CEPEO Working Paper No. 25-12

Measuring systematic gaps in teacher judgement: A new approach

Oliver Cassagneau-Francis
UCL

Lindsey Macmillan
UCL

Richard Murphy
University of Texas at Austin

Gill Wyness
UCL

We propose a new approach to test for systematic biases in teacher evaluations. We exploit a setting where teachers were required to assign students both grades and rankings within each grade. Comparing students immediately adjacent to grade boundaries, we apply a local randomization approach to estimate imbalance in student characteristics. Our findings reveal systematic bias favoring higher income and female students. These grading decisions carry real consequences: students just above the grade threshold are significantly more likely to attend university. Our approach can be applied whenever there is a system with many thresholds and subjective rankings.

VERSION: October 2025

Suggested citation: Cassagneau-Francis, C., Macmillan, L., Murphy. R., & Wyness, G. (2025). *Measuring systematic gaps in teacher judgement: a new approach.* (CEPEO Working Paper No. 25-12). UCL Centre for Education Policy and Equalising Opportunities. https://EconPapers.repec.org/RePEc:ucl:cepeow:25-12.

Centre for Education Policy & Equalising Opportunities

Disclaimer

Any opinions expressed here are those of the author(s) and not those of UCL. Research published in this series may include views on policy, but the university itself takes no institutional policy positions.

CEPEO Working Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

Highlights

- Despite concerns about the validity of subjective academic evaluations, many countries – including Australia, the UK, the USA, and Portugal – are increasingly relying on these metrics.
- Proponents of teacher-assessed grades argue several advantages: they may reduce stress on students and teachers, allow for broader curricula that more adequately capture students' skills and personality, and are less prone to measurement error. On the other hand, they can be less informative of academic success at university, and they make it harder to draw comparisons between schools, or over time.
- Potentially the most important issue, however, is that the subjective nature of teacher assessments leaves the door open for teacher bias.
- In this paper, we propose a new method to test whether teachers favour some students (by gender, ethnicity, SES) over others when awarding grades.
- Our approach exploits the Covid-19 induced cancellation of A-level exams in the UK. In place of exams, teachers were required to assign students both grades and rankings within each grade. We use this ranking to test for systematic biases in teacher evaluations. We identify the most marginal students at each grade boundary, and implement a simple test comparing the share of student types immediately above/below the boundary – e.g. comparing the top-ranked students of one grade with the lowest-ranked student of the next grade.
- We find evidence of discontinuities in favour of female students, and against those receiving Free-School Meals, suggesting that teachers exhibit bias in these directions. We show that these biases have direct consequences for being accepted at university, and of being accepted to ones first choice course.
- Our results imply that governments considering increasing reliance on these assessments should take steps to mitigate this bias.

Why does this matter?

It is essential that policymakers considering abolishing or reducing/diminishing exams understand the limits of alternative approaches

Measuring systematic gaps in teacher judgement: A new approach

Oliver Cassagneau-Francis* Richard Murphy† Lindsey Macmillan*
Gill Wyness‡

13th October 2025

Abstract

We propose a new approach to test for systematic biases in teacher evaluations. We exploit a setting where teachers were required to assign students both grades and rankings within each grade. Comparing students immediately adjacent to grade boundaries, we apply a local randomization approach to estimate imbalance in student characteristics. Our findings reveal systematic bias favoring higher income and female students. These grading decisions carry real consequences: students just above the grade threshold are significantly more likely to attend university. Our approach can be applied whenever there is a system with many thresholds and subjective rankings.

Keywords: Teacher bias; Gender; Stereotypes; Proportions; Test Optional
JEL codes: C10, C25, I23, I24, J15

*Centre for Education Policy and Equalising Opportunities (CEPEO), UCL.
†UT Austin, NBER, IZA, CESifo
‡CEPEO and CEP, LSE
We also thank Ofqual, DfE and UCAS for making this valuable dataset available to researchers. This work was generously supported by an ADR UK Research Fellowship. The usual disclaimers apply.

1 Introduction

Despite concerns about the validity of subjective academic evaluations, many countries – including Australia, the UK, the USA, and Portugal – are increasingly relying on these metrics.[1] This shift often involves moving away from standardized test scores in favour of teacher assessments, such as GPA, particularly in the university application process (Dessein et al., 2025).[2] Proponents of subjective teacher-assessed grades argue several advantages: they may reduce stress on students and teachers (Holbein and Ladd, 2017), allow for broader curricula that more adequately capture students' skills and personality (Kautz et al., 2014), and are less prone to measurement error (Rimfeld et al., 2019). Yet evidence also suggests that these non-blind metrics may have undesirable features: they can be less informative of academic success at university, and they make it harder to draw comparisons between schools, or over time (Chetty et al., 2023; Friedman et al., 2025; Goodman, 2016). Potentially the most important issue, however, is that the subjective nature of teacher assessments leaves the door open for teacher bias (Lavy, 2008; Lavy and Sand, 2015; Carlana, 2019; Terrier, 2020; Avitzour et al., 2020; Burgess et al., 2022; Cassagneau-Francis and Wyness, 2025). This has the potential to reinforce stereotypes and exacerbate existing achievement inequalities, subsequently impacting students' investment in their own human capital.

The main challenges faced by studies measuring teacher bias is that they either require primary data collection, or rely on assumptions to compare achievements on blind and non-blind assessments (Hinnerich et al., 2011; Hanna and Linden, 2012; Alesina et al., 2018; Carlana, 2019; Lavy, 2008; Lavy and Sand, 2015; Terrier, 2020; Burgess et al., 2022; Graetz and Karimi, 2022; Lavy and Megalokonomou, 2024). In this paper, we develop a novel approach that does not require new data collection, can be applied in many settings including outside of education,[3] and which does not require blind assessments or the standard assumptions of the literature.

Our approach exploits the Covid-19 induced cancellation of high-stakes, standardized, end of secondary education (A-level) exams in the UK, and the nature of the teacher-assessment which replaced them. Instead of sitting (externally graded) exams, students were assigned a letter grade by their teachers in each A-level subject they were taking (typically three).[4] In addition to assigning grades to students, teachers were also asked to

[1] Subjective evaluations play a pivotal role in many high-stakes decisions, including hiring and promotion in the labor market (Li and Kahn, 2018; Taylor and Yildirim, 2011), bail and sentencing in the justice system (Chyn et al., 2025; Ichino et al., 2003; Arnold et al., 2018; Cohen and Yang, 2019), and assessments of pupil performance in education (Burgess et al., 2022; Lavy, 2008).

[2] See Leonhardt (2024) for a discussion of the SAT in the US, and Portuguese Ministry of Education (2023) for Portugal, and Moss et al. (2021) for the UK.

[3] Our method can be implemented in any setting with endogenous thresholds and subjective scores or rankings, for example in managerial performance metrics, job interviews and subjective competitions.

[4] These teacher-assessed grades were officially known as "Centre Assessed Grades" or CAGs and were

rank students within each letter grade. This provides a ranking of every A-level student by subject and grade in each secondary school.[5] We use this ranking to test for systematic biases in teacher evaluations. We identify the most marginal students at each grade boundary, and implement a simple test comparing the share of student types immediately above/below the boundary – e.g. comparing the top-ranked students of one grade with the lowest-ranked student of the next grade. The intuition is that we should not expect to see a jump in students of one characteristic (e.g. female students) at the bottom of the A grade, versus the top of the B grade – such a jump would be evidence of bias. This parsimonious approach is an application of local randomization (LR, Cattaneo et al., 2024) – a form of regression discontinuity with discrete mass points. In our setting, we do not require projecting outcomes to the cutoff, as the adjacent students are by definition the most marginal students. We have access to a unique, detailed administrative dataset (the "GRADE" dataset), which contains these subjective rankings alongside student demographics, for all state students in England (140,000), providing sufficient statistical power.

Our new approach makes several contributions to the existing literature measuring bias. Studies of bias fall into two camps: those measuring bias directly and those using indirect approaches. Studies that use the direct approach either use field experiments where the same exam scripts are graded blindly and non-blindly (Hinnerich et al., 2011), or randomly assign characteristics (Hanna and Linden, 2012), or implement Implicit Association Tests (IAT) (Alesina et al., 2018; Carlana, 2019). These approaches have the advantage of measuring bias directly, but they involve collecting primary data, which results in limited populations, limiting the external validity of the estimates, and making scalability challenging. For example, if a government wanted to estimate teacher bias across their country, they would have to perform tests on every teacher. Ours is a generalizable, direct measure of bias that does not require specific survey instruments — indeed, it can be implemented in any setting with endogenous grade boundaries and underlying test scores, or ranking of students.

The indirect approach to measuring teacher bias was developed by Lavy (2008), and has now become the convention (Lavy and Megalokonomou, 2024). In this approach, researchers compare gaps (e.g. male versus female) in student performance from (non-blind) teacher assessment, to those from (blind) external assessment. The key assumption

originally going to be adjusted by an algorithm to account for differences across schools, combating grade inflation and moderating teacher predictions (House of Commons Education Committee, 2020b). The algorithm did what it was designed to do, lowering grades below the teacher predictions. However, students were not happy to see their grade reduced in this way. After a national outcry the algorithm was scrapped and students were awarded the teacher-assessed CAG as their final grade (House of Commons Education Committee, 2020a).

[5]This ranking was initially going to be fed into the algorithm to adjust CAGs, and there was detailed guidance given to teachers both on how to assign grades and how to rank teachers (ofqual, 2020).

is that any observed differences in these gaps must be due to teacher bias. These papers have generally found teachers favour female over male students, and white students over non-white students (Lavy, 2008; Lavy and Sand, 2015; Terrier, 2020; Burgess et al., 2022; Graetz and Karimi, 2022; Lavy and Megalokonomou, 2024).

This approach has many advantages, including its ease of implementation. However, it requires a specific setting and assumptions that are not required by our method. To implement, it is necessary that pupils sit the same or similar tests, both in a blind and non-blind setting, at similar ages, which is rare. A second issue concerns the assumption that both blind and non-blind assessments are measuring the same ability. But if, as is often the case, blind exams are written and non-blind exams are oral this could be measuring different types of ability (Hirnstein et al., 2023). A third issue concerns the assumption that blind and non-blind assessments have the same mapping of ability to assessment performance by student type. In other words, that blind and non-blind tests would generate the same gaps in performance in the absence of teacher discrimination. Yet there is evidence that females perform worse in the high-pressure environment of standardized assessments versus in-class tests, (Cai et al., 2019; Galasso and Profeta, 2024; Arenas and Calsamiglia, 2025) suggesting this assumption could be problematic. Finally, this approach requires that blind and non-blind assessments have the same measurement error for each characteristic. Yet if females have higher average ability than males, this would make it less likely that high achievement among females in blind tests would be driven by measurement error (Zhu, 2024; Delaney and Devereux, 2025). A key contribution of our method is that it only requires a non-blind assessment and so does not require these assumptions relating two forms of assessment. Our findings of bias in favour of female students, and against low SES students are in line with the teacher bias literature (Lavy, 2008; Burgess and Greaves, 2013).

A further contribution is that our paper allows, for the first time, estimates of the heterogeneity of teacher bias by student characteristics, student achievement, and subject of study. While many papers have uncovered heterogeneity in one or two of these elements, no single study has measured bias across such a wide array of dimensions. Overall, our results provide new evidence of biases in how teachers assign high-stakes grades. These findings imply that education policymakers should take caution when considering adopting teacher assessment in place of external exams, particularly in high-stakes settings.

We are also able to provide evidence that teacher bias is more prevalent in subjects where teachers have more 'discretion'. We show that bias is lower in subjects where prior attainment is more informative of future performance. The patterns of the heterogeneity observed are not consistent with simple models of taste or statistical discrimination, implying that teachers' actions are complex and there are multiple mechanisms at work.

To carry out our analysis, we obtained access to GRADE[6] data, a unique administrative dataset which contains these subjective rankings alongside student data including demographics, prior achievement, and applications to college for all students in England – allowing us to estimate bias across different student characteristics, and to establish the consequences of this bias.

We find evidence of discontinuities in favour of female students, and against those receiving Free-School Meals (FSM, an indicator of socio-economic disadvantage), suggesting that teachers exhibit bias in these directions. These findings are in line with work by Lavy (2008), Burgess and Greaves (2013), and others which rely on a stronger set of assumptions. We show that the extent of this bias varies by grade boundary, across student types. For example, teachers are biased against FSM students except at the highest grade boundary, where we find pro-FSM bias, while the largest (pro-female) gender gaps occur at the C/D grade boundary, which is generally considered the pass/fail threshold in these exams. This finding implies that teachers are aware of the implications of their decisions on student outcomes. Importantly, we only find these discontinuities in the characteristics of adjacent students at grade boundaries, while away from these boundaries, we see no such discontinuities (see Figure 4).

We show that these biases have direct consequences: being just over a grade boundary threshold increases the chance of being accepted at university by 3.5 percentage points, and of being accepted to ones first choice course by 2.3 percentage points. This is particularly important at the top of the grade distribution, with those just above the threshold 6.6 percentage points more likely to be accepted to their first choice course than those just below the boundary.

The standard assumption behind the local randomization approach is that individuals are as good as randomly assigned each side of a threshold. However, we have shown that individuals are not randomly allocated across thresholds since teachers systematically favor some students over others. Therefore we rely on a weaker assumption, that *potential* outcomes are as good as random across a threshold. A threat to this assumption would be if there is a relationship between student attainment and student characteristics that is driving our results.[7] We test that our results are not driven by this type of threat in three ways: first, we condition on various measures of prior achievement; second, we create a cardinal measure of latent achievement and restrict analysis to adjacent students with the same latent achievement; and third we use a less restrictive approach that only exploits crowding around certain boundaries. In all cases we find no evidence that our results are driven by achievement gradients.

[6]The Grading and Admissions Dataset for England (Ofqual et al., 2025).

[7]For example female students generally perform better than male students at A-level, (EPI, 2021).

The rest of this paper proceeds as follows. In section 2 we describe key aspects of the UK education system and the context surrounding the cancellation of exams in 2020. Section 3 describes our dataset and empirical approach in more detail. We present and discuss our results in section 4, and demonstrate the robustness of our results in section 5. Section 6 concludes.

2 Background and context

In this section we describe the institutional context of exams and higher education admissions in England, and then describe COVID-19 induced exam cancellations and their replacement with teacher-assessed grades.

2.1 Institutional context

High school students in the UK take a set of compulsory standardised exams at age sixteen called GCSEs, and then some continue on an academic track to study a further set of exams known as A-levels. These are the main qualification for admittance to university programs.[8] Students typically take around ten GCSE subjects, and then specialize in studying towards three A-level subjects.

Figure 1 shows the timeline of educational decisions that students who wish to attend university generally face. At the beginning of their final year of high school students apply to university courses and during the spring they receive "conditional" offers.[9] These conditional offers prescribe the results they should achieve in their A-levels to confirm their place. In a standard academic year (figure 1a), students then take their A-level examinations in May-June of their final year, receive their results in August and based on this performance start at a university in October. Note that in the English system, students apply to university-major combinations (known as "courses"), such as economics at LSE, or mathematics at UCL.

2.2 Changes due to the COVID-19 pandemic

The COVID-19 pandemic meant that the majority of schools closed in March 2020, and they did not reopen for exams. In June it was decided that all standardised exams in

[8]Nearly 80% of university students (Cavaglia et al., 2024) use this academic track. The next popular approach is a more vocational track taken at further education (FE) college. Although many students do attend university with vocational qualifications, the traditional, and most popular route to university is by studying A-levels in high school or sixth form college, a route taken by . We focus on the set of students who take A-levels in this paper.

[9]The English system is relatively unusual by global standards in that students apply to university, and receive offers, before they have taken their entry exams. Instead, they apply on the basis of predicted grades provided by their teachers. For more details see Murphy and Wyness (2020).

