# Measuring systematic gaps in teacher judgement: A new approach

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

21st February 2025

**PRELIMINARY DRAFT. PLEASE DO NOT CITE.**

**Abstract**

Do teachers mark some pupils more generously than others? We propose a new approach to this longstanding question, by exploiting a unique situation where teachers were required to assign grades and the rankings of students within grades for a high-stakes assessment. We use this to test for imbalance of student characteristics across grade boundaries by comparing the top ranked students of one grade, to the lowest ranking students of the next grade. Due to the discrete nature of ranks, we implement an extension to the RDD framework called local randomization. This does not require the standard assumptions used in teacher bias literature. We find evidence of teacher generosity on average favoring higher income, female, white students. We go on to show that these grades are meaningful, being just above the grade boundary significantly increases a students probability of going to university, and being accepted at their first choice of degree.

**Keywords**: Teacher judgements; teacher bias; ethnicity; gender; exams.
**JEL codes**: I21; I23; I24; I28

*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

Many countries are considering increasing their reliance on teacher-based assessment in place of standardized testing for the university application process (Australia, UK (British Educational Research Association 2021), USA (Leonhardt, 2024), Portugal (Portuguese Ministry of Education, 2023)).Proponents of teacher-assessed grades argue they reduce stress for students and teachers, allow for broader curricula, are less prone to measurement error, and are highly correlated with test scores (Rimfeld et al., 2019). However, there are also those advocating for a move in the other direction as standardized assessments are better predictors of performance at college than high-school GPA (Chetty et al., 2023). The key concern with regards to subjective teacher assessments, is that they can exacerbate existing racial and socio-economics gaps. There is a literature establishing evidence of bias in subjective teacher assessments, a bias that is avoided by blind and externally marked exams (Burgess and Greaves, 2013; Carlana, 2019; Terrier, 2020; Burgess et al., 2022).

In this paper, we provide new evidence of the existence of teacher bias by providing a new approach that does not require the standard assumptions of the literature. Our analysis exploits the Covid-19 induced cancellation of standardized end of secondary education (A-level) exams in the UK, and the nature of the teacher-assessment which replaced them. Students were assigned a letter grade by their teachers in each A-Level subject they were taking (typically 3), known as CAGs.[1] In addition to assigning CAGs to students, teachers were also asked to rank students within each letter grade. This allows for the ranking of every A-level student by subject in each secondary school.[2] We combine this subjective ranking of students (Grading and Admissions Data for England (GRADE) dataset) with administrative data on student characteristics, including demographics, prior achievement, and their applications to college. This allows us to test if there are discontinuous jumps in the characteristics of students across every school-subject-grade boundary. We implement a implement a Local Randomisation (LR, Cattaneo et al., 2024) approach comparing the mean characteristics of the individuals each side of the threshold. We utilize the feature which comes from ranking students that boundaries will have a mass of observations immediately adjacent to them, allowing us to identify the most marginal students.

[1]These teacher-assessed grades were officially called "centre-assessment grades" (henceforth CAGs), and were 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).

[2]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).

The intuition behind this approach is as follows. While there might be an underlying relationship between student attainment and these characteristics (for example female students generally perform better than male students at A-level, EPI, 2021), these relationships should be continuous. Therefore we should not see jumps or drops in the share of students across all schools with certain characteristics across grade boundaries. For example, while there may be more female students achieving A grades than male students, there should not be a higher concentration of female students immediately above the A grade threshold. Specifically we implement a LR approach with teacher-assessed rankings as the running variable, using only the most marginal students at grade-subject to test for imbalance in student characteristics.

We find evidence of discontinuities in favour of female students and white students, and against those receiving free-school meals, across a range of grade boundaries. The extent of these discrepancies is sizable, ranging between 1 and 4 percentage points. For example, we observe a 3 percentage point increase in the proportion of females at the bottom of the B grade, compared to the top of grade C, suggesting generosity in favour of females at the C-B grade boundary. The largest gaps occur at the D/C grade boundary, which is generally considered the pass fail threshold. When considering the extent of this bias by subject, we find that teachers tend to favour the gender that is more represented. Female students are 1 percentage points more likely to be ranked immediately above the threshold in sociology, while they are 0.5 percentage points less likely in physics. The biases are consistently larger for characteristics which are more easily observable (gender, ethnicity) compared to those which are not (Free School Meal Eligibility). Such sizable discontinuities across grade boundaries indicate biases in how teachers assigned these high-stakes grades in 2020.

The next section establishes that these small differences in rankings, resulting in different grades have important implications to student life courses. By comparing the university application outcomes of those immediately above to immediately below a grade boundary, we show that students are 3.3 percentage points more likely to be accepted to a university course if they were assessed to be above a grade boundary. These students were also 3 percentage points more likely to be accepted on to their most preferred course.

The discrete nature of using rank as a running variable poses some non-standard advantages and disadvantages. The primary advantage is that we always have students immediately each side of a subject-grade boundary, as long as there are any students with each grade. This means that we have many (271,278) marginal observations, which allows for focusing just on the most marginal students. Focusing on the most marginal students means we do not need to make functional form assumptions about the relationship between students' rank and characteristics away from the threshold in order to project to the threshold. The primary disadvantage is that the ordinal nature of the rank

means that we do not know 'how far' a student is from a boundary in terms of the teacher assessment of performance, only their ranking. To account for this we implement three distinct approaches which rely on different assumptions. Primarily, and most straightforwardly we condition on prior attainment of each student measured two years previously in standardized examinations. Second, we estimate a propensity for students to be adjacent to a boundary, and only estimate on the subset of students with a similar propensity to be around a particular threshold, and who are adjacent to a threshold. Third, we show that our estimates are robust to increasing the minimum number of students in adjacent grade-school-subject. This approach does not require functional form assumptions about the relationship between prior and current academic performance. The intuition is that in schools where there are more students with A and B grades, the most marginal students are likely to be closer in academic ability than in schools with fewer students.

As the teachers had complete discretion in terms of ranking students, both the running variable and the cutoffs can be considered endogenous. Indeed it is exactly these endogenous choices that teachers make with respect to student characteristics that we are attempting to capture. A teacher may decide on the location of a threshold so that the marginal female is above it. We perform a series of robustness tests to ensure that these discontinuities are not due to spurious correlations, including subject-boundary fixed effects, and establishing that there are no discontinuities in student characteristics away from the threshold.

