(0.013)	(0.012)	(0.016)	(0.013)	(0.010)	(0.015)
N	96058	80459	39504	108477	52723	64096	51528	75221	98676	55620
R^2	0.46	0.47	0.35	0.52	0.50	0.38	0.43	0.38	0.37	0.30

Notes: Full specification from the IV regressions using subject-specific grade 6 score as an instrument for achieved grade at A-level. The depended variable is the is the A-level predicted grade. All regressions include school-by-grade fixed effects. Predicted and achieved grades are scored between E to A* and transformed to the following scores per grade: A*=6, A=5, B=4, C=3, D=2, E=1. The standard errors are clustered at the school-by-cohort level and reported in parentheses. Significance levels are denoted as: *** p<0.01, ** p<0.05, * p<0.10. The sample comprises university applicants between 2017-2019. *Source: GRADE 2017-2019*