Figure 1: Timeline of educational decisions

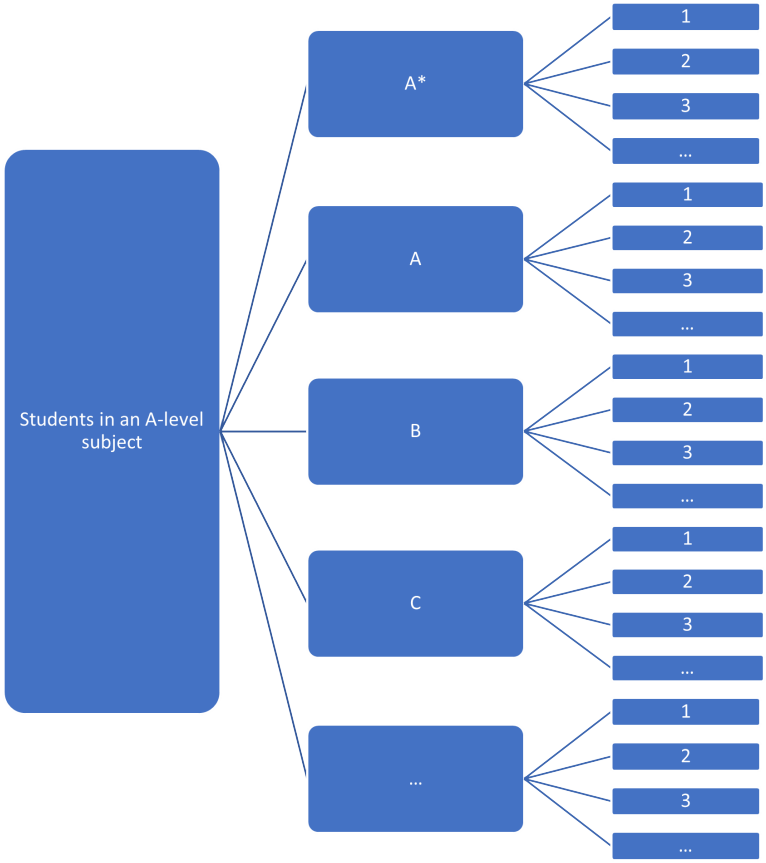
(a) Normal timeline

End of year 11	Sit GCSEs
Year 13 (age 17/18)	Apply to universities & receive "conditional" offers
June of year 13	Take A level exams
August after year 13	Receive A level results & confirm place at university

(b) COVID-19 pandemic affected timeline in 2020

End of year 11	Sit GCSEs
Year 13 (age 17/18)	Apply to universities
March 2020	COVID-19 pandemic - all schools close
June 2020	A-level exams cancelled
August 2020	A-level results based on CAGs & confirm place at university

Figure 2: Ranking of students within school, subject and grade



the UK would be canceled, including A-levels (See figure 1b). Given the importance of A-level qualifications, it was decided to replace the exams with teacher assessed grades, so that students would obtain qualifications that were informative of their ability.

After a short consultation, the exam regulator Ofqual decided on a process to determine these grades. Ofqual's guidance was that "[e]xam boards will ask exam centres to generate, for each subject [...] Centre Assessment Grades (CAG) for their students [...] and then to rank order the students within each of those grades" (Ofqual, 2020, p. 5). These CAGs were decided on by the students' teachers. The guidance instructed teachers to grade students based on "the grade that each student is most likely to have achieved if they had sat their exams" (Ofqual, 2020, p. 5). Thus, every school was required to submit teacher-assessed grades and a rank order of their pupils within each grade, for every A-level subject the student was taking. This process is illustrated in Figure 2.

While teachers were not given a reason for why they need to provide rank information, the intention of the Department for Education (DfE) in requesting this rank information was to combat anticipated grade inflation as a result of using teacher assessment instead of externally marked exams. The DfE intended to use the rank information, along with student prior achievement, and previous school-subject achievement distributions to monotonic-

ally transform the teacher assessed grades. Indeed, the algorithm was implemented and did what it was supposed to do — generally lowered grades below the teacher-assessed CAGs. However, upon the release of the adjusted grades there was a national outcry that an algorithm had determined (and lowered) students' A-level grades and hence their future life-chances.[10] After three days, the DfE rescinded the augmented grades and instead students were awarded either their CAG, or the algorithm grade, depending on which was the highest of the two. This resulted in the vast majority of students receiving their CAG.

Although the student ranks within grade were rarely used, they are central to our estimation strategy, as we explain in our empirical approach below.

3 Data and empirical approach

Our aim is to exploit the ranking of students within grades discussed in the previous section as a means of identifying teacher bias. To this end, we use the linked-administrative GRADE dataset (Ofqual et al., 2025), which was set up to allow researchers to study the events of 2020. Specifically, we will examine the students around each school-subject-grade boundary, to look for changes in the concentration of student characteristics at either side of each boundary. Thus, we require information on the grades and rankings for each student, as well as their characteristics, and school attended, all of which is present in GRADE.

3.1 Dataset

GRADE contains pupil-level data on age 16 (GCSE) and age 18 (A-level) exams and qualifications data from the Office for Qualifications (Ofqual), linked to a rich set of pupil characteristics and background information from the Department for Education (DfE)'s National Pupil Database (NPD), in turn linked to data on each pupils' university applications (where applicable) from the University and College Admissions Service (UCAS).

The dataset contains both the GCSE and A-level scores of the students for the 2020 cohort and the prior 2019 cohort.[11] Crucially, the dataset contains the CAGs and teacher rankings awarded in 2020, for every student in each subject, grade and school. We restrict our analysis to the cohort who were subject to teacher assessment of their A-level grades (i.e. the 2020 cohort). As well as observing the grades and rankings for these students, we also have detailed data on their prior attainment at age 16 (i.e. their raw GCSE scores, which are used to derive the GCSE grades, which were unaffected by the pandemic).

[10] See the Wikipedia article "2020 United Kingdom school exam grading controversy" for more details.

[11] We use the 2019 cohort to create a typical mapping between GCSE and A-level achievement, to create latent achievement measure used in a robustness test.

As we are interested in inequalities in teacher generosity by pupil types, our variables of interest are the gender of the pupil (female/male), and their socio-economic status.[12] For socio-economic status we use information on the free school meals (FSM/non-FSM) status of the student. In the UK, around 7% of high school students are eligible for free school meals, and while FSM eligibility is not entirely determined by household income, it is a strong indicator of poverty.

Table 1 panel (a) presents population shares and numbers of students for each of our categories of interest. While there is complete data on the gender of students, information on students free school meal status (FSM) is incomplete. For consistency across models we include only those students with gender and FSM status in our analysis sample.[13] Around 55% of students are female, representing the greater tendency of female students to pursue A-levels.[14] Only 7% are FSM eligible. The shares of these characteristics in this subsample are the same as in the population.

In panel (b) of Table 1, we show how our sample restrictions reduce our sample size, focusing now on numbers of observations, where an observation is student $\times$ subject. The restrictions are as follows: (i) we start with all observations for students with information on their gender and FSM status; (ii) we remove students with no GCSE marks; (iii) we remove observations which are not ranked adjacent to a grade boundary (i.e. not ranked 1 or -1); (iv) we remove observations for which there is no observation on the other side of the grade boundary; (v) our weighting procedure means we drop any observations in homogeneous school-subject combinations with respect to our dependent variables, gender and FSM. Finally in panel (c), we present shares and numbers of observations in our analysis sample, broken down by subject and grade boundary.

3.2 Empirical strategy

As described above, in 2020 teachers determined each students' grade, and then ranked each student within their grade level, for every subject. Our empirical strategy leverages these rankings, testing for imbalances in the concentration of student types around grade thresholds.

The intuition behind this approach is straightforward: without bias the concentration of student characteristics around each grade boundary should be continuous; there should not be a jump in the proportion of students of a particular characteristic immediately

[12] We present our results for student ethnicity in Appendix B as these results do not pass some of our key robustness tests.

[13] Note that the UK has a small number of private (i.e. fee-paying) schools, who operate outside of the state sector. These schools are not required to report FSM or ethnicity, and so a large proportion of the missing data on FSM is from pupils in these private schools.

[14] e.g.https://ffteducationdatalab.org.uk/2021/09/which-a-level-subjects-have-the-best-and-worst-gender-balance/

Table 1: Descriptive statistics

(a) Numbers of students and population shares

Dependent variable (X):	Female	FSM eligible
Share	.55	.06
Number of students	237,037	210,013

(b) Number of observations and step-by-step sample selection

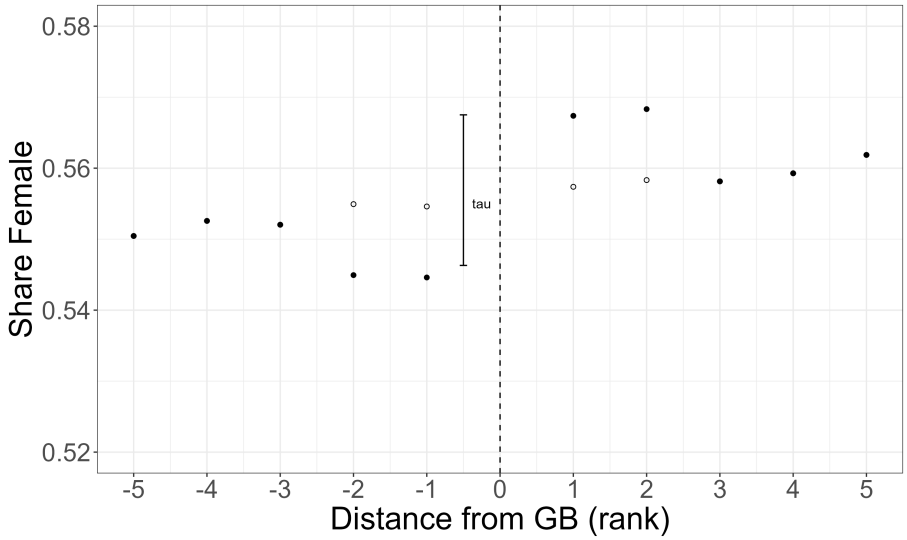
	No. of obs.	% female	% FSM
w/ info. on all X's	961,105	0.56	0.06
− missing GCSE	926,563	0.56	0.06
− ranked 2+ from GB	198,051	0.55	0.07
− unbalanced GBs	172,338	0.55	0.06
By X			
− single-sex cells	145,880	0.54	−
− all FSM/non-FSM cells	88,094	−	0.13

(c) Share and numbers of marginal observations by subject and grade boundary

	Female		FSM	
	N	Share	N	Share
By grade boundary				
A/A*	27,912	0.58	16556	0.09
B/A	37,962	0.57	22188	0.12
C/B	38,968	0.54	23038	0.14
D/C	28,364	0.49	17728	0.14
E/D	12,674	0.45	8584	0.14
By subject				
Biology	10,014	0.61	6,416	0.11
Business Studies	65,12	0.40	3,794	0.10
Chemistry	8,702	0.51	5,600	0.11
Economics	5,166	0.29	3,182	0.12
English Literature	8,122	0.74	5,690	0.13
History	9,148	0.55	5,306	0.12
Mathematics	11,150	0.37	7,904	0.09
Physics	7,376	0.22	4,166	0.11
Psychology	10,526	0.72	8,044	0.11
Sociology	7,694	0.74	6,014	0.13

Notes: Panel (a) shows the numbers of students in our "potential" samples, i.e. those students for whom we have information on their grades and on whether they are female or FSM eligible. We also present the shares of these characteristics. In panel (b) we show the numbers of resulting observations, i.e. subject×student, and the impact of the restrictions we impose to reach our final samples. Note we use the same "potential" sample for female as for FSM, dropping any students missing information on their FSM status. Panel (c) breaks these totals down by grade boundary and subject.

Figure 3: Stylized setup



Source: Authors construction.
Notes: The points represent the mean share of female students at a given rank. The hollow points show the "true" trend, whereas the filled spots show the trend if teachers are systematically boosting female students' grades.

around a threshold. If there were a discontinuity in the share of students of a certain type across a grade threshold that would imply that teachers are awarding some types of students more generous predictions than others. For example, if we observe a higher proportion of female students at the bottom of the A rankings, and a lower proportion at the top of the B rankings, this implies there may be systematic bias in favor of female students, with teachers pushing these students from a B to an A.

We illustrate this intuition in Figure 3, which shows (for a stylized example) the share of female students by distance from the grade boundary (where -1 equals the rank immediately below the grade threshold, and 1 equals the rank immediately above the grade threshold). Here, the unfilled series represents a case of no teacher bias. There is a continuous share of female students around the threshold, with no discontinuity in the share of females at the boundary. The filled series shows what we would see in the case of teacher manipulation — a discontinuity in the proportion of females ranked just above versus just below the grade cutoff. Our aim is to estimate τ — the difference in shares of student types of adjacent students either side of the boundary. Of course if there is a positive gradient in achievement by type (e.g. if females are generally higher achieving than males) some part of this τ will be due to these differences in ability between males and females. We will present multiple empirical tests in the robustness section to establish that the effect of ability is negligible around the threshold, once we control for prior attainment.

We operationalise our strategy by implementing a Local Randomisation (LR) approach (Li et al., 2021). This method employs the ranking as a discrete running variable, with

cutoffs at each grade boundary. We then test for differences in the proportion of student characteristics for students immediately each side of the boundary. This is akin to a standard RD "manipulation test". The core difference to a regression discontinuity approach is that by using the proportions of student types of students immediately adjacent to the boundary, we do not need to estimate the share at the cutoff, as the students are by definition the most marginal students. Indeed, projecting to the threshold, at rank zero, would provide incorrect estimates.

A key requirement of this test is to have sufficient data close to cutoffs to reliably estimate the shares. Due to our large sample sizes, we have over 70,000 populated school-subject-grade boundaries.[15] This allows us to use the smallest window $(-1, 1)$ around boundary b in subject s and in school j.

We use the subsample of students ranked immediately adjacent to a populated subject-grade boundary within their school (i.e. ranked first or last within a school-subject-grade cell), with a simple specification:

$$X_i = \beta_0 + \tau D_{ijsb} \ + \varepsilon_{ijsb} \tag{1}$$

Where X_i is an indicator of student i's characteristic (female, FSM), and D_{ijsb} is an indicator for student i, in school j and subject s being to the right hand side of boundary b. The parameter of interest is τ, which represents the percentage point difference in the share of a characteristic on one side of the boundary compared to the other.

Certain structural features limit our ability to detect bias. First, as shown in Table 1, there is a lack of diversity in post-16 education. For example, only 7% of our students are FSM, and fewer of these students are found at the top grades. There is also student sorting into schools — both through school segregation by socio-economic status, and through single-sex schools which appear in our sample. Finally, there is sorting by subject, with female students being more likely to choose humanities subjects at A-level, and male students more likely to choose maths and sciences. In these situations our measurement of bias will be attenuated. To combat this, we weight each school-subject-boundary observation by the variance of the share of the student type within the school-subject $W_{js} = \overline{X}_{js} \cdot (1 - \overline{X}_{js})$. This will down weight (or drop) cases where the school-subject-grade in question is disproportionately (or fully) one characteristic (e.g. in subjects dominated by female students, or in all-girls schools).

Local Randomisation treats units very close to the cutoff as if they were randomly assigned to be at one side or another, as in a natural experiment. However, in our case, we are

[15] A school-subject-grade boundary is populated if there is at least one student assigned to each adjacent grade in that school and subject.

looking for evidence of manipulation. Hence, by definition, we are looking for cases which are *not* as good as random. A key concern with our approach is that differential shares of student types each side of the boundary are not actually reflecting manipulation, but are simply reflecting differences in ability. For example, if female students are higher achieving, we would expect a marginally higher concentration of female students on the right hand side (RHS) of the boundary. In this case, τ would be picking up the underlying achievement gradient by student characteristic, in addition to any teacher manipulation. We primarily account for this by conditioning on prior achievement T, using the student's mean, standardized within subject, GCSE points:

$$X_i = \beta_0 + \tau D_{ijsb} \ + \beta_1 T_i + \varepsilon_{ijsb} \tag{2}$$

Conditioning on prior attainment should ensure that we achieve balance in potential outcomes around the threshold — e.g. that we have two students with the same probability of getting an A grade, but one of them gets assigned a B grade because of teacher bias. We will show in the robustness section that our results are robust to alternate measures of prior achievement and alternate approaches for achieving balance in potential outcomes around the threshold. As analysis of this sort requires large sample sizes, for our main analysis we combine the thresholds for all subjects and all grade boundaries together, for a total of 72,940 thresholds. We then examine discontinuities for each subject and grade boundary separately.

4 Results

In this section we first present local randomisation (LR) estimates averaged across all subjects and boundaries, then consider discontinuities by subject, and grade boundary separately.

4.1 Stacking subjects and grade boundaries

Table 2 presents weighted LR estimates averaged across all subjects and grade boundaries. Standard errors are clustered at the school level.

In columns 1 and 3, we show the raw average differences in proportion of each characteristic (female, FSM) from the lower (left hand side, henceforth LHS) side of the boundary to the upper (right hand side, RHS) side. In column 1, the coefficient of 0.032 implies that across all schools and all subject-grade boundaries, females are 3.2 percentage points (p.p.) more likely to be ranked on the RHS of a boundary compared to the LHS. While FSM students are 1.7p.p. under represented.