The key contribution of this paper is that it provides a new approach to the established methods in the teacher bias literature. Many of these papers uses a double-differences (DiD) methodology (Blank, 1991; Breda and Ly, 2015; Burgess and Greaves, 2013; Falch and Naper, 2013; Goldin and Rouse, 2000; Lavy, 2008; Lavy and Sand, 2018; Terrier, 2020), using the difference in performance between blind and non-blind assessment metrics between one group and another. The key assumption is that the blind (standardised tests) and non-blind (teacher) assessment are attempting to measure the same underlying latent characteristic. If that assumption does not hold, then the parameter of interest with this DiD approach will also contain differences in other characteristics. For example if verbal examinations are non-blind and written examinations are blind, then any difference in the gender gap across these examination types could be due to bias, or to differences in written-verbal abilities by gender. In contrast our LR approach only has one dimension through which students are measured, which is the non-blind teacher assessment. We do not require that the blind and any non-blind assessment are measuring the same student aspect of achievement. Any differences in the concentration of a characteristic around a grade threshold will be due to decisions made by teachers, thereby producing a direct measure of teacher bias. Whilst not directly comparable, our estimates are consistent with those found in the established literature. Teachers are biased towards white and

female students.

The second contribution of our paper is that it allows, for the first time, estimates of the heterogeneity of teacher bias by student achievement. Prior research which used the differences in achievement for identification prevents this type of analysis. The we find that teacher bias is largest in the middle of the distribution around a traditional pass-fail boundary implies that teachers are considering the implications of their decisions on student outcomes. This provides pause that such effects are due to unconscious bias.

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. Section 5 concludes.

# 2 Background and context

In this section we first describe the institutional context of exams and higher education admissions in the UK relevant for this paper, and then describe the Covid-19 induced exam cancellations and replacement with the teacher-assessed grades which are central to our analysis.

## 2.1 Institutional context

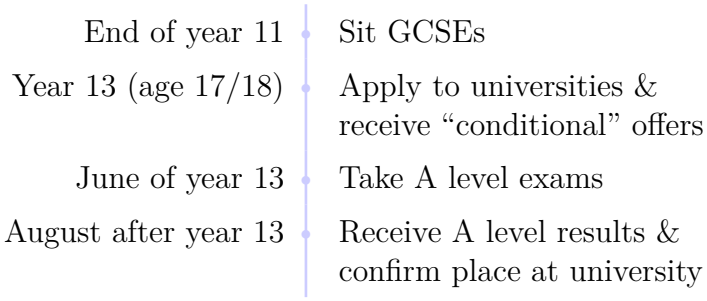
High school students in the UK take standardised exams at age sixteen called GCSEs, and then continue on either an academic track to study A-levels or follow a (more) vocational track and attend a further education (FE) college. We focus on A-levels in this paper. Figure 1 shows the timeline of educational decisions that students who wish to attend university generally face. At the beginning of their final year of high school students apply to universities and during the spring they receive conditional offers. These conditional offers prescribe the results they should achieve in their A-levels to confirm their place. During a standard academic year, 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.

## 2.2 Changes due to Covid-19

The Covid-19 pandemic meant that the majority of schools closed in March 2020, and they did not reopen in time for exams. This meant that all standardised exams in the UK were cancelled in 2020, including GCSEs and A-levels. The timetable of these cancellations for the cohort who were to be taking their A-levels in 2019 is in figure 1b. As is clear from figure 1b, exams were cancelled at the last minute. Given the high-stakes nature of both

Figure 1: Timeline of educational decisions

(a) Normal timeline




(b) Pandemic affected timeline in 2020

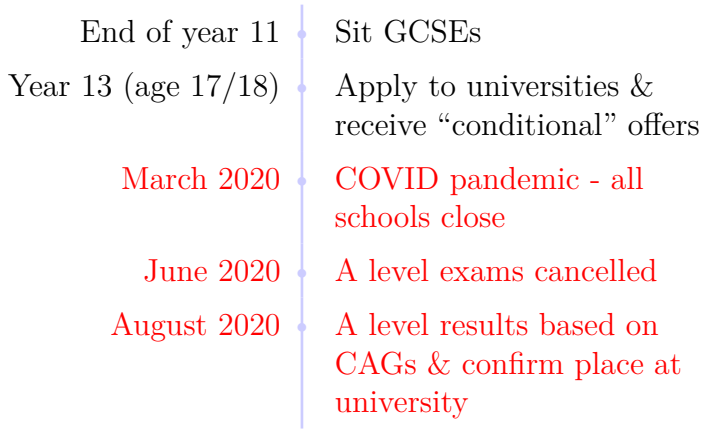
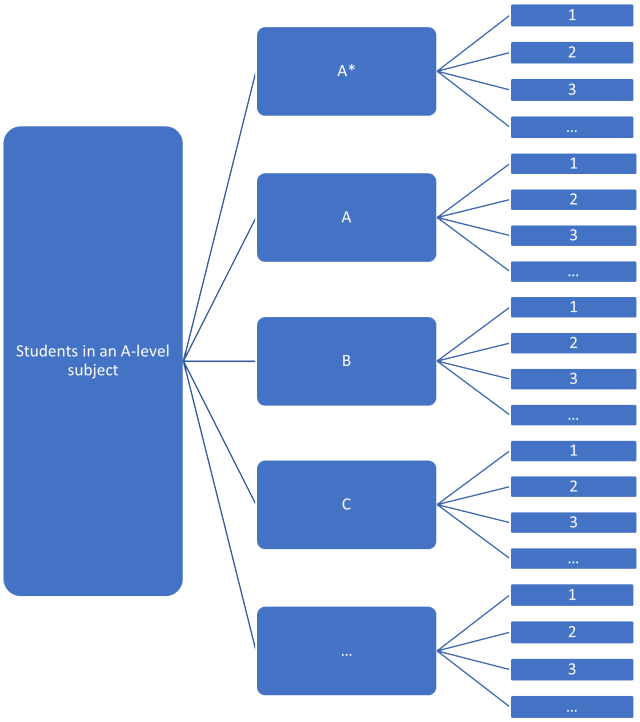


Figure 2: School subject grading




GCSE and A level exams, it was decided to replace the exams with teacher predicted grades, so that students would still end up with a set of exam results.

After a short consultation, the exam regulator Ofqual decided on a process to determine these grades. Ofqual's guidance was that "Exam 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." Teachers were asked to grade student on the following basis "the grade that each student is most likely to have achieved if they had sat their exams".

Every school was required to submit teacher assessed grades and a rank order of their pupils within each grade. This process is illustrated in Figure 2.

While teachers were given no reasoning for why they need to provide rank information, the intention of the Department for Education (DfE) requesting this rank information was to combat anticipated grade inflation as a result of using teacher assessment. The DfE intended to use the rank information, along with student prior achievement, and previous school-subject achievement distributions to monotonically transform teacher assigned grades. Indeed, the algorithm was implemented and did what it was supposed to do — lowered grades below the teacher predictions. However, upon the release of the augmented grades there was a national outcry that an algorithm had decided students' A-level grades. After 3 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 never actually used, they are central to our estimation strategy, as we explain in our empirical approach below.

# 3 Data and empirical approach

## 3.1 Dataset

Our analysis makes use of the Grading and Admissions Data for England (GRADE) dataset. This dataset contains pupil-level data on GCSE and 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). Four cohorts of data are provided (2017 to 2020) allowing researchers to compare the covid-19 affected cohort (2020) to three previous cohorts (2017, 2018 and 2019). The dataset contains both the GCSE and A level scores of the students in the normal years, including rarely available raw data on the actual marks awarded by exam markers, rather than purely the grades themselves, and the CAGs and teacher rankings awarded in 2020. For the purposes of this paper, we restrict our analysis to the cohort who were subject to teacher rankings (i.e. the 2020 cohort).

As we are interested in inequalities in teacher generosity by pupil characteristics, our variables of interest are the gender of the pupil (Female/Male), their ethnicity (White/other ethnic group), and their socio-economic status. For socio-economic status we use information on the free school meals (FSM) status of the student.

## 3.2 Empirical strategy

We aim to exploit the 2020 exam cancellation to allow us to explore discrepancies across pupil groups in teacher judgement.

As described above, in 2020, teachers were asked to rank each student within a subject and grade level. To assess the extent to which there are discrepancies by pupil characteristics, we will examine these rankings, testing for imbalances in the density of students from different groups around grade thresholds using a Local Randomisation (LR) approach. Specifically, our method employs this 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 operationalised by using the subsample of students ranked immediately adjacent

to a subject-grade boundary within their school (i.e. first or last within a school-subject-grade cell), with a simple specification:

$$Y_{isg} = \beta_0 + \tau D_{isg} + \varepsilon_{isg} \tag{1}$$

where $Y_i$ is an indicator of student $i$'s characteristic (female, white, FSM), D is an indicator for if student $i$ is to the right hand side of a boundary in subject $s$ at grade boundary $g$. The parameter of interest is $\tau$, which represents the percentage point difference in the share of a characteristic on one side of the boundary compared to the other.

The intuition behind this approach is straightforward: there should not be a discontinuous jump in the proportion of students of a particular characteristic around these thresholds. If there were statistically more students of a certain type marginally below a grade threshold than above it that would imply that teachers are awarding some (otherwise identical) groups of students more generous predictions around grade thresholds. For example, if we observe a higher proportion of female students towards the bottom of the A rankings, and a lower proportion towards the top of the B rankings, this implies there may be systematic bias in favor of female students, pushing these students from a B to an A.

In all cases, we require that any school included must have at least 3 students either side of the boundary. As analysis of this sort requires large sample sizes, as such for our main analysis we combine all subjects together and grade boundaries together. To account for differences in characteristics across subjects and grade boundaries we present estimates which include subject, grade boundary, or subject-grade boundary fixed effects (2). We then explore discontinuities for each subject and boundary separately.

$$Y_{isg} = \beta_0 + \tau D_{isg} + D_{sg} + \varepsilon_{is} \tag{2}$$

Using this approach, we ask whether there are systematic differences in how teachers predicted the grades of certain types of students in 2020 — specifically looking at gender, and FSM status. These differences could be due to conscious or unconscious bias. Note, even if teachers are awarding grades on the basis on conduct in class, there is no reason to believe that this would change discontinuously around a boundary.

$$Y_{isg} = \beta_0 + \tau D_{isg} + X_i + D_{sg} + \varepsilon_{is} \tag{3}$$

A potential concern is that students adjacent to a grade boundary may not be equally distant in terms of academic achievement. A as consequence we are concerned that $\tau$ may be picking up the underlying achievement gradient by student characteristic e.g. female

Table 1: Descriptive statistics

(a) Population share and numbers of students

| | Female | FSM eligible | White |
|---|---|---|---|
| Share | .55 | .07 | .73 |
| Number of students | 222,643 | 198,320 | 198,320 |

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

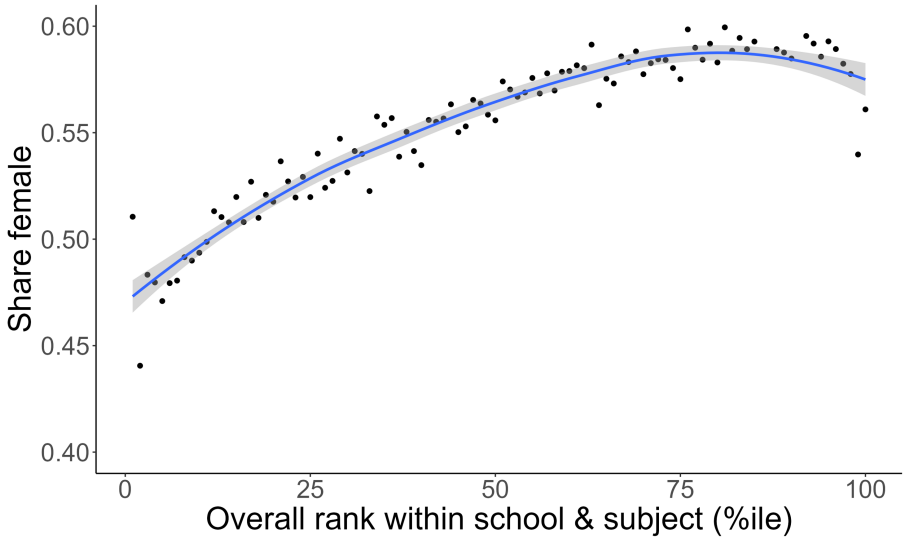
| | Female | FSM eligible | White |
|---|---|---|---|
| *Share* | | | |
| Overall | .53 | .08 | .75 |
| *Number of observations* | | | |
| Overall | 258,246 | 228,570 | 228,625 |
| *By subject* | | | |
| Mathematics | 18,395 | 15,910 | 15,913 |
| Psychology | 16,685 | 14,925 | 14,927 |
| Biology | 16,809 | 14,658 | 14,663 |
| Chemistry | 15,382 | 13,347 | 13,348 |
| History | 14,396 | 13,113 | 13,113 |
| Sociology | 12,104 | 11,718 | 11,720 |
| English literature | 14,320 | 12,544 | 12,545 |
| Business studies | 11,154 | 9,384 | 9,386 |
| Physics | 14,064 | 12,224 | 12,228 |
| Economics | 9,457 | 7,646 | 7,647 |
| *By grade boundary* | | | |
| A/A* | 54,311 | 45,287 | 45,292 |
| B/A | 67,048 | 57,096 | 57,106 |
| C/B | 67,082 | 58,731 | 58,737 |
| D/C | 53,342 | 48,063 | 48,072 |
| E/D | 29,495 | 27,523 | 27,531 |