Table 2: Discontinuous change in student types at grade boundaries

Dependent variable (X_i):	Female		FSM	
	(1)	(2)	(3)	(4)
RHS of cutoff (τ)	0.031	0.022	−0.016	−0.009
	(0.003)	(0.003)	(0.003)	(0.003)
Constant	0.514	0.488	0.194	0.213
	(0.002)	(0.002)	(0.002)	(0.002)
T_i		✓		✓
N	145,880	145,880	88,094	88,094

Notes: This table present the results of our main specification, both raw (equation 1, odd columns) and conditional on prior achievement (equation 2, even columns) pooling observations across all subjects and grade boundaries. We estimate a weighted OLS weighting observations by how close the share of X is to 50% in that school and subject. Standard errors are in parentheses below the estimates.

In columns 2 and 4 we condition on prior achievement. The magnitude of the coefficients falls, but remains significant. Females are 2.3p.p. more likely to be ranked just above the threshold compared to below, and FSM students are 0.9p.p. less likely to be ranked just above. This reduction in magnitude is due to differential achievement being part of the unconditional τ (as discussed in Section 3). It is important to note that the share of A-level students who are FSM eligible is only 7%. Considering these differences in terms of percentage changes, then at the boundaries there are females are 4.2% more likely to be on the RHS, and FSM are 12.9% less likely to be on the RHS.

This simple conditioning on prior achievement is making a functional form assumption that the mapping between prior and current achievement is constant across genders, schools and subjects. We will relax this assumption in the robustness section.

Despite conditioning on prior achievement, a concern may remain that τ is still capturing characteristic-achievement gradients. To test for this directly we implement placebo tests, where we compare adjacent ranks that do not straddle a grade boundary. This test is displayed in figure 4. Here, the x-axis represents the rank-pairs distance from the grade-boundary, e.g. 0 represents the rank-pair that straddles the boundary — students ranked -1 and 1 relative to the cutoff — and 1.5 represents students ranked 1 and 2 above the cutoff, etc. Here, the τ at zero is our main estimate from Table 2. If there was a persistent positive characteristic-achievement gradient then estimates away from zero would be significantly and consistently positive.

For females we observe a distinct large positive estimate at zero, reflecting more 2.3 percentage point more females on the RHS. Critically however, we do not observe any significant discontinuities away from the grade boundary. That we only observe non-zero values at the threshold implies that our main estimates are not primarily driven by gender

achievement gradients. The second panel of figure 4 shows estimates for FSM students The estimate at zero is negative reflecting that FSM are less likely to be on the RHS. Reassuringly, the estimates for the change in concentration away from the threshold are all statistically indistinguishable from zero.[16]

4.2 Heterogeneity by grade boundary

Next, we present the extent of bias by grade boundary. These results are shown in Figure 5, which presents the change in the share female/FSM across grade boundaries, and at ranks away from the grade boundary.

We observe clear jumps in the proportion of female students at all of the grade boundaries, implying that teachers are pushing females over the line to a higher grade at every grade boundary. The extent of the bias is decreasing in the grade of the student, with the exception of the lowest boundary. The differences are largest at the C/D grade boundary with a 4.2p.p. gap in the proportion of females marginally achieving a C grade compared to a D grade. In contrast the difference at the A*/A grade boundary is only 1.7 percentage points. Again we see that at ranks away from the grade boundary there is no significant change in the share female conditional on prior achievement.

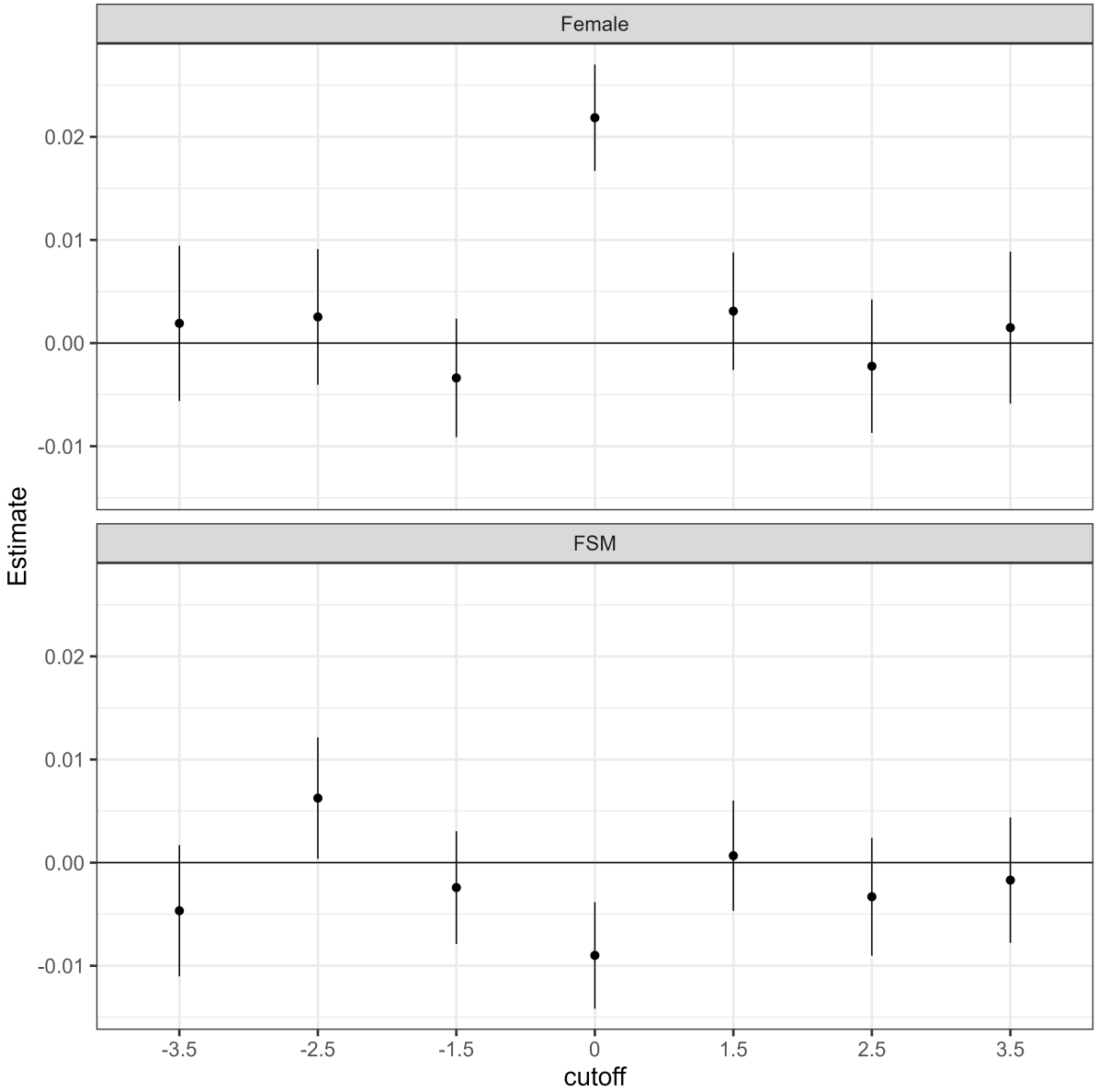
The pattern for FSM bias across boundaries is revealed to be quite different in Figure 5. From Figure 4 we saw on average teachers are biased against FSM students, now we observe that teachers are biased in favor of them at the highest grade boundary by 1.9 percentage points. For the remaining grade boundaries teachers are biased against FSM students, and become increasingly biased against them, with the largest bias against of 2.6p.p. at the lowest grade threshold (D/E). This distinct heterogeneity by grade boundary could be interpreted at teachers recognizing "diamonds in the rough". Note, the under-representation of FSM students on the RHS at the D/E boundary is made up for by their over-representations at the ranks just below the boundary.

4.3 Heterogeneity by subject

Estimates presented so far are an average across all subjects, but teachers bias may vary by subject (Breda and Hillion, 2016). Figure 6 shows there is considerable variation in

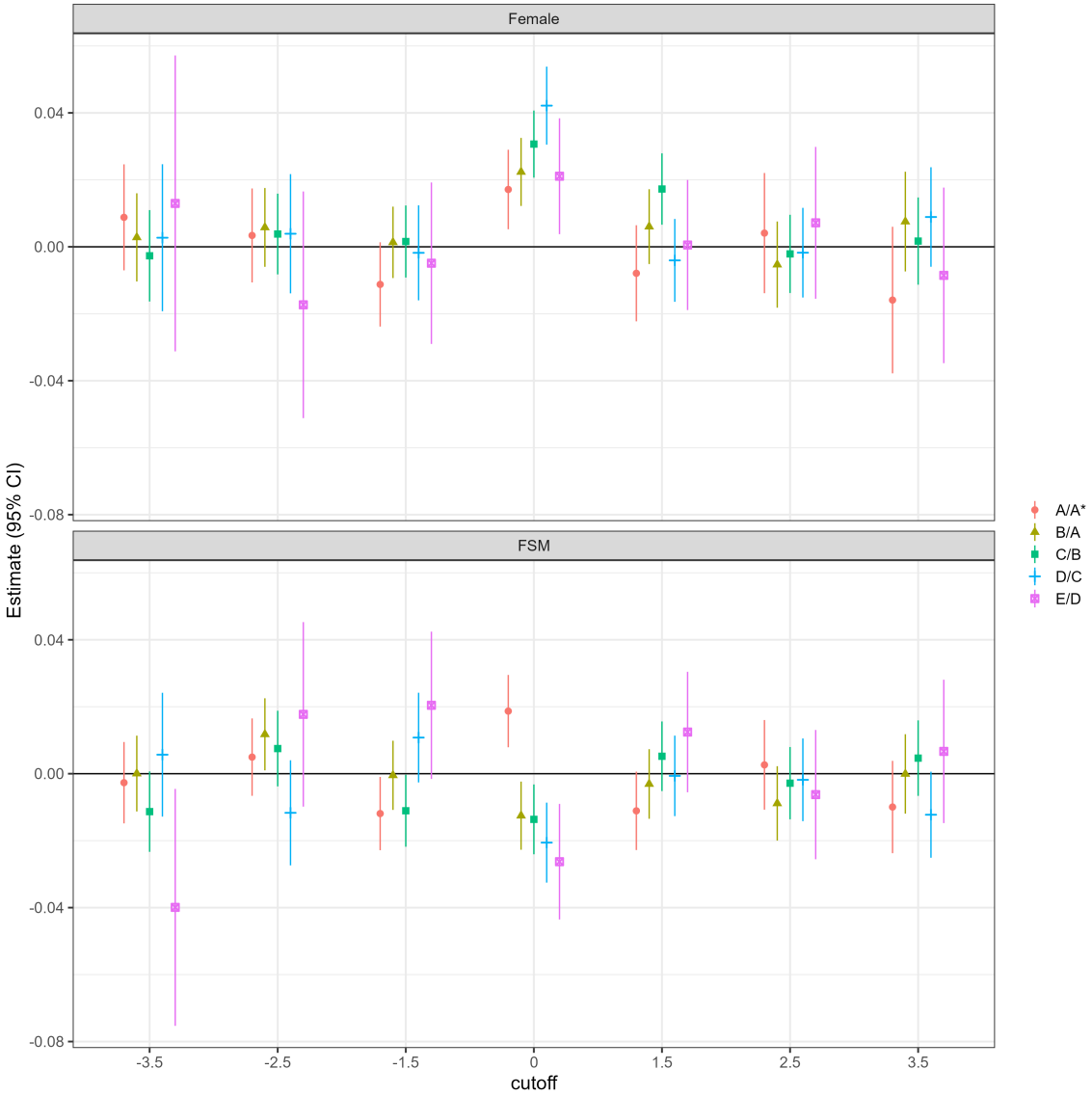
[16]In Appendix figure B.1 presents the equivalent placebo test for students having a White ethnicity. This implies that there are 1.2 percentage point more white students on the RHS than the left. However, the estimates away from the threshold are positive, two significantly. This may imply that τ even conditional on prior achievement is contaminated up the ethnicity achievement gradient e.g. white students have higher growth rates in achievement. However, it may also represent general discrimination against non-white students. We present evidence in Appendix B that it is likely due to the former. As our ethnicity estimates do not pass two of our three robustness tests, we do not present our estimates for this student type. Nevertheless, we take this as evidence that this test is informative in determining the validity of the approach in a setting.

Figure 4: Placebo tests by distance from Grade Boundary



Notes: This figure presents our main estimates from table 2, alongside "placebo" estimates for pairs of adjacently ranked observations that do not straddle grade boundaries. For example, while the point at 0 represents the estimated effect at a true grade boundary, the estimate at -1.5 represents the estimated effect if we place a "placebo" grade boundary between students ranked 1 and 2 within a grade. The extending bars represent 95% confidence intervals. All estimates are conditional on prior attainment and weighted by how close the share of X is to 50% within a school-subject.

Figure 5: Estimated bias (τ) by grade boundary (with placebo tests)



Notes: The coloured points show the estimated pro-female (a) / FSM (b) bias (τ) with different shapes (and colours) corresponding to different grade boundaries or "placebo" boundaries. The true grade boundary corresponds to 0 on the x-axis. The extending bars represent 95% confidence intervals. All estimates are conditional on prior attainment and weighted by how close the share of X is to 50% within a school-subject.

τ by A-level subject conditional on prior attainment. The bias in favor of females is the largest in the subjects where the teachers have more discretion – i.e. where marking is more subjective (art, law and English) compared to subjects that are more likely to be objectively marked (mathematics, chemistry, economics). Similarly the bias against FSM students is also largest in the subjects where teachers have more discretion. We discuss the implications of this in more depth in the next sub-section.

4.4 Discussion of mechanisms

This paper has documented systematic patterns in teacher generosity when assigning high-stakes grades. Teacher bias can operate through two distinct mechanisms – taste-based discrimination (Becker, 2010), where teachers favor students of a certain race, gender, or other characteristic, or statistical discrimination (Arrow, 1972; Phelps, 1972) where teachers have a noisy signal of student achievement and use stereotypes or group-level averages to decide marginal cases. Understanding which of these two mechanisms is responsible for the bias we observe is important for policy conclusions. If bias is taste-based, it suggests a need for bias training, monitoring and accountability mechanisms. Whereas, if bias is statistical, this would point towards improving information available to teachers about student ability (e.g. through better formative assessments).

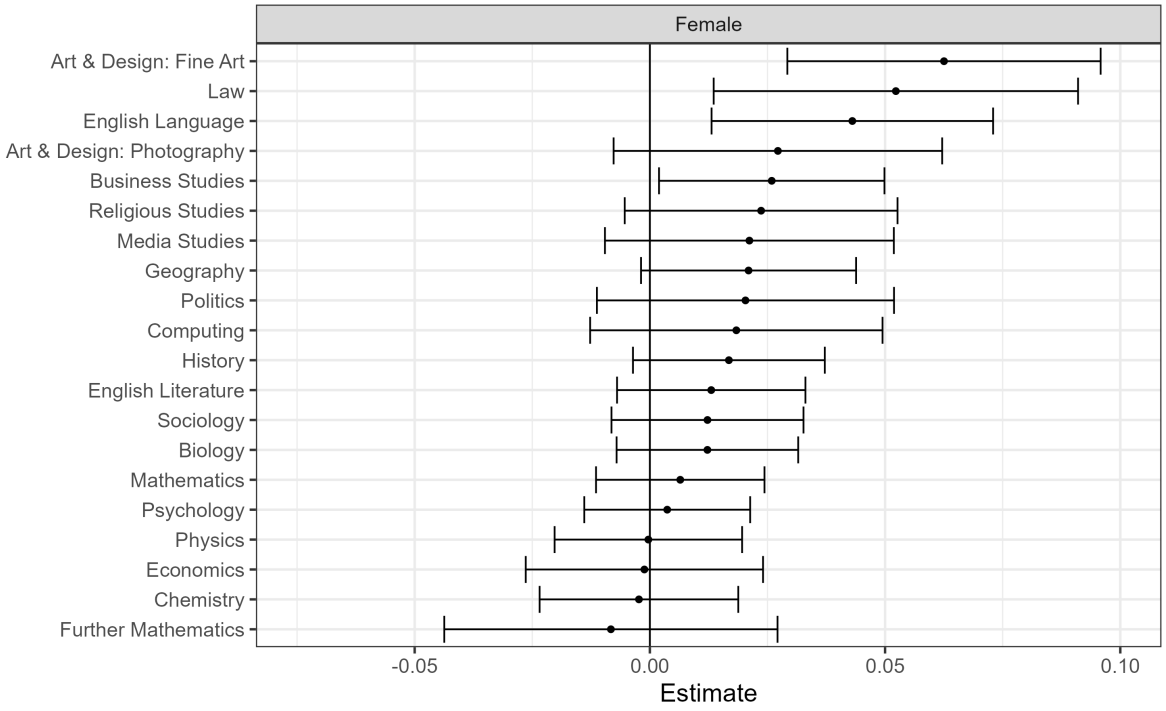
To date the literature exploring whether teacher bias can be explained by statistical or tasted based discrimination is mixed (Lavy, 2008; Burgess and Greaves, 2013; Botelho et al., 2015; Kisfalusi et al., 2021; Terrier, 2020). The heterogeneity of our estimates across subjects and grade boundaries can provide new insights to this discussion.

There is considerable variation in our estimates of τ across subjects for bias with respect to gender and FSM status (Section 4.3). The extent of this bias is greater in subjects where teachers have more discretion (English language, media studies) compared to where they have less (mathematics, physics). To objectively quantify the extent of discretion across subjects we estimate how informative prior measures of achievement are, for each subject. We estimate a simple bivariate regression of achievement at age 18 on achievement at age 16 and recover the R-squared parameter. For this exercise we use a previous cohort (2019 – the year before our main estimates) that was assessed externally, rather than by teachers. In figure 7 we plot the absolute value of τ estimates by subject against these R-squared measures (our rationale for using the absolute value of τ is that we are interested in the extent of bias, not the direction).

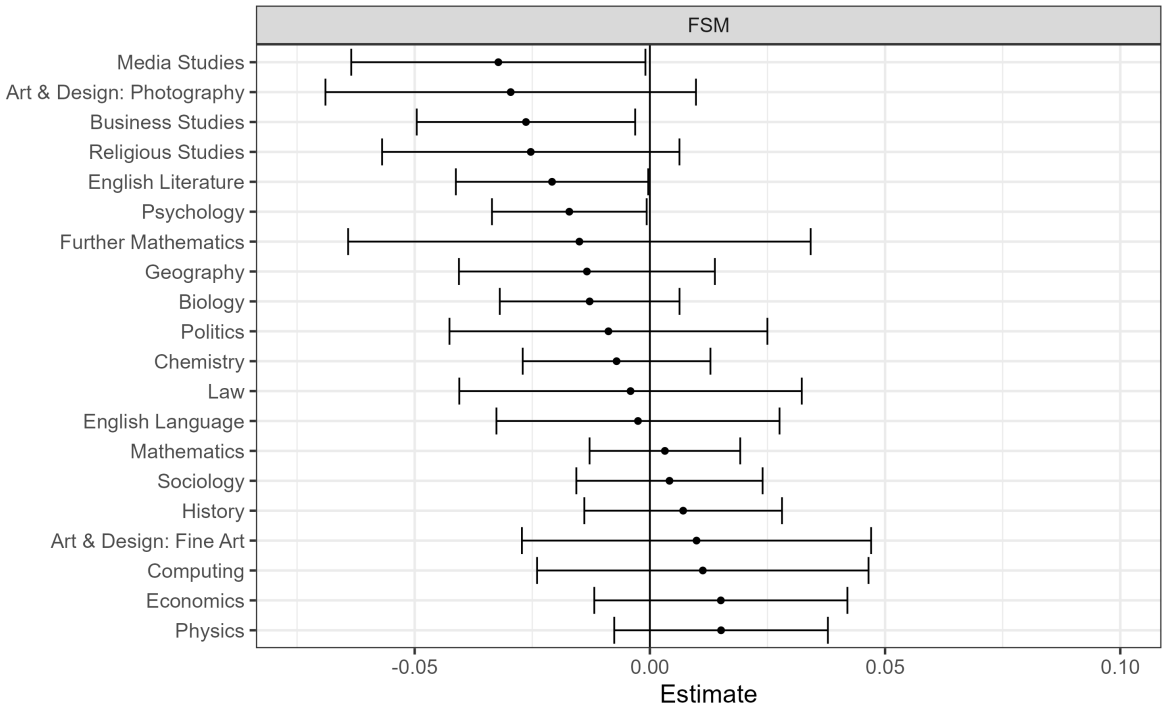
We observe a negative correlation across subjects between R-squared and absolute bias for both FSM and gender. In addition, the bias is in favor of the type that historically does well that subject (Appendix A.4). This is consistent with statistical discrimination. For subjects with low informativeness, teachers may rely more on stereotypes or

Figure 6: Stacking GBs by subject

(a) Female



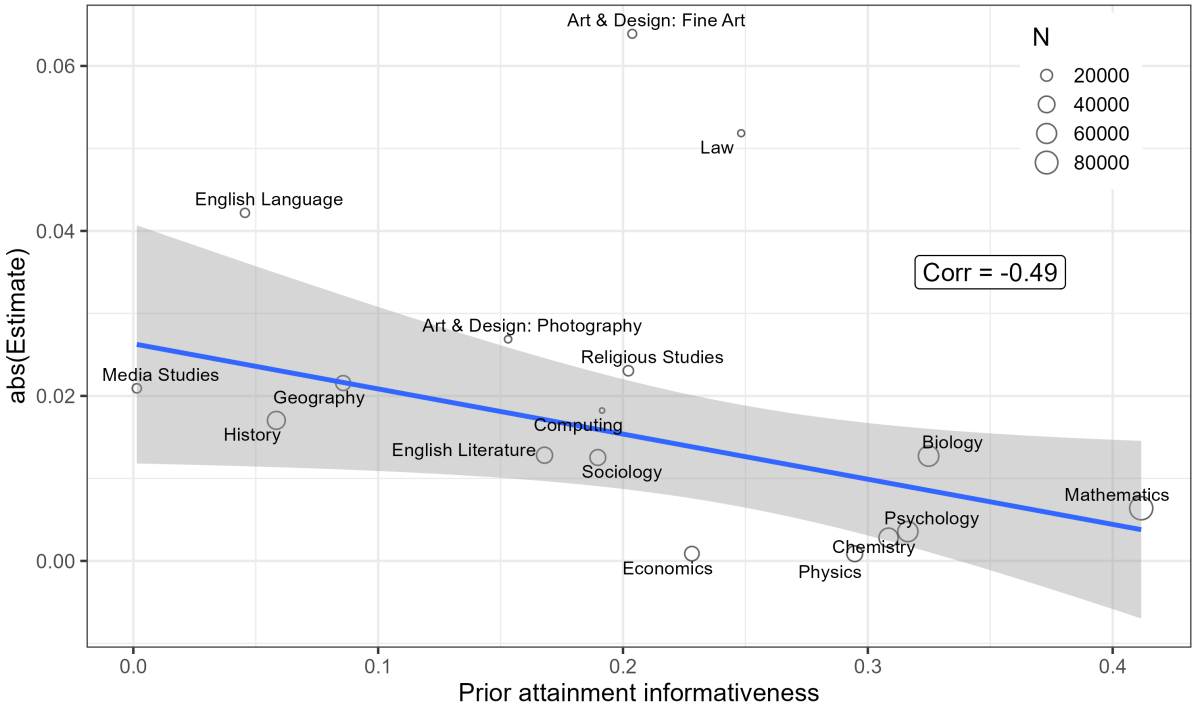
(b) FSM



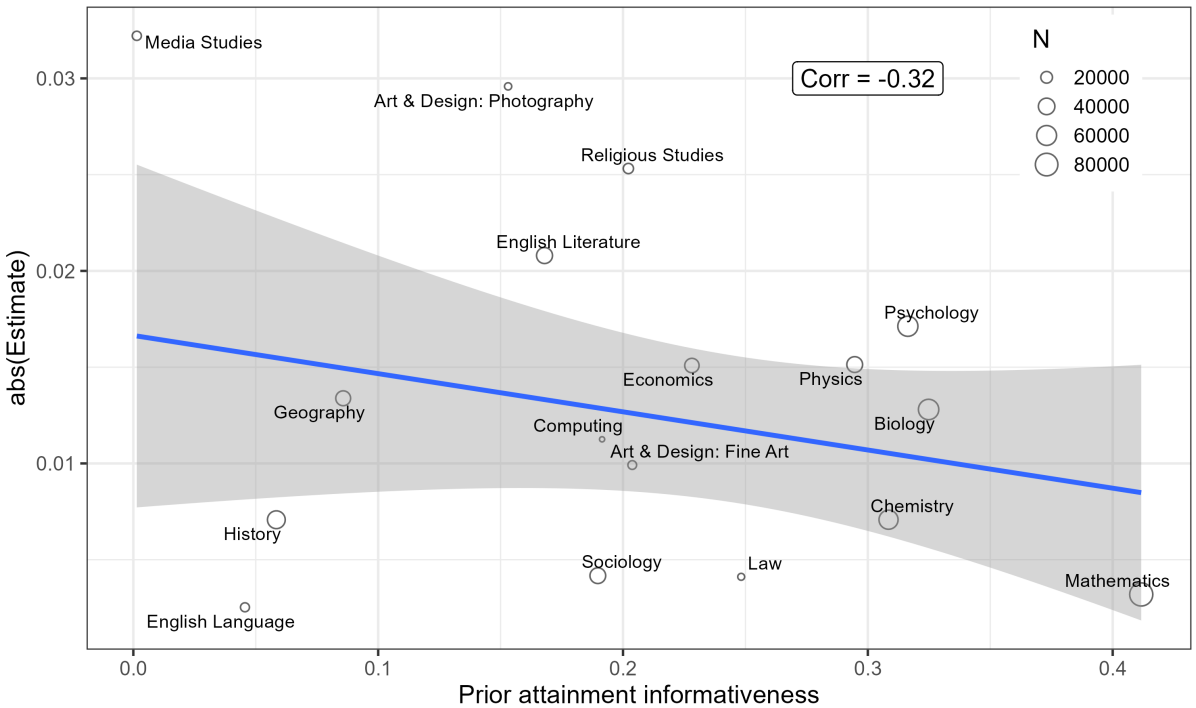
Notes: The points show the estimated pro-female bias (τ) for each of the top-20 subjects taken at A-level in 2020. The extending bars represent 95% confidence intervals. All estimates are conditional on prior attainment and weighted by how close the share of X is to 50% within a school-subject.

Figure 7: Bias estimates (τ) versus prior attainment informativeness by subject

(a) Female



(b) FSM



Notes: These panels show scatter plots of prior attainment informativeness on the absolute value of estimated bias by subject. We measure prior attainment informativeness as the R^2 from a regression of age 18 attainment on age 16 attainment for the previous year's cohort. The blue line shows the linear best fit with 95% CIs shaded. We also include the correlation as an inset. Panel (a) uses the estimated pro-female bias estimates, and panel (b) the FSM bias estimates.

group averages. Conversely, a key feature of simple taste-based discrimination is that it is unresponsive to information, and so one would expect zero correlation if taste-based discrimination.[17]. In addition to providing evidence distinguishing between the types of bias, this figure also provides evidence that τ is recovering the extent of teacher bias, rather than an ability gradient. The intuition is that, a priori there is no reason to believe that the extent of ability bias would vary systematically by R-squared.

In contrast to the variation across subjects being consistent with statistical discrimination, the variation across grade boundaries (Section 4.2) is not consistent with either basic model. For FSM students we observe a positive bias for the highest achieving students at the A*/A boundary, but increasingly negative for the rest. This runs contrary to what we should expect from statistical discrimination. On average FSM students are lower achieving than non-FSM students. This means that if teachers were applying a simple statistical discrimination model we would expect bias against FSM students everywhere. The fact that the bias is positive at the highest grade boundary and then gets increasingly negative as grade boundary decreases would require teachers to have a much more reliable measure of FSM student ability than of non-FSM students. These results are also not consistent with a simple taste-based model, which would have a constant level of bias regardless of grade boundary.

For female students we observe the positive bias is largest in middle of the distribution at the C/D boundary. This is also not consistent with statistical discrimination regardless of teachers' beliefs about mean performance of genders, or the reliability of measures. On the other hand, the consistent positive bias in favour of females is compatible with taste-based discrimination, but the simple model does not easily explain the heterogeneity.

Considering the variation in τ holistically (i.e. in favour of females, and against FSM students), there is evidence that the heterogeneity across subjects is consistent with statistical discrimination. However, simple versions of either model cannot easily explain the heterogeneity across grade boundaries. These effects could be driven by complex versions of either model, for example where teachers only favour high-achieving low-income students. But once we allow for adding complexities, it becomes harder to discern whether the bias is statistical or taste-based in nature. What can be said is that the actions of teachers cannot easily be classified into purely taste-based or statistical discrimination, and they vary by student type. However, our finding that the extent of the bias increases with discretion does suggest that one way to reduce teacher bias would be to increase the use of standardized assessments.

[17] A more complex model could be made that in subjects with low informativeness, teachers have more opportunity to impose taste-based discrimination, but this would require teachers to care about plausible deniability.

4.5 Impact on Future Outcomes

We now establish the consequences of biases by demonstrating how small differences in rankings impact future university outcomes, given that A-level grades are a key determinant of which degree courses students are accepted to. Now, instead of estimating the imbalance around the thresholds, we use our approach to estimate the impact on university application outcomes.

Table 3 shows the impact for those immediately above a grade boundary, on being accepted to any university course, accepted to their first choice, and accepted to their insurance choice.[18] Focusing first on the pooled estimates in column one, students with a higher grade in one of their three A-level courses are around 3.5p.p. more likely to be accepted onto any course conditional on applying, and 2.3p.p. more likely to be accepted onto their firm, or *preferred* course, conditional on being accepted at university.

We then break down these estimates by grade boundaries, to capture differential effects at different points in the grade distribution. For example, missing out on an A* may not matter much for gaining a place at university overall, but might mean a student misses out on their firm (first) choice, while the extensive margin may be more relevant at lower grades. These hypotheses are reflected in our results, with the highest increase in probability of acceptance on the extensive margin occurring at the A/B and B/C grade boundaries (5.4-5.6p.p.) with minimal impact at the highest grade boundary (1.7p.p.). While we see the largest impact on the intensive margin, getting their first choice, at the highest grade boundary (6.6p.p.).

[18] As students in the UK generally apply and receive offers from university before completing the exams that these offers will hinge upon, they choose a "firm" and an "insurance" choice from their offers. The "firm" choice is their first choice that they will attend if they achieve the required grades in their offer, while the insurance is a back up option in case they fail to achieve the grades required for their first choice.

Table 3: Consequences of achieving a higher grade on being:

(a) accepted anywhere | applying

Grade boundary:	All	A*/A	A/B	B/C	C/D	D/E
	(1)	(2)	(3)	(4)	(5)	(6)
τ	0.035	0.017	0.054	0.056	0.034	0.046
	(0.003)	(0.005)	(0.005)	(0.006)	(0.008)	(0.013)
Constant	0.599	0.684	0.624	0.588	0.580	0.561
T_i	✓	✓	✓	✓	✓	✓
N	99,938	24,038	29,682	25,684	14,930	5,604

(b) accepted at "firm" choice | accepted

Grade boundary:	All	A*/A	A/B	B/C	C/D	D/E
	(1)	(2)	(3)	(4)	(5)	(6)
τ	0.023	0.066	0.037	0.021	0.029	0.023
	(0.003)	(0.006)	(0.007)	(0.008)	(0.011)	(0.019)
Constant	0.506	0.629	0.577	0.534	0.485	0.424
T_i	✓	✓	✓	✓	✓	✓
N	66,586	19,714	20,620	15,344	8,064	2,844

Notes: This table presents estimates of the effect of being ranked one above a grade boundary versus being ranked one below a grade boundary (+1 vs −1) on the probability of being accepted anywhere for university (panel a) conditional on having applied, being accepted in your firm choice (panel b), for the pooled sample (column 1) and by grade boundary (columns 2-6). Standard errors are in parentheses below the estimates.

5 Robustness

5.1 Conditioning

A potential problem with our approach is that we may be picking up achievement gradients. If the ability distribution varies by student characteristics, e.g. the female achievement distribution is to the right of the male distribution (see figures A.2 and A.4 in the appendix), then these achievement gaps will be correlated with student characteristics. So τ will be a combination of gender bias and achievement differences.

To empirically account for this, in our preferred specification (equation 2) we condition on a measure of prior achievement, average GCSE score, to account for pre-existing characteristic-achievement gradients. Here we attempt to improve our predictive accuracy by conditioning on GCSE scores in respective subjects, rather than just the average.

The results of this conditioning can be found in Table 4, with the first two columns showing our baseline estimates from Table 2. The next four columns show how our estimates change when we: (i) restrict the sample to the subset of student-subject observations who take the same subject A-level and at GCSE (column 3);[19] (ii) condition on prior attainment by subject (column 4); (ii) condition on average prior attainment (column 5); and (iv) condition on both (column 6). In each case, we find results of a similar magnitude to our preferred specification.