students are higher achieving, and so we would expect a marginally higher concentration of female students on the RHS of the boundary. We primarily account for this by conditioning on a cubic polynomial for prior achievement $X$.

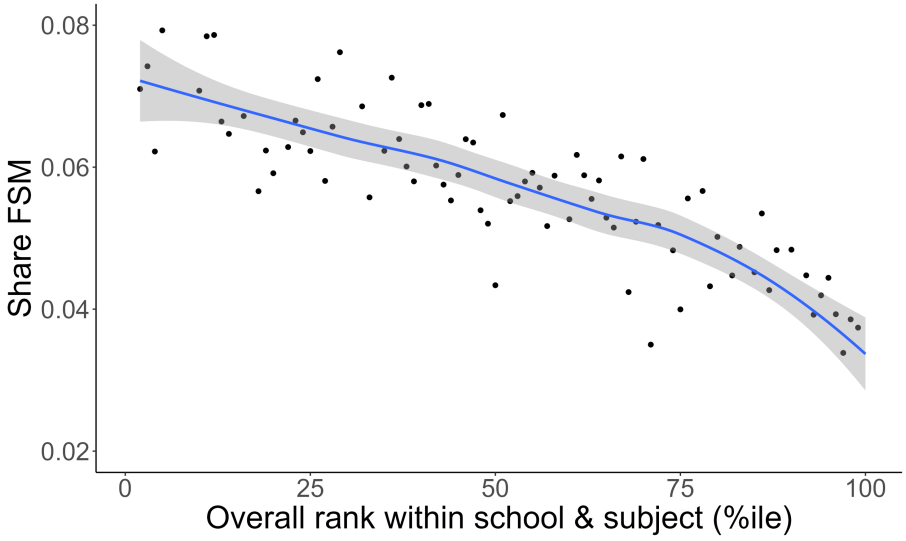
Table 1 shows how our sample breaks down by key characteristics in 2019 and 2020. While there is complete data on the gender of students, information on students free school meal status, and ethnicity are incomplete. Thus, for our results using FSM students, we use only our partial sample. The shares of these characteristics in this subsample are the same as in the population.

Figures 3a – 3c concatenate the rankings across grades within a school-subject, to provide an overall percentile rank within school-subject, to illustrate the underlying relationships

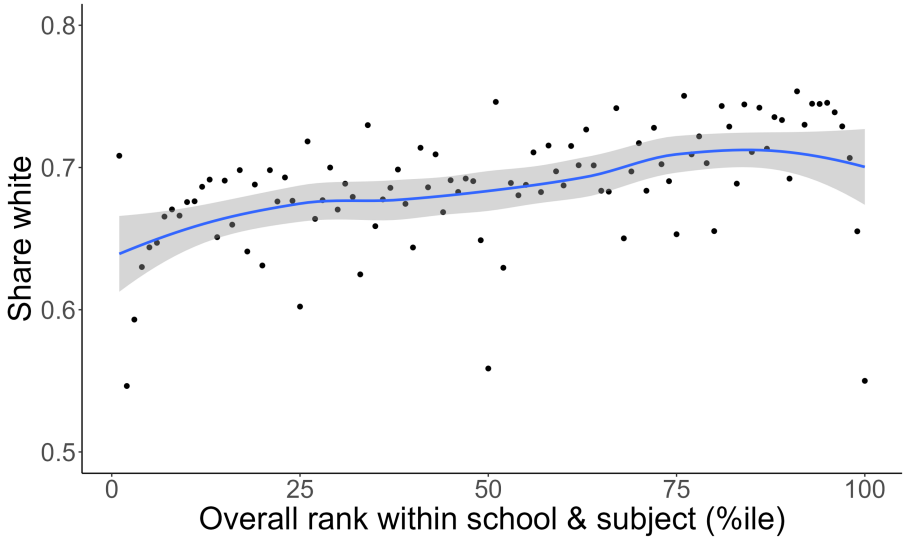
Figure 3: Characteristics by percentile (2020)




(a) Gender




(b) FSM




(c) White

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.[3]

# 4 Results

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

In column 1 of tables 2 a-c, we show the raw average differences in proportion of each respective characteristic from the lower (left hand side, henceforth LHS) side of the boundary to the upper (right hand side, RHS) side, across all subjects and grade boundaries. The coefficient of 0.026 implies that across all schools and all subject-grade boundaries females are 2.6 percentage point more likely to be ranked on the RHS of a boundary compared to the LHS. White students are also over represented on the RHS (1.5 percentage points), while FSM students are 0.6 percentage points under represented.

To account for differences in the proportion of respective characteristics across subjects, and at each grade boundary, the subsequent columns additionally condition on subject, grade boundary, and subject-grade boundary fixed effects (Cols 1-4). The inclusion of any of these sets of fixed effects makes little difference to the estimated parameters.

Column 5 conditions on prior achievement, to account for expected distance from the grade boundary. There is 1.1 percentage point more females above the threshold compared to below. Similarly white students are 1 percent more represented, and FSM students are 0.3 percent under represented. While the magnitude of each coefficients has reduced, they remain significant. Standard errors are clustered at the school level. 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.

Figure 4 presents the extent of female bias across grade boundaries, and discontinuities across ranks away from the grade boundary, conditional on prior achievement. 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 represents students ranked 1 and 2 above the cutoff. If there was no teacher bias all estimates would cross 0 on the y-axis. However, we observe clear jumps in the proportion of female students at the D versus E, C versus D and B versus C boundaries, implying

[3] Although this broad categorisation ignores important difference across students from minority ethnic groups who do perform vastly differently, relative both to other minority ethnic groups and to white students.