Simply conditioning on prior achievement may be unsatisfying for two reasons. First, it may be the case that age 16 scores (even within subject) are a poor predictor of future achievement (Wyness et al., 2023). Second, this approach may be too restrictive in terms of functional form assumptions.

We therefore propose two alternative, novel approaches which make fewer assumptions.

5.2 Latent Achievement

The ranking of students does not contain cardinal information on how far in terms of absolute achievement the adjacent students are from the threshold. We create a subsample of students who are similar in terms of potential outcomes, by creating a cardinal metric of latent achievement based on their prior (GCSE) achievement in respective subjects.

For each student, we estimate the probability, ϕ_α, of achieving grade α separately for each subject-grade boundary, conditional on flexible measures of prior achievement. For example, for the A-B boundary in English, we estimate the probability of achieving the

[19] Several A-level subjects included in our main results do not have a corresponding GCSE subject — or at least not one that is widely offered to students — and hence many students taking these A-levels will not have a GCSE in the same subject.

Table 4: Robustness — Conditioning

	Main Sample		Same Subject Sample		
	(1)	(2)	(3)	(4)	(5)
Female (τ)	0.031	0.022	0.034	0.023	0.023
	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)
N	145,880	145,880	76,064	76,064	76,064
FSM (τ)	-0.016	-0.009	-0.017	-0.008	-0.007
	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)
N	88,094	88,094	41,527	41,527	41,527
T_i		✓		✓	✓
T_{is}					✓

Notes: This table presents our main estimates reproduced from table 2 in columns 1 (without prior attainment controls, T_i) and two (with T_i controls), alongside estimates using a more restricted sample (columns 3-6): only those A-level CAGs for which the student sat a GCSE in the same subject. This additional restriction allows us to control for same subject prior attainment (T_{is}).

higher of the grades, ϕ_A, for everyone who ultimately attained an A or B in English. We use the prior 2019 cohort of students to establish the mapping between GCSE and A levels. Our specification is estimated separately for each subject-grade boundary pair, including a cubic in prior attainment (mean GCSE marks), and a school-subject specific measure of value-added. This allows for a very flexible mapping function. Figure 8a shows the propensity scores of students ranked just above the grade boundary in blue (1) and those ranked just below in red (-1). When restricting to students adjacent to the boundary, there are still some small differences in the propensity score distributions.

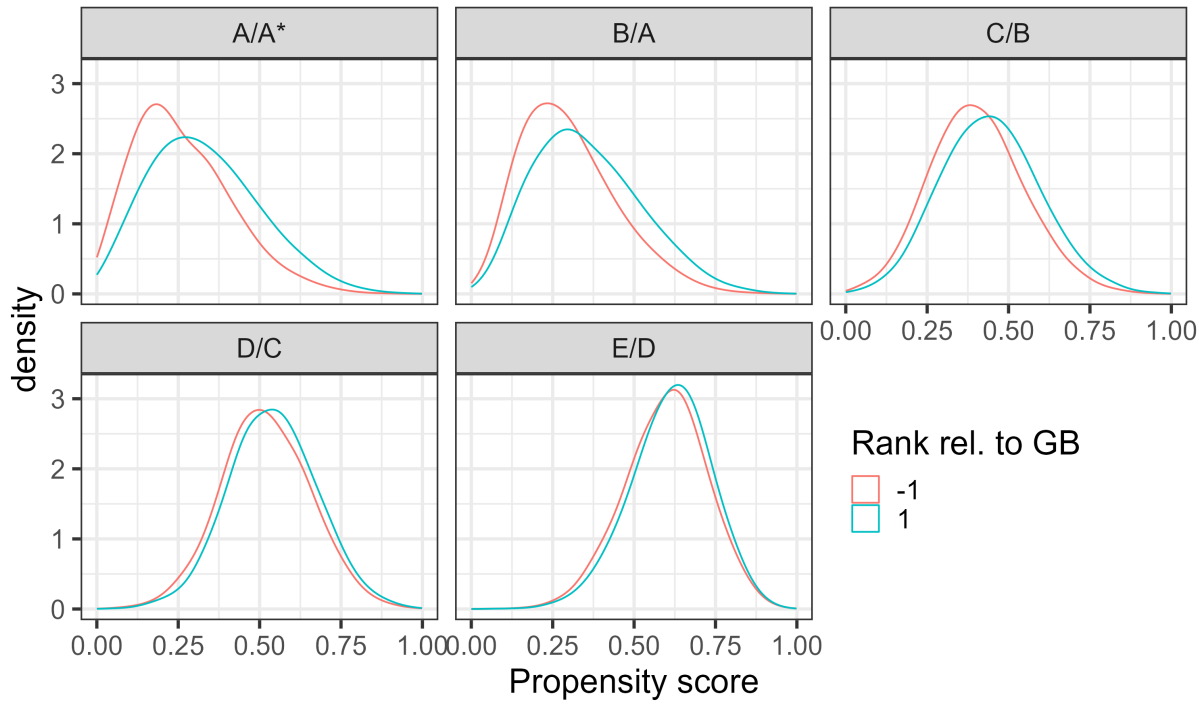
To account for this, we can limit the sample to individuals with similar propensity scores. For example, we can restrict the sample to those who have a propensity of $\phi_\alpha(0.45, 0.55)$, which reduces our number of observations from 158,800 observations to 28,553.[20] Figure 8b shows the propensity to achieve a higher grade for students adjacent to a boundary with this restriction.

In Figure 9 we present our estimates with increasingly restrictive bandwidths. The x-axis shows the size of the bandwidth around 0.5, running from 0.5 either side (i.e. probability between 0 and 1, to match our main estimate) down to 0.05 either side ($\phi_\alpha \in [0.45, 0.55]$). Critically, the figure shows that our conditional estimates are stable even as we narrow the bandwidth, and the unconditional estimates are closer to the conditional estimate. With the tightest bandwidth, the estimates begin to lose stability as we use fewer grade boundaries that have more similar students.

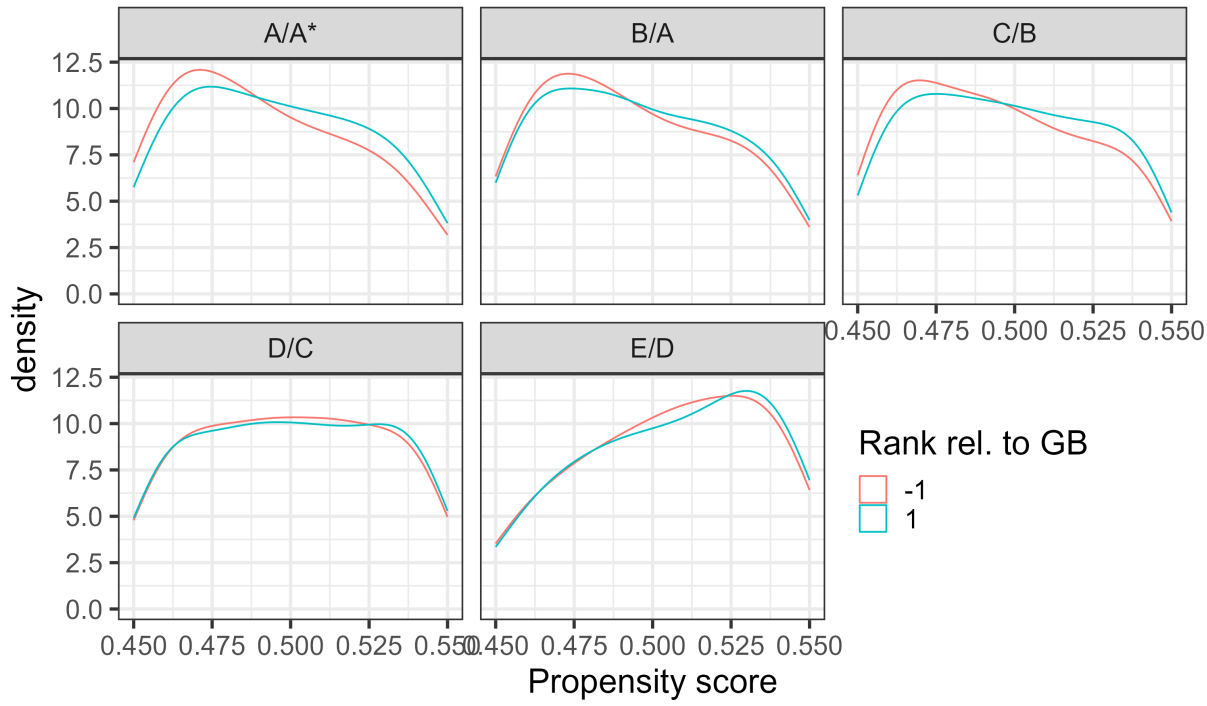
[20]These counts are pre-balance restrictions and pre-dropping of single-X school-subjects.

Figure 8: Histogram of Propensity of Attaining the Higher Grade of Boundary Adjacent Students, by Grade Boundary

(a) All propensity scores: $\phi_\alpha \in [0,1]$



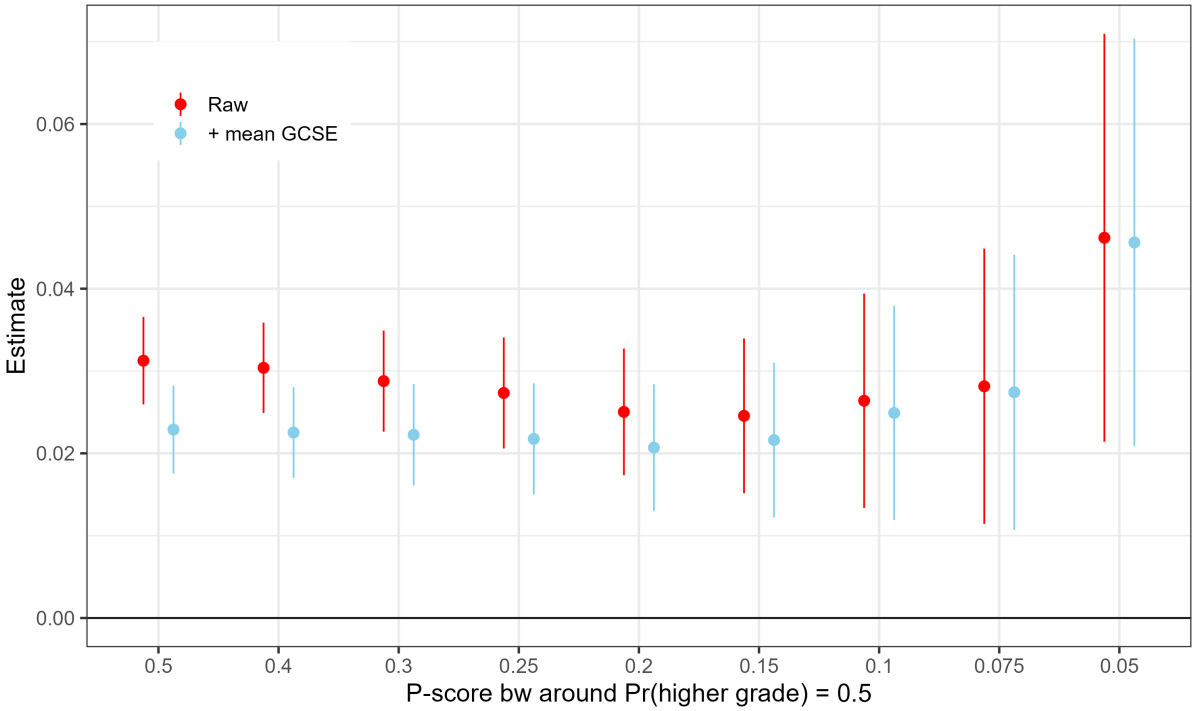
(b) Restricted propensity scores: $\phi_\alpha \in [0.45, 0.55]$



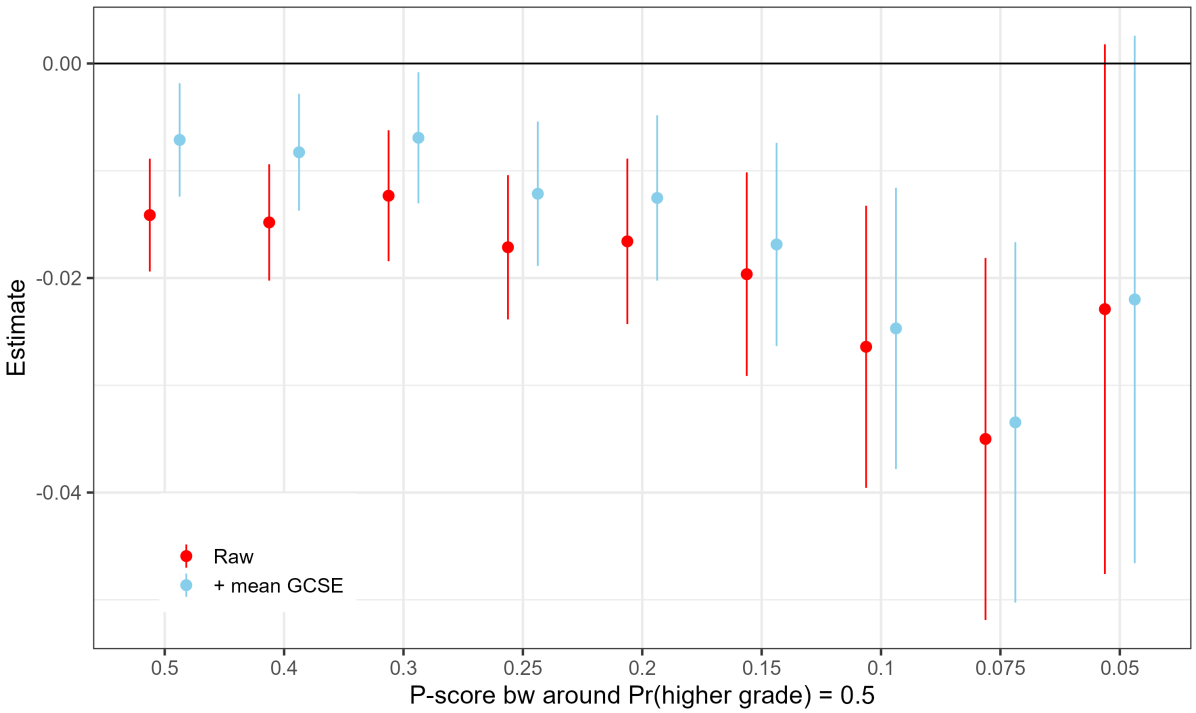
Notes: The blue line shows the histogram of propensity scores for all students ranked one above the grade boundary (+1). The red line shows the histogram of propensity scores for students one below the grade boundary (−1). In panel (a) all observations ranked adjacent to a boundary are included, while panel (b) restricts the sample to those with $\phi_\alpha \in [0.45, 0.55]$.

Figure 9: Latent achievement bandwidth analysis

(a) Female τ



(b) FSM τ



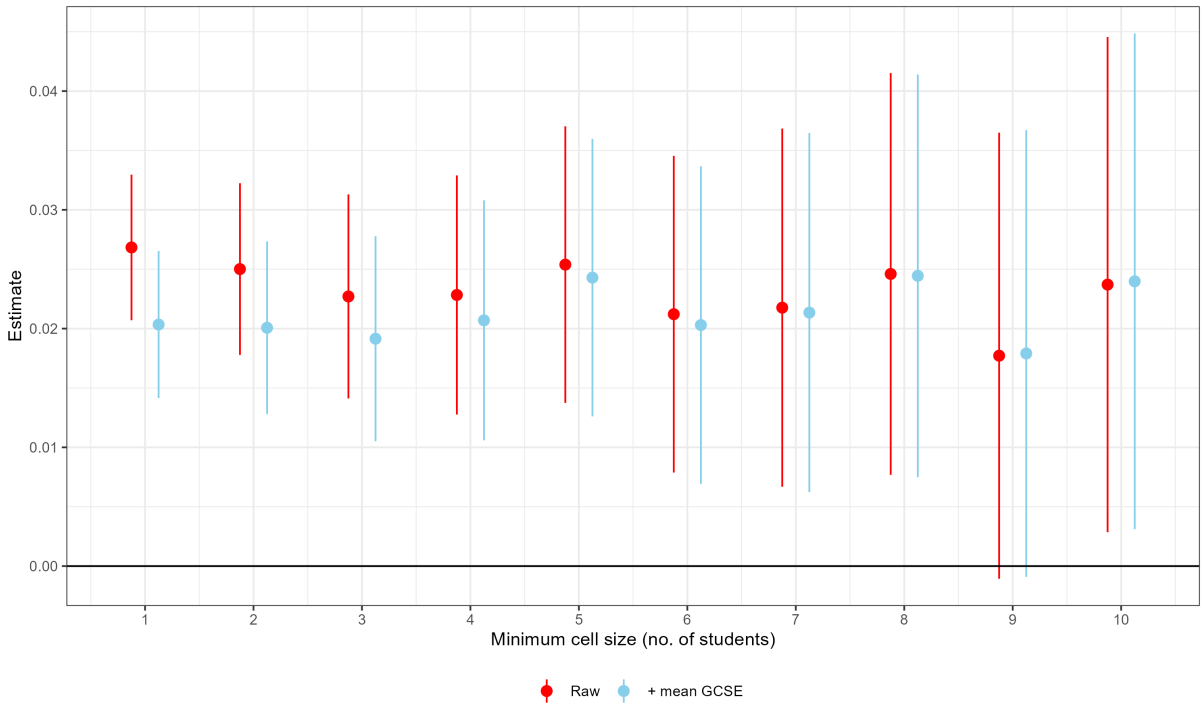
Notes: These panels show the results of a bandwidth analysis exercise where we restrict our sample to a smaller and smaller window propensity score window around the grade boundary. Raw estimates are shown in red, while those conditional on prior attainment are in blue and red, with 95% confidence intervals also shown. In panel (a) female share is the dependent variable, while in panel (b) it is FSM. The same figure with white as a dependent variable in the appendix (figure B.3).

5.3 Crowding

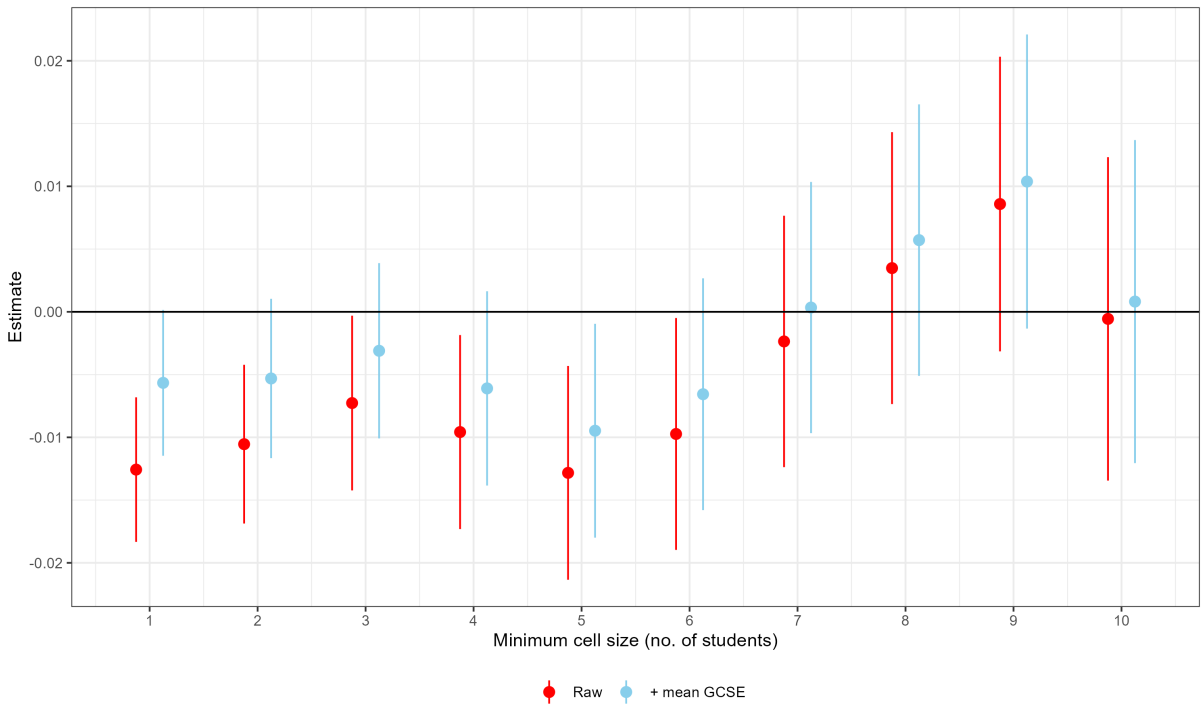
Our third approach for dealing with endogenous boundaries makes even fewer functional form assumptions. Our main specification contains all grade boundaries with at least one adjacent student. However, if there is only a single student of each grade — e.g. only one A student and one B student then it is unlikely they are of similar abilities. In settings with more students in adjoining grades the more likely it is that the marginal students are of similar abilities. Therefore, we re-estimate τ with sub-samples with increasing numbers of students in adjacent grades. This requires no require functional form assumptions between past and present achievement. The results of this exercise can be found in Figure 10. Our baseline estimates are those with at least one student either side, as shown by the estimates on the furthest left. Once again, for female students, the raw estimates converge on our conditional estimates, and the conditional estimates are unchanged as we move to using more marginal students. This coefficient stability implies that τ is not being driven by differences in potential outcomes. For FSM students the estimates are stable until over 80% of boundaries are dropped, up to requiring 7 students in adjoining subject-grades. The bias against FSM students then shrinks and becomes positive. This may be due to the composition of schools changing once we restrict to larger number of students — large schools, in more urban areas, with higher shares of FSM students. Teachers from these schools could be less biased against FSM students as they are exposed to greater numbers of them.

Figure 10: Crowding analysis — increasing number of students in adjacent grades

(a) Female: τ



(b) FSM: τ



Notes: These panels show the results of a crowding analysis exercise where we restrict our sample to students in larger and larger cell sizes (i.e. classes). Raw estimates are shown in red, with those conditional on prior attainment in blue, and 95% confidence intervals displayed. Female share is in panel (a) and FSM in (b). Our estimated with share of white students as a dependent variable are in the appendix, figure B.4.

6 Conclusion

This paper establishes the existence of biases in the way teachers assign grades in a high-stakes assessment. Our method exploits a situation where teachers were asked to rank students within the grades they assigned, allowing us to precisely identify marginal students as defined by subjective teacher assessment. Leveraging the discrete nature of ranks, we implement a Local Randomisation approach to detect bias in teacher-assigned grades. Our intuition is straightforward – we should not expect to find the share of students of a certain group (e.g. females) change disproportionately just above or just below a grade threshold.

We find the share of females to be disproportionately higher on the right hand side of grade boundaries, implying a teacher bias in favour of females. For FSM students, our results imply teachers are biased against them, apart from at the highest grade boundary. Our results suggest teachers are awarding students grades based on their characteristics, rather than solely on the basis of their academic performance. The results are consistent with – and thus provide a new way to validate – the existing literature. Critically, our estimates do not rely on the same assumptions that the standard method requires, that blind and non-blind assessment are measuring the same underlying latent characteristic, or that students exert the same effort and have the same extent of measurement error. If these assumptions do not hold, then this difference-in-difference approach will not recover the parameter of interest. By contrast our local randomisation approach only has one dimension through which students are measured, so any differences in the concentration of a characteristic around a grade boundary must be due to decisions made by teachers, thereby producing a direct measure of teacher bias.

Our results imply that teacher assessments favour some groups of students over others, and that governments considering increasing reliance on these assessments should take steps to mitigate this bias, either by issuing better guidance and information for teachers, or by using externally marked assessments.

The empirical test developed in this study can be used to test for bias in other settings. Our test for bias is appropriate whenever there is subjective assessment of multiple agents/objects in a system with thresholds. The approach is ideal when there is subjective ranking and multiple thresholds, but the same intuition can be applied to settings with cardinal subjective measures of achievement. Our new approach can therefore be used to explore bias in settings beyond education, as varied as subjective competitions, managerial performance metrics or judicial decisions.

References

Alesina, A., M. Carlana, E. La Ferrara, and P. Pinotti (2018): "Revealing stereotypes: Evidence from immigrants in schools," Tech. rep., National Bureau of Economic Research.

Arenas, A. and C. Calsamiglia (2025): "Gender differences in high-stakes performance and college admission policies," *Management Science*.

Arnold, D., W. Dobbie, and C. S. Yang (2018): "Racial bias in bail decisions," *The Quarterly Journal of Economics*, 133, 1885–1932.

Arrow, K. J. (1972): "Some mathematical models of race discrimination in the labor market," *Racial discrimination in economic life*, 187–204.

Avitzour, E., A. Choen, D. Joel, and V. Lavy (2020): *On the origins of gender-biased behavior: The role of explicit and implicit stereotypes*, National Bureau of Economic Research, issue: w27818).

Becker, G. S. (2010): *The economics of discrimination*, University of Chicago press.

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

Breda, T. and M. Hillion (2016): "Teaching accreditation exams reveal grading biases favor women in male-dominated disciplines in France," *Science*, 353, 474–478.

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

Burgess, S., D. Hauberg, B. Rangvid, and H. Sievertsen (2022): "The importance of external assessments: High school math and gender gaps in STEM degrees," *Economics of Education Review*, 88, 102267.

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

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

Cassagneau-Francis, O. and G. Wyness (2025): "The merits of teacher assessment versus external exams to measure student achievement," *IZA World of Labor*.

Cattaneo, M. D., N. Idrobo, and R. Titiunik (2024): "A Practical Introduction to Regression Discontinuity Designs: Extensions," *Elements in Quantitative and*

Computational Methods for the Social Sciences, iSBN: 9781009441896 9781009462327 9781009441902 Publisher: Cambridge University Press.

Cavaglia, C., L. Macmillan, K. Maragkou, R. Murphy, and G. Wyness (2024): "The mismatch earnings penalty," Tech. rep., UCL Centre for Education Policy and Equalising Opportunities.

Chetty, R., D. J. Deming, and J. N. Friedman (2023): "Diversifying Society's Leaders? The Determinants and Causal Effects of Admission to Highly Selective Private Colleges," .

Chyn, E., B. Frandsen, and E. Leslie (2025): "Examiner and judge designs in economics: a practitionerâs guide," *Journal of Economic Literature*, 63, 401–439.

Cohen, A. and C. S. Yang (2019): "Judicial politics and sentencing decisions," *American Economic Journal: Economic Policy*, 11, 160–191.

Delaney, J. M. and P. J. Devereux (2025): "Teacher Bias and Evaluation Differences in Test Scores: Different Methods for Different Questions," *Oxford Bulletin of Economics and Statistics*.

Dessein, W., A. Frankel, and N. Kartik (2025): "Test-optional admissions," *American Economic Review*, 115, 3130–3170.

EPI (2021): "Analysis: A Level Results 2021," .

Friedman, J. N., B. Sacerdote, D. O. Staiger, and M. Tine (2025): "Standardized test scores and academic performance at ivy-plus colleges," Tech. rep., National Bureau of Economic Research.

Galasso, V. and P. Profeta (2024): "Gender differences in math tests: The role of time pressure," *The Economic Journal*, 134, 3461–3475.

Goodman, S. (2016): "Learning from the test: Raising selective college enrollment by providing information," *Review of Economics and Statistics*, 98, 671–684.

Graetz, G. and A. Karimi (2022): "Gender gap variation across assessment types: Explanations and implications," *Economics of Education Review*, 91, 102313.

Hanna, R. N. and L. L. Linden (2012): "Discrimination in Grading," *American Economic Journal: Economic Policy*, 4, 146–168.

Hinnerich, B. T., E. Höglin, and M. Johannesson (2011): "Are boys discriminated in Swedish high schools?" *Economics of Education review*, 30, 682–690.

Hirnstein, M., J. Stuebs, A. Moè, and M. Hausmann (2023): "Sex/gender differ-

ences in verbal fluency and verbal-episodic memory: a meta-analysis," *Perspectives on Psychological Science*, 18, 67–90.

Holbein, J. B. and H. F. Ladd (2017): "Accountability pressure: Regression discontinuity estimates of how No Child Left Behind influenced student behavior," *Economics of Education Review*, 58, 55–67.

House of Commons Education Committee (2020a): "Getting the grades they've earned Covid-19: the cancellation of exams and 'calculated' grades," Tech. rep.

——— (2020b): "Getting the grades they've earned: COVID-19: the cancellation of exams and 'calculated' grades: Response to the Committee's First Report," .

Ichino, A., M. Polo, and E. Rettore (2003): "Are judges biased by labor market conditions?" *European Economic Review*, 47, 913–944.

Kautz, T., J. J. Heckman, R. Diris, B. Ter Weel, and L. Borghans (2014): "Fostering and measuring skills: Improving cognitive and non-cognitive skills to promote lifetime success," *National Bureau of Economic Research*.

Kisfalusi, D., B. Janky, and K. Takács (2021): "Grading in Hungarian primary schools: Mechanisms of ethnic discrimination against Roma students," *European Sociological Review*, 37, 899–917.

Lavy, V. (2008): "Do gender stereotypes reduce girls' or boys' human capital outcomes? Evidence from a natural experiment," *Journal of Public Economics*, 92.

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

Lavy, V. and E. Sand (2015): "On the origins of gender human capital gaps: Short and long term consequences of teachers," *stereotypical biases*, publisher: National Bureau of Economic Research.

Leonhardt, D. (2024): "The Misguided War on the SAT," *The New York Times*.

Li, F., A. Mercatanti, T. MÃ€kinen, and A. Silvestrini (2021): "A regression discontinuity design for ordinal running variables: Evaluating central bank purchases of corporate bonds," *The Annals of Applied Statistics*, 15.

Li, Danielle, M. H. and L. B. Kahn (2018): "Discretion in hiring," *The Quarterly Journal of Economics*, 765, 800.

Moss, G., H. Goldstein, S. Hayes, B. M. Chereau, P. Sammons, G. Sinnott, and G. Stobart (2021): "High standards, not high stakes: An alternative to SATs that will transform England's testing & school accountability system in primary edu-

cation & beyond [BERA Expert Panel on Assessment Report]," *British Educational Research Association*.

Murphy, R. and G. Wyness (2020): "Minority report: the impact of predicted grades on university admissions of disadvantaged groups," *Education Economics*, 28, 333–350, publisher: Routledge _eprint: https://doi.org/10.1080/09645292.2020.1761945.

ofqual (2020): "Summer 2020 grades for GCSE, AS and A level, Extended Project Qualification and Advanced Extension Award in maths," .

Ofqual (2020): "Summer 2020 grades for GCSE, AS and A level, extended project qualification and advanced extension award in maths," .

Ofqual, UCAS, and Department for Education (2025): "GRading and Admissions Data England-Ofqual-DfE-UCAS," .

Phelps, E. S. (1972): "The statistical theory of racism and sexism," *The american economic review*, 62, 659–661.

Portuguese Ministry of Education (2023): "New conditions for completion of secondary education and access to higher education," .

Rimfeld, K., M. Malanchini, L. J. Hannigan, P. S. Dale, R. Allen, S. A. Hart, and R. Plomin (2019): "Teacher assessments during compulsory education are as reliable, stable and heritable as standardized test scores," *Journal of Child Psychology and Psychiatry*, 60, 1278–1288, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jcpp.13070.

Taylor, C. R. and H. Yildirim (2011): "Subjective performance and the value of blind evaluation," *The Review of Economic Studies*, 78, 762–794.

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

Wyness, G., L. Macmillan, J. Anders, and C. Dilnot (2023): "Grade expectations: how well can past performance predict future grades?" *Education Economics*, 31, 397–418.

Zhu, M. (2024): "New Findings on Racial Bias in Teachers' Evaluations of Student Achievement," Tech. rep., IZA Discussion Papers.

7 Appendix

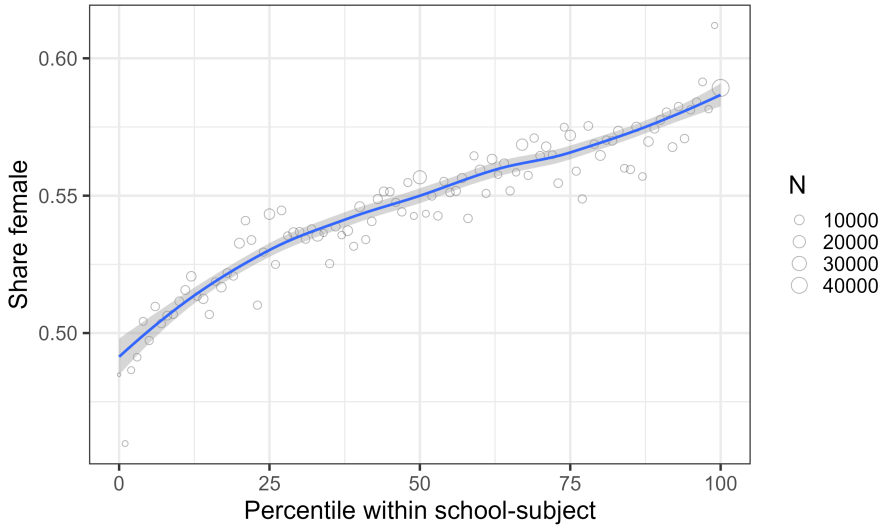
A Attainment gradients

Figures A.2a – A.2c concatenate the rankings across grades within a school-subject, to provide an overall percentile rank within school-subject, to illustrate the underlying relationships between student characteristics and teacher assessments. The positive gradient for share of female, shows that a higher share of high ranked students are female. There is a positive gradient but less strong for white students. In contrast there is a negative gradient for the share of FSM students.