Table 2: Stacked subject and grade boundary estimates

(a) Gender

| *Dep. var.:* Female | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| RHS of GB ($\tau$) | 0.026*** | 0.025*** | 0.026* | 0.025*** | 0.011*** |
| | (0.002) | (0.003) | (0.007) | (0.002) | (0.002) |
| Mean GCSE | | | | | 0.099*** |
| | | | | | (0.001) |
| Constant | 0.540*** | | | | |
| | (0.001) | | | | |
| Subject FEs | | ✓ | | | |
| GB FEs | | | ✓ | | |
| Subject×GB | | | | ✓ | |

*Notes:* $^{\dagger}p < .1$, $^{*}p < .05$, $^{**}p < .01$, $^{**}p < .001$. $N = 258,246$ for all specifications. Columns (1)-(5) all estimate variations of equation (1) with no controls in column (1), adding subject and GB FEs in (2) and (3), subject-GB FEs in (4) and controlling for mean GCSEs (our measure of prior attainment) with no FEs in (5).

(b) FSM

| *Dep. var.:* FSM | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| RHS of GB ($\tau$) | −0.006*** | −0.006*** | −0.006** | −0.006*** | −0.003*** |
| | (0.002) | (0.003) | (0.001) | (0.001) | (0.001) |
| Mean GCSE | | | | | −0.019*** |
| | | | | | (0.000) |
| Constant | 0.075*** | | | | |
| | (0.001) | | | | |
| Subject FEs | | ✓ | | | |
| GB FEs | | | ✓ | | |
| Subject×GB | | | | ✓ | |

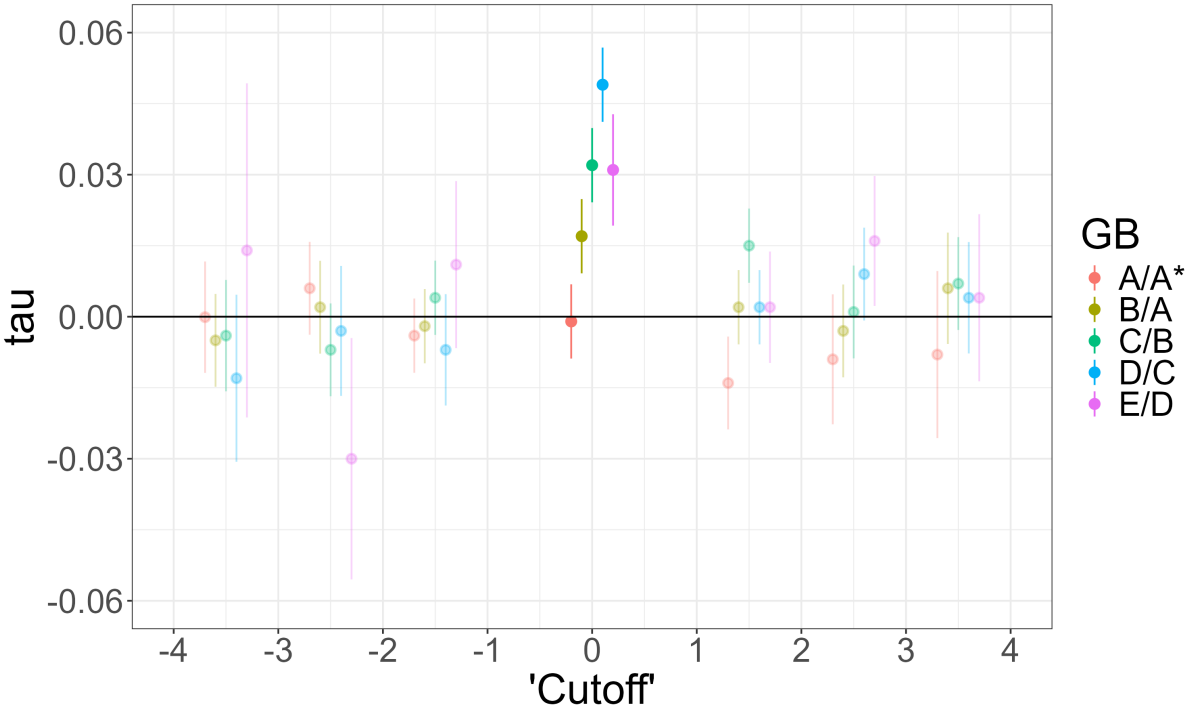
*Notes:* $^{\dagger}p < .1$, $^{*}p < .05$, $^{**}p < .01$, $^{**}p < .001$. $N = 228,570$ for all specifications. Columns (1)-(5) all estimate variations of equation (1) with no controls in column (1), adding subject and GB FEs in (2) and (3), subject-GB FEs in (4) and controlling for mean GCSEs (our measure of prior attainment) with no FEs in (5).

(c) Ethnicity

| *Dep. var.:* White | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| RHS of GB ($\tau$) | 0.015*** | 0.014** | 0.015* | 0.014*** | 0.010** |
| | (0.002) | (0.003) | (0.004) | (0.002) | (0.002) |
| Mean GCSE | | | | | 0.032*** |
| | | | | | (0.001) |
| Constant | 0.746*** | | | | |
| | (0.001) | | | | |
| Subject FEs | | ✓ | | | |
| GB FEs | | | ✓ | | |
| Subject×GB | | | | ✓ | |