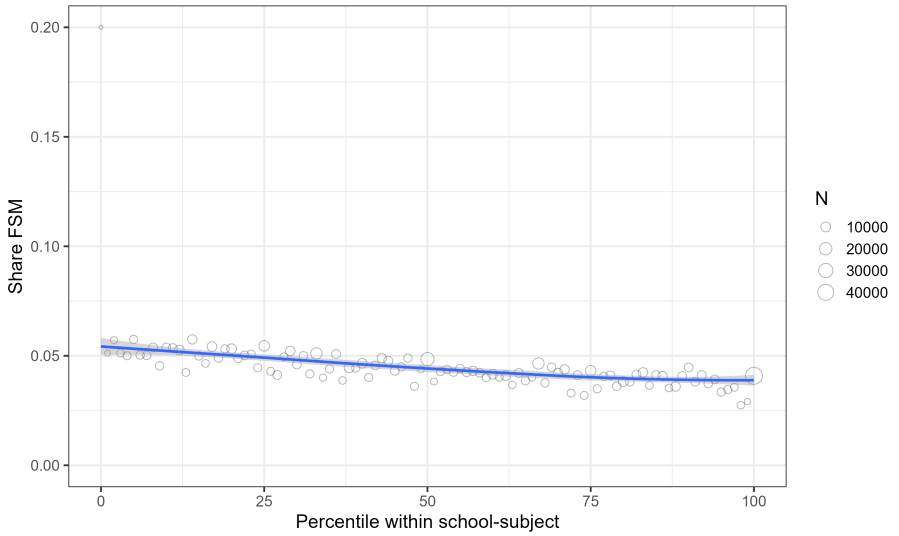
We repeat the exercise for each of the top 20 subjects in figure A.2.

Figure A.1: Overall attainment-characteristic gradients (2019)

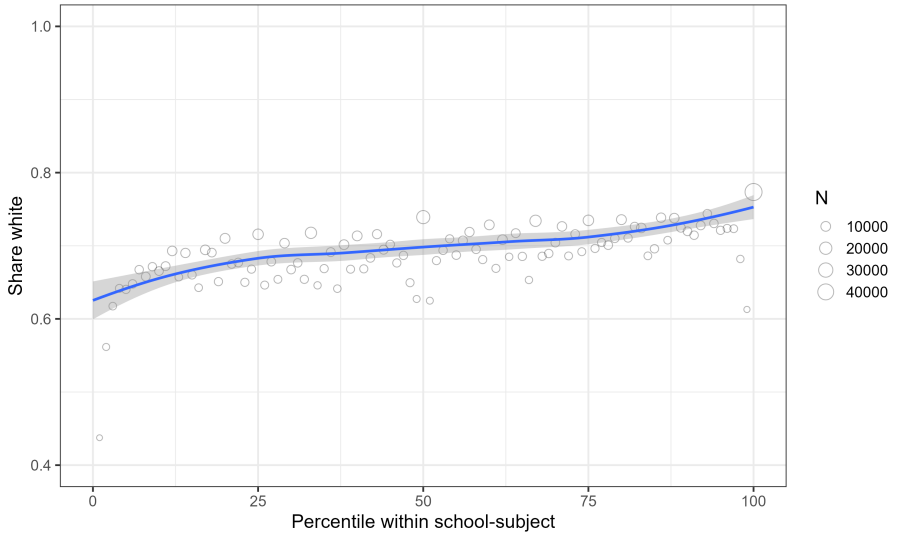
(a) Female



(b) FSM



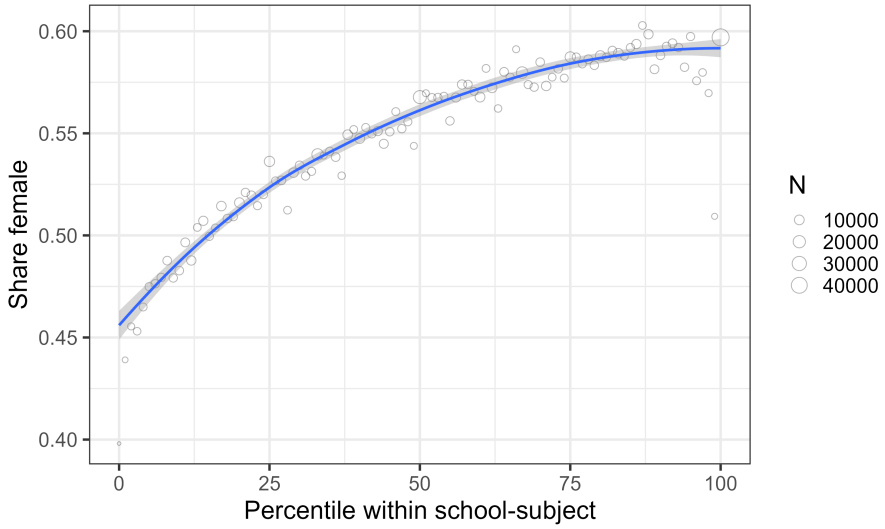
(c) White



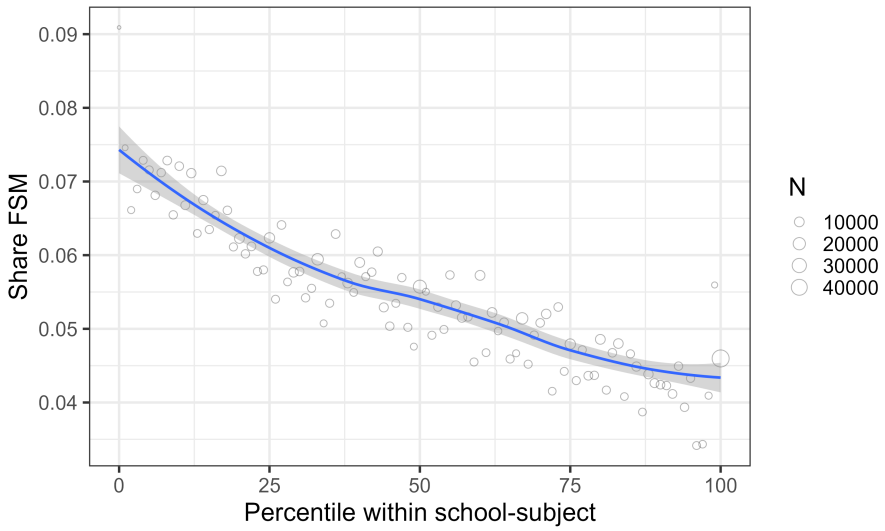
Notes: These plots show the share of characteristic X among all students for each percentile in a school and subject. Then a line of best fit is drawn using local regressions.

Figure A.2: Overall attainment-characteristic gradients (2020)

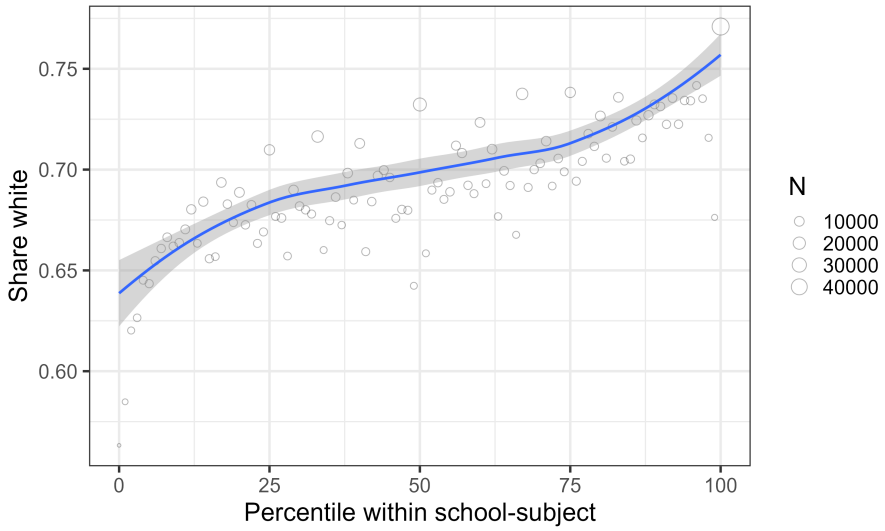
(a) Female



(b) FSM



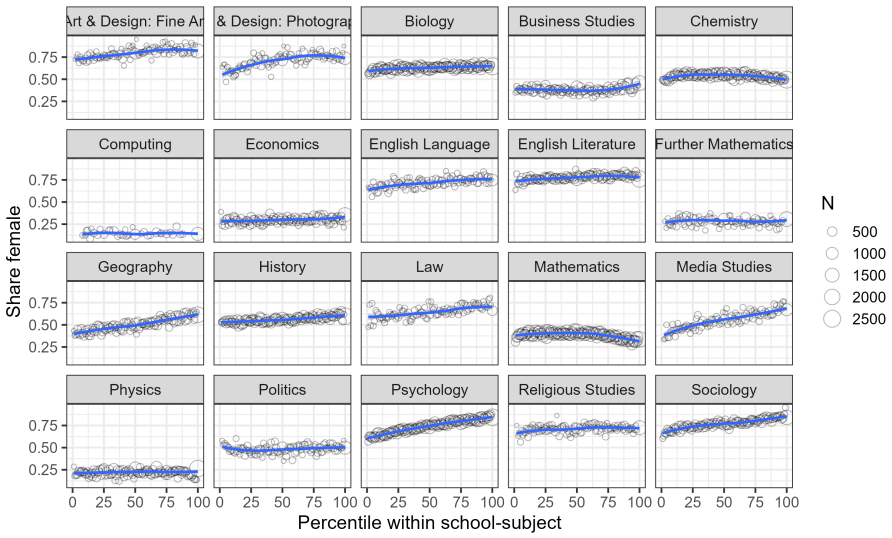
(c) White



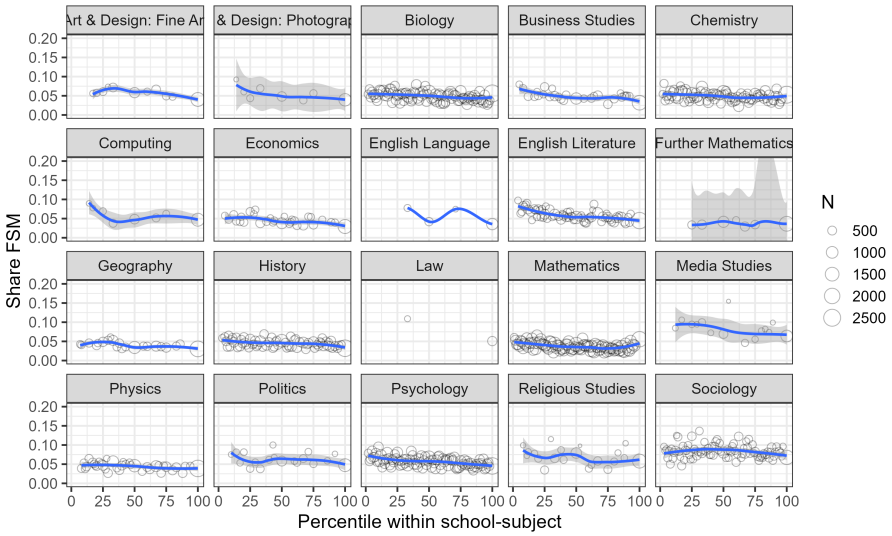
Notes: These plots show the share of characteristic X among all students for each percentile in a school and subject. Then a line of best fit is drawn using local regressions.

Figure A.3: Attainment-characteristic gradients by subject (2019)

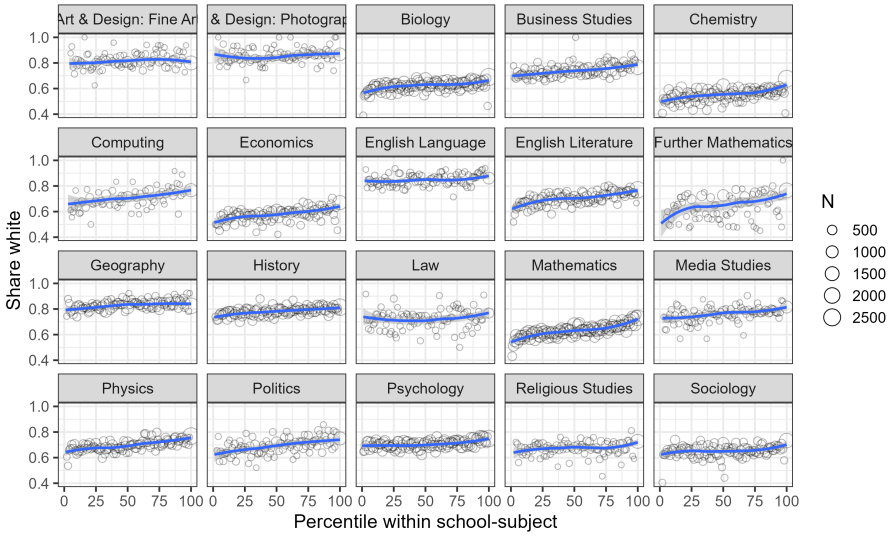
(a) Female



(b) FSM



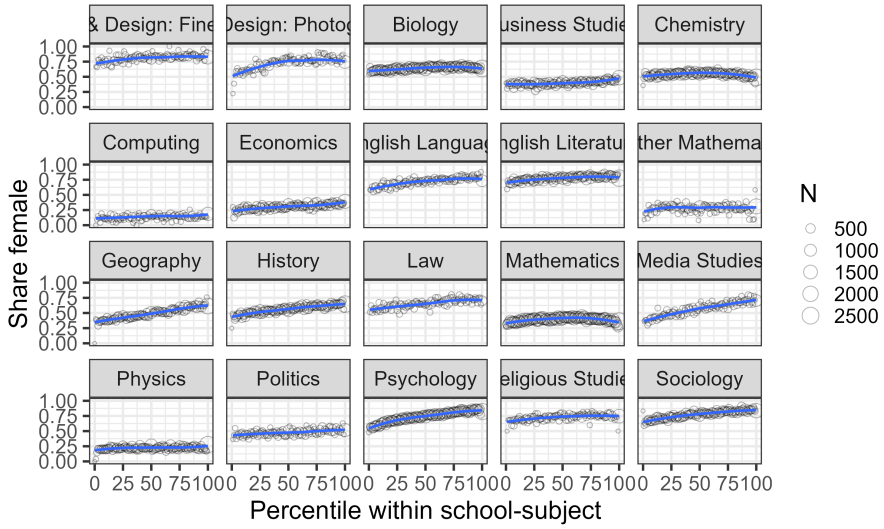
(c) White



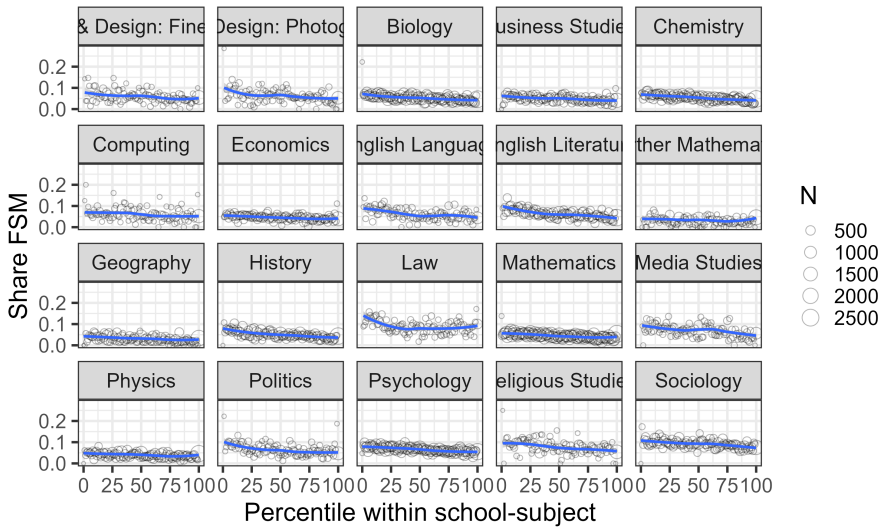
Notes: These plots show the share of characteristic X among all students for each percentile in a school and subject. Then a line of best fit is drawn using local regressions.