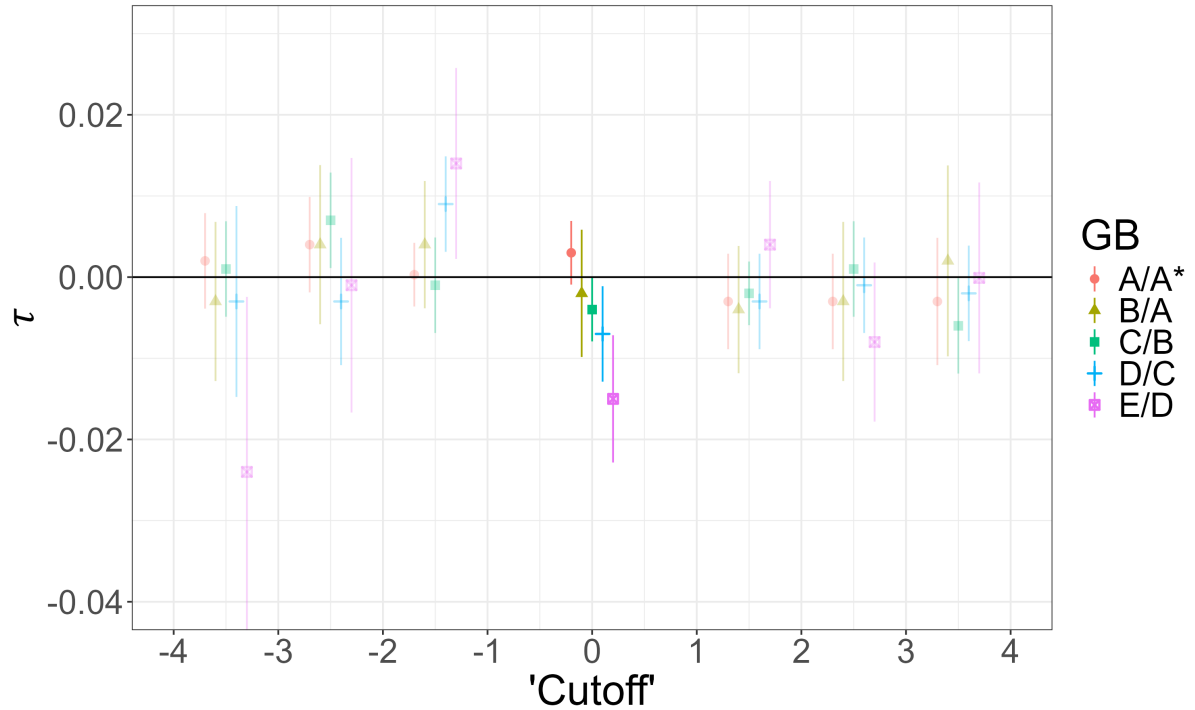
*Notes:* $^{\dagger}p < .1$, $^{*}p < .05$, $^{**}p < .01$, $^{**}p < .001$. $N = 228,570$ for all specifications. Columns (1)-(5) all estimate variations of equation (1) with no controls in column (1), adding subject and GB FEs in (2) and (3), subject-GB FEs in (4) and controlling for mean GCSEs (our measure of prior attainment) with no FEs in (5).

Figure 4: Estimated pro-female bias ($\tau$) by grade boundary (with placebo tests)



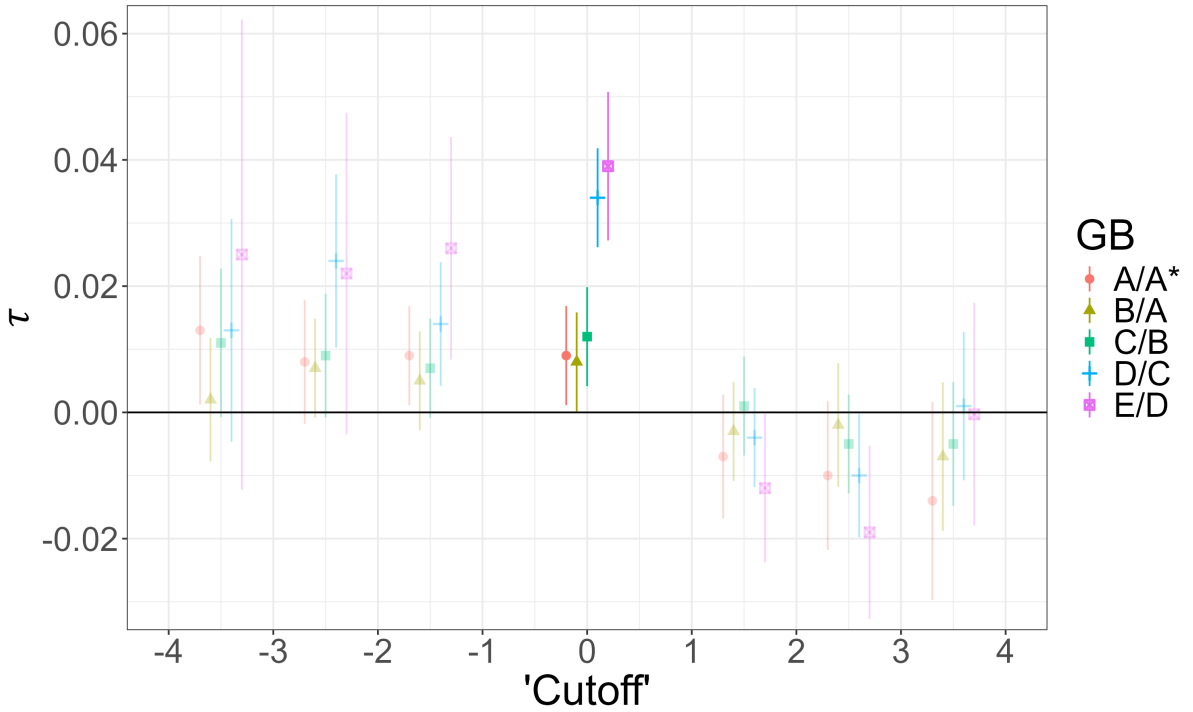

*Notes:* The coloured points show the estimated pro-female bias ($\tau$) with different shapes (and colours) corresponding to different grade boundaries. The lineranges are 95% confidence intervals. Fully opaque points/lines are estimates across a true boundary, while the transparent points/lines are placebo tests at other adjacent ranks.

Figure 5: Estimated pro-FSM bias ($\tau$) by grade boundary (with placebo tests)




*Notes:* The coloured points show the estimated pro-female bias ($\tau$) with different shapes (and colours) corresponding to different grade boundaries. The lineranges are 95% confidence intervals. Fully opaque points/lines are estimates across a true boundary, while the transparent points/lines are placebo tests at other adjacent ranks.

Figure 6: Estimated pro-White bias ($\tau$) by grade boundary (with placebo tests)




*Notes:* The coloured points show the estimated pro-female bias ($\tau$) with different shapes (and colours) corresponding to different grade boundaries. The lineranges are 95% confidence intervals. Fully opaque points/lines are estimates across a true boundary, while the transparent points/lines are placebo tests at other adjacent ranks.

that teachers are more generous at grading females at these parts of the distribution, pushing them over the line to a higher grade. The differences are largest at the C/D grade boundary with a 4.7 percentage point gap in the proportion of females marginally achieving a C grade comapred to a D grade. We see no difference in the proportion female at the A* versus A boundary.

The estimates at non-zero cutoffs in figure 4 represent placebo tests. One possible limitation of our method is that rather than capturing bias we are just capturing the characteristic-achievement gradients displayed in figures 3a-3c. In a standard RD design, continuous functional form assumptions would help to account for these relationship between the running variable and the outcome. The assumption with LR is that there is no meaningful gradient between the observations on the LHS and RHS. In our setting however, the use of ordinal rank as a discrete running variable puts this assumption at risk. Therefore to account for this we both control for (by including prior attainment as previously discussed) and test for this underlying relationship. Our test takes the form of placebo tests, where we compare adjacent ranks that do not straddle a grade boundary. We do not observe many significant discontinuities away from the grade boundary. If there was a strong achievement gradient in favour of a particular gender then these estimates would be non-zero. That we only observe non-zero values at the threshold implies that

our main estimates are not primarily driven by gender achievement gradients.

Figures 5 and 6 present the same estimates for the pro-FSM and pro-white bias. The pattern in terms of the relative magnitudes varies by characteristic. Teachers increasingly become more biased against FSM students at lower grade boundaries, with the largest biased against of 1.5 percentage points at the lowest grade threshold (E/D). With regards to race, there is little difference across the three highest boundaries (A/A* B/A, C/B), but teachers are significantly more pro-white (around 4 percentage points) for the lowest two grade boundaries (D/C, E/D).

## 4.1 Grade Boundaries Are Endogenous

Teacher's subjective assessments not only determine the ranking of students, but also where the boundaries occur. In order to avoid marginal cases teachers may wish to create thresholds where there are clear achievement differences in students ability. This in and of itself is would not cause a problem with our approach. However, if ability distribution varies by student characteristics, e.g. the female achievement distribution is to the right of the male distribution, then these achievement gaps will be correlated with student characteristics. So $\tau$ will be a combination of gender bias and achievement differences.

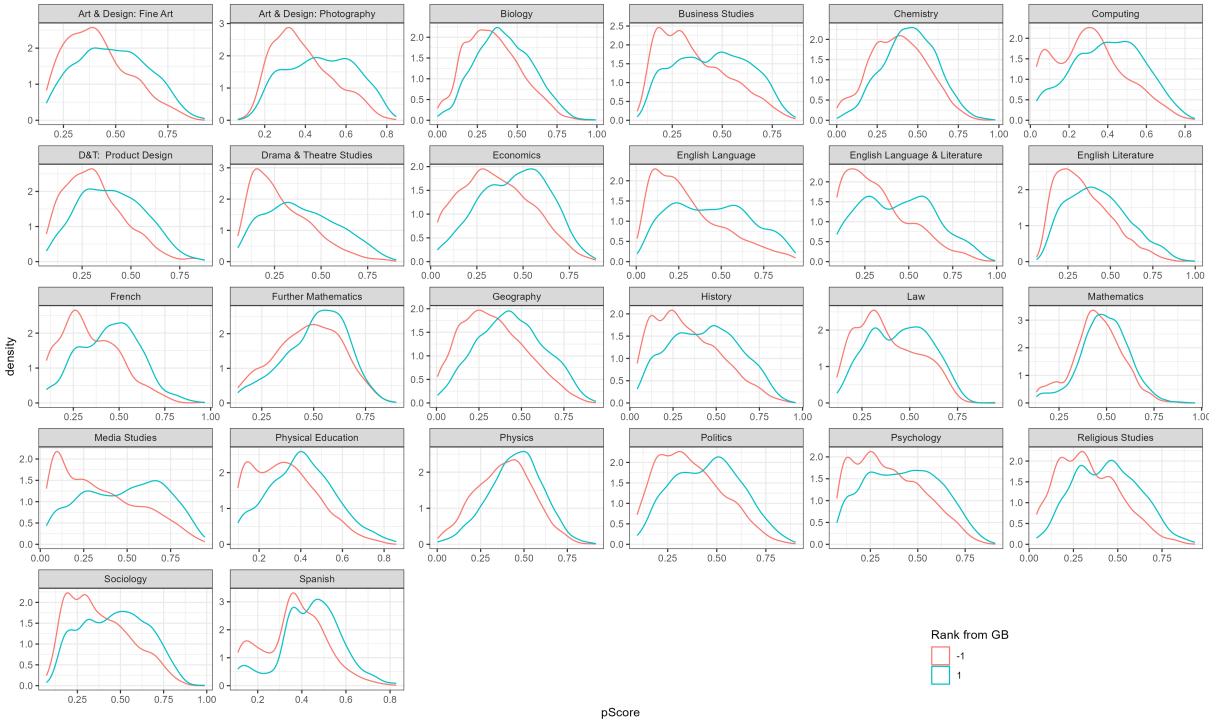
To empirically account for this we have conditioned on a cubic-polynomial of prior achievement, to account for pre-existing characteristics achievement gradients (the proportion of females is increasing in achievement). However if one was concerned that this is too restrictive in terms of functional form assumptions, we propose two alternative measures which make fewer assumptions.

First, we make an addition sample restriction after creating a continuous latent achievement measure. The ranking of students does not contain any cardinal information for how far in terms of absolute achievement the adjacent students are the the threshold. To only consider students who are similar in terms of ability we create a cardinal metric of latent achievement based on their prior achievement in respective subjects.

For each student, we estimate the probability,$\phi_\alpha$, of achieving grade $\alpha$ separately by 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 higher of the grades,$\phi$, 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 Alevels. Panel A of Figure **??** shows the propensity scores of students ranked just above the grade boundary in blue (1) and those ranked just below in red (-1). Even when restricting to students adjacent to the boundary, there are large difference in propensity scores.