Figure A.4: Attainment-characteristic gradients by subject (2020)

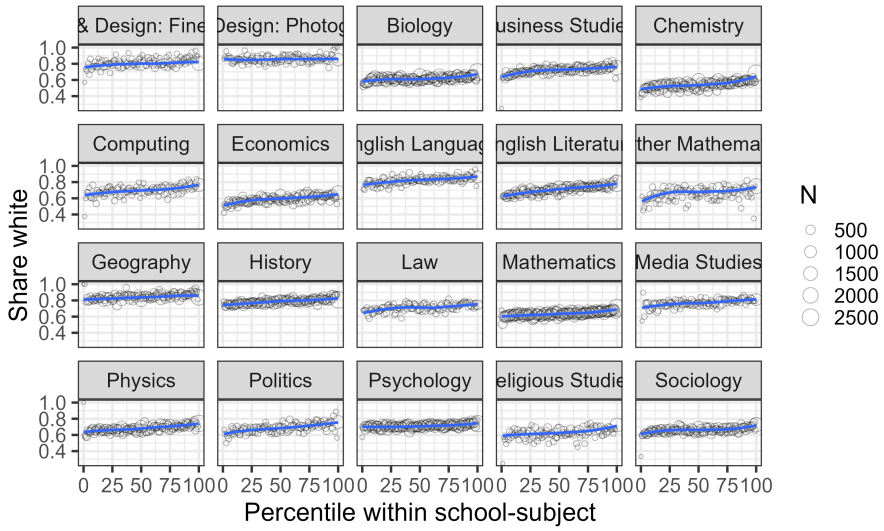
(a) Female



(b) FSM



(c) White



Notes: These plots show the share of characteristic X among all students for each percentile in a school and subject. Then a line of best fit is drawn using local regressions.

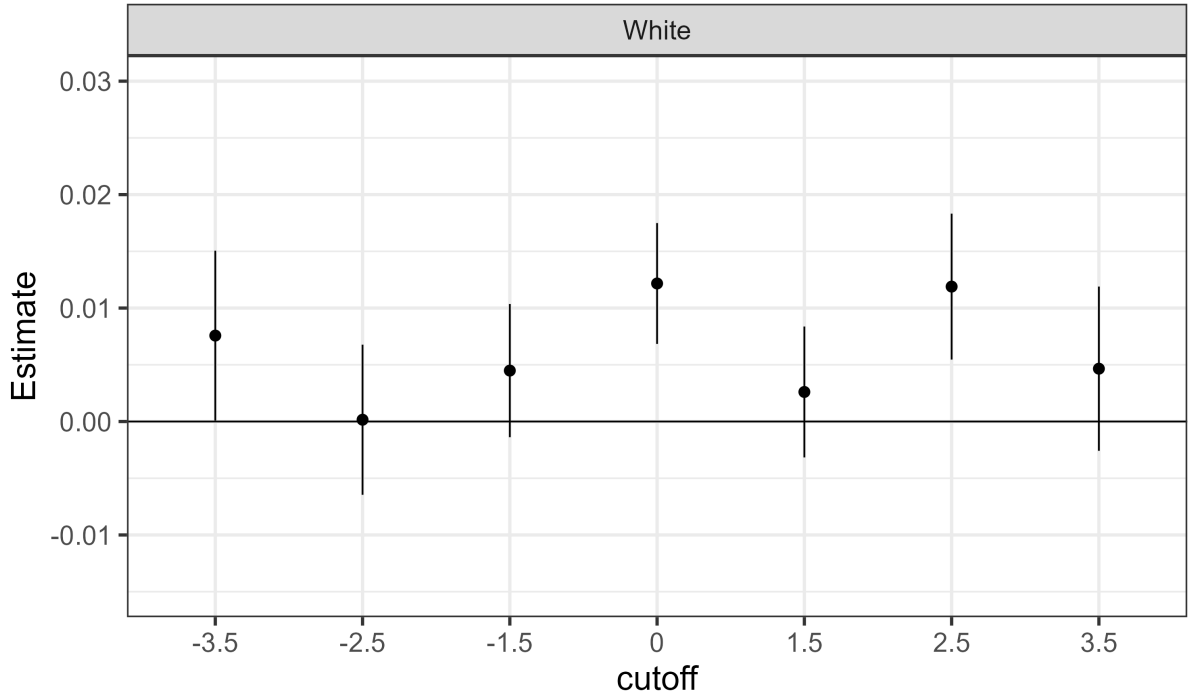
B Results by ethnicity

Table B.1: Main results + conditioning robustness: ethnicity

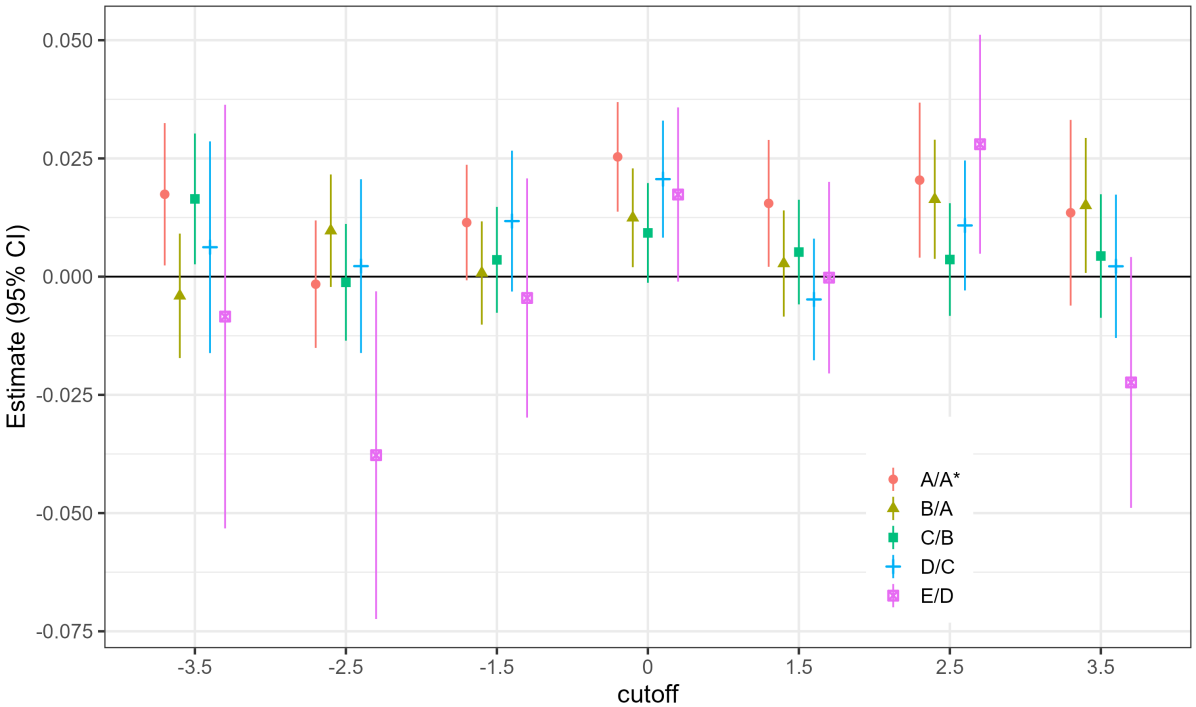
	Main Sample		Same Subject Sample		
	(1)	(2)	(3)	(4)	(5)
White (τ)	0.019	0.012	0.022	0.014	0.013
	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)
T_i		✓		✓	✓
T_{is}					✓
N	126,818	126,818	61,846	61,846	61,846

Figure B.1: Placebo tests by distance from grade boundary: ethnicity

(a) Pooled sample



(b) By grade boundary



Notes: This figure presents our main estimates from table 2, alongside "placebo" estimates for pairs of adjacently ranked observations that do not straddle grade boundaries. For example, while the point at 0 represents the estimated effect at a true grade boundary, the estimate at -1.5 represents the estimated effect if we place a "placebo" grade boundary between students ranked 1 and 2 within a grade. The extending bars represent 95% confidence intervals. All estimates are conditional on prior attainment and weighted by how close the share of X is to 50% within a school-subject.

Figure B.2: Estimates by subject: ethnicity

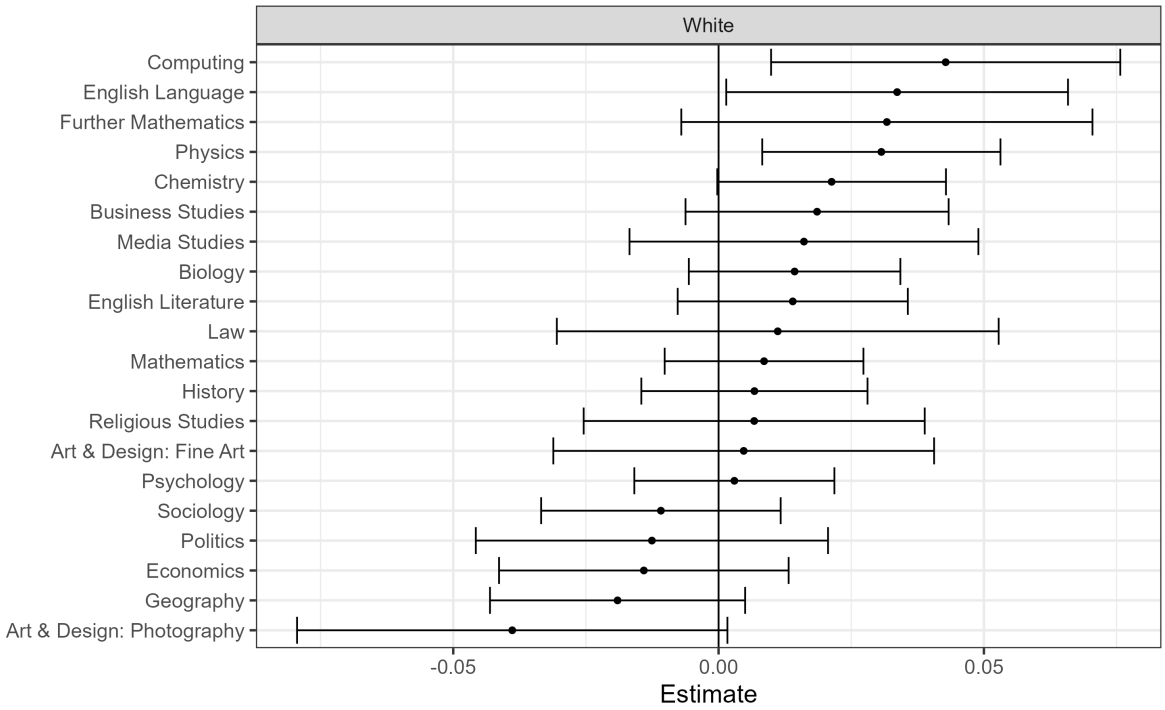


Figure B.3: White τ

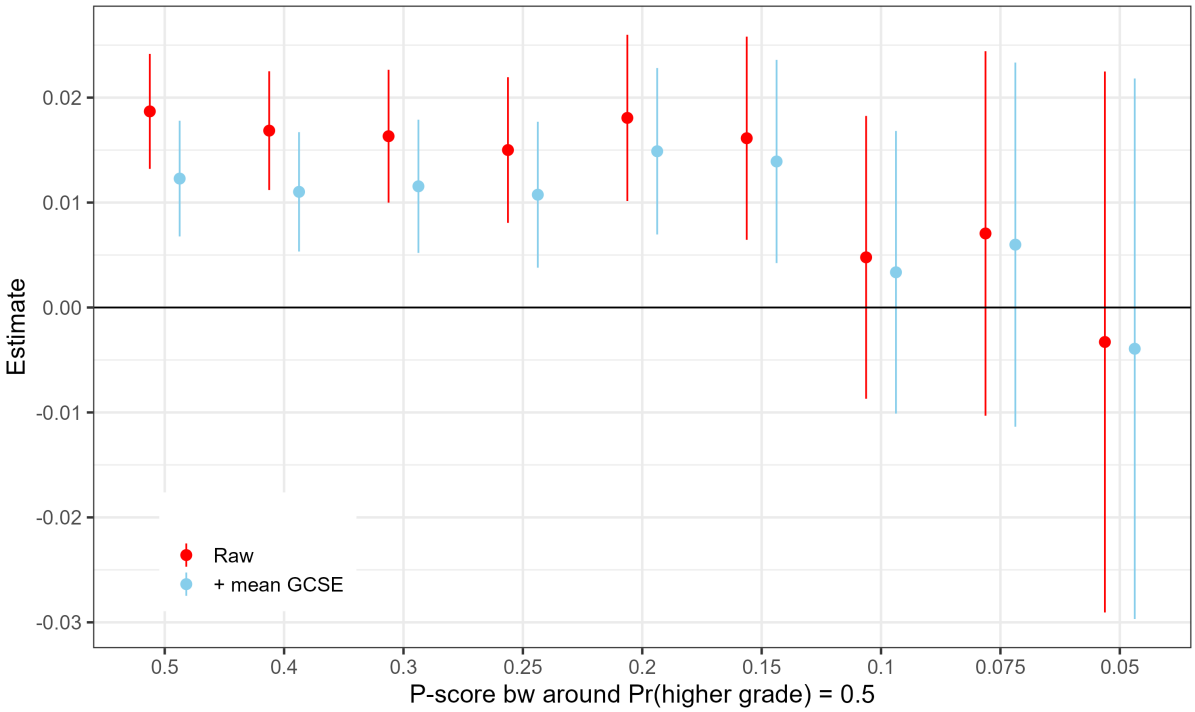
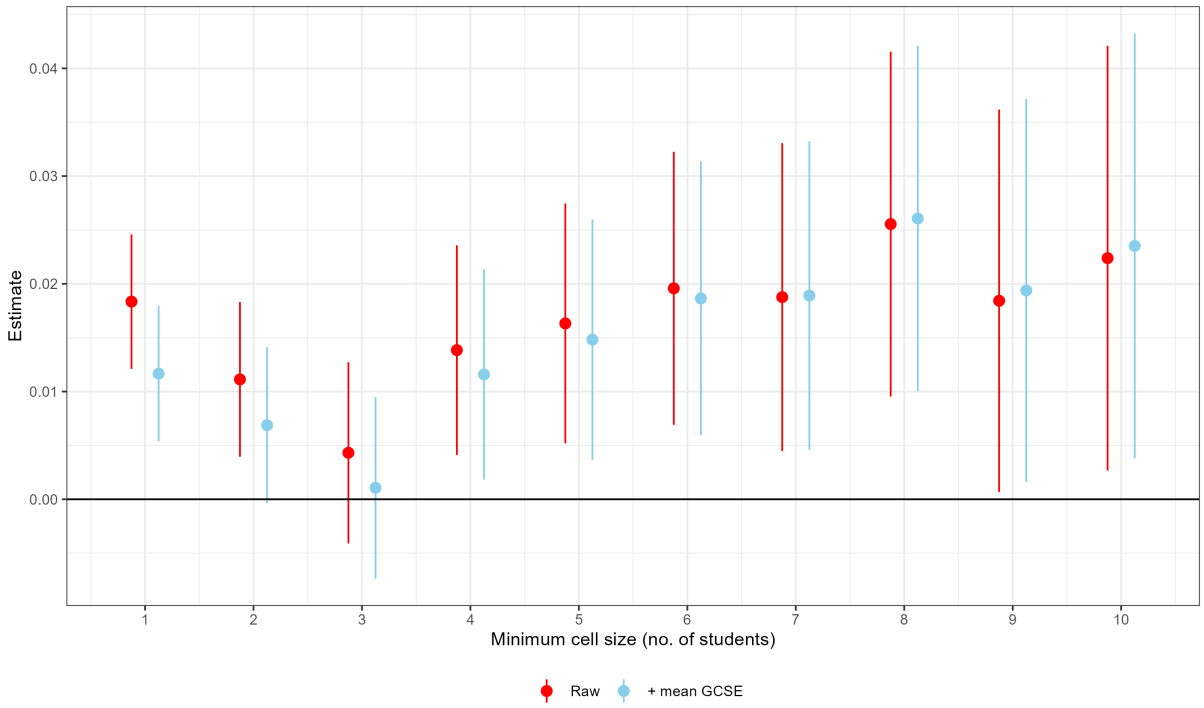


Figure B.4: White: τ versus minimum adjacent students



ucl.ac.uk/ioe/cepeo