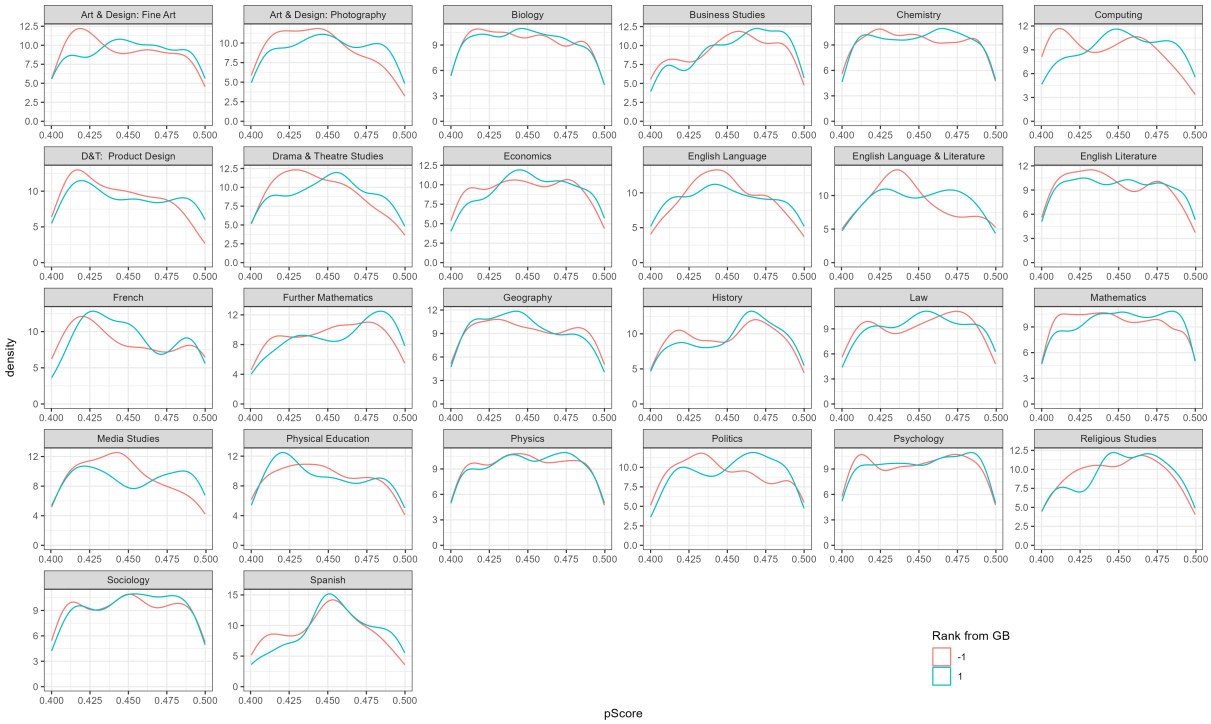
We then restrict the sample to those who have a propensity of achieving grade s $\phi_\alpha(0.45, 0.55)$.

Figure 7: Histogram of Propensity of Attaining the Higher Grade of Boundary Adjacent Students, by Subject Grade Boundaries




*Notes:* The blue line shows the histogram of propensity scores for students one above the grade boundary. The red line shows the histogram of propensity scores for students one below the grade boundary.

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




*Notes:* The blue line shows the histogram of propensity scores for students one above the grade boundary. The red line shows the histogram of propensity scores for students one below the grade boundary. Restricted to students with propensity scores between 0.45-0.55.

(a) Double RD estimates

| *Characteristic:* | Female | FSM | White |
|---|---|---|---|
| RHS of GB ($\tau$) | 0.014* | -0.007** | 0.015** |
| | (.020) | (.003) | (.006) |
| Observations | 20,470 | 20,310 | 18,500 |

*Notes:* $^{*}p < .1$, $^{**}p < .05$, $^{***}p < .01$. The sample includes any school with at least 3 students either side of the boundary. We keep students up to 1 either side, with a propensity of 0.45-0.55 to estimate regression coefficients.

Panel B of Figure **??** shows the propensity of achieving the higher grade for students adjacent on the LHS and RHS with this restriction. This limits the sample to individuals with similar propsencities. We re-estimate $\tau$ with this subsample. We currently restrict this analysis to English and Maths to ensure compatablity between GCSE and Alevels. This reduces our sample from 228,570 observations to 20,468. The corresponding estimates are shown in Table 9a, and are insignficatly different from our main estimates. This imposes fewer functional form assumptions away from the threshold, and by definition there are only small latent achievment gaps.

Our third approach for dealing with endogenous boundaries is to make even fewer functional form assumptions. The main specification contains all grade boundaries with at least one adjacent student. However, if there is 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 $\tau$ with sub-samples with increasing numbers of adjacent students. Coefficient stability implies that $\tau$ not driven by ability This requires no require functional form assumptions between past and present achievement.

## 4.2 Impact on Future Outcomes

A-Levels are a key determinant of which degree courses students are accepted to. In order to establish the consequences of the biases established in the previous section we now establish how these small differences in rankings, resulting in different grades, impact the success of student applications. Instead of estimating the imbalance around the thresholds, we use the LR approach to estimate the impact on university application outcomes. Table XXX shows the impact for those immediately above a grade boundary, on being accepted to any university course, and to be accepted to their most preferred course.

Students with a higher grade in one of the A-level courses are around 3 percentage points more likely to be accepted onto any course, and on to their preferred course. As non-

white students are 1 percentage point under represented above the grade boundary. There are approximate 60,000 non-white students around the boundaries, of these 600 were experianced a lower grade, and so 18 would not be accepted on any university course due to teacher bias. It is important to remember that we are only picking up the micro-bias immediately around the threshold. Any larger biases, which move students multiple ranks are not accounted for.

# 5 Conclusion

This paper establishes the existence of biases in how teachers assigned 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 exactly identify marginal students. Leveraging the discrete nature of ranks we implement a Local Randomisation approach to detect bias in teacher-assigned grades.

We find sizable gaps across a range of grade boundaries in the proportion of students by gender, socio-economic status, and ethnicity. 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 the existing literature, including that teachers tend to be biased towards the gender that is under represented in a subject. Critically, our estimates do not rely on the same assumption that the standard method requires, that the blind (standardised tests) and non-blind (teacher) assessment are attempting to measure the same underlying latent characteristic. If that assumption does not hold, then the parameter of interest with this DiD approach will also contain differences in other characteristics. In contrast our LR approach only has one dimension through which students are measured, so any differences in the concentration of a characteristics around a grade boundary will be due to decisions made by teachers, thereby producing a direct measure of teacher bias. Our results imply that teacher assessments might favour some groups of students over others, and care should be taken to avoid this, either by issuing better guidance and information for teachers or by using externally marked assessment.

# References

Blank, R. M. (1991): "The Effects of Double-Blind versus Single-Blind Reviewing: Experimental Evidence from The American Economic Review," *The American Economic Review*, 81, 1041–1067, publisher: American Economic Association.

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, 53–75.

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.

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

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.

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," .

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

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

Goldin, C. and C. Rouse (2000): "Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians," *American Economic Review*, 90, 715–741.

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," .

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 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.

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

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

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.

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