Systematic gaps in teacher judgement: A new approach

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

2nd October 2024

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 PLEASE DO NOT CITE.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 randomisation. This does not require the standard assumptions used in teacher bias literature. We find evidence of teacher generosity on average favouring higher income, female, white students. However, there is large variation in gender bias across subjects. Teachers tend to favour a gender more the less that gender is represented in a subject.

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 verses standardised testing for students transition to post secondary education. For example, in the US several universities dropped their SAT requirement during Covid-19 and have not reinstated it (Leonhardt, 2024), and Portugal has reduced the weight of final exams in the final classification at the end of high school from 30% to 25% (Portuguese Ministry of Education, 2023). However, there are also those advocating for a move in the other direction. A small number of high-profile US colleges (including MIT and Dartmouth) have *reinstated* admissions tests, amid concerns that GPA scores exacerbate existing disadvantage gaps and recent evidence that standardised tests are better predictors of performance at college than high-school GPA (Chetty et al., 2023). In the UK, there is an ongoing debate over the fairness of the current university admissions system which currently relies on a combination of teacher predictions and standarised examinations, with critics arguing that these grades might favour certain student types (Murphy and Wyness, 2020).[1]

These mixed attitudes towards standardised testing reflect the mixed evidence on the relative merits of teacher-assessed grades versus standardised tests. Performing an analysis of 59 countries using PISA data, Bergbauer et al. (2021) find that the introduction of standardised testing leads to improved student performance in low- and medium-performing countries, while in high-performing countries the expansion of standardised testing appears harmful.

Proponents of teacher-assessed grades argue they allow for broader curricula than can be assessed with externally graded exams, and that are less stressful for both teachers and students than exams, while still correlating very highly with test scores (Rimfeld et al., 2019). Teacher assessments can also measure performance over an extended period and hence are not driven by performance at a single point in time. However, their opponents point to evidence of bias in (subjective) teacher assessments, 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 contribute to the evidence base on bias in teacher-assessed grades. Our analysis exploits the Covid-19 pandemic-induced cancellation of A-level exams in the UK in 2020, exams which were replaced at the last minute by teacher-assessed grades, known as CAGs.[2] A-level qualifications are high-stakes for many students as they determine

[1] The UK Department for Education recently ran a consultation over potential reforms to the university admissions system to move to a post-qualification admissions (PQA) system.

[2] 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,

where (and whether) students will enrol for their university degrees. In addition to assigning CAGs to students, teachers were also asked to rank students within grade and subject within each school.[3] We obtained access to a rich student level dataset (GRADE) which contains the teacher ranking of all students in all subjects that would have taken their A-level examinations in the summer of 2020. We have linked this data with prior measures of student achievement.[4] Our method employs this ranking as the running variable in a regression discontinuity design (RDD) framework, with cutoffs at each grade boundary. We then implement a Local Randomisation (LR, Cattaneo et al., 2024) test to verify whether the concentration of student characteristics is equal across these grade boundaries.

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 with certain characteristics at grade boundaries. For example, while there may be more female students achieving A grades than male students, there should not be a bunching of male students at the top of B grades (and a corresponding "bunching" of female students at the bottom of A grades).

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. For example, if there is only one student with a C grade in a subject, they will be the marginal student for the C/D and B/C grade boundaries. To account for this we condition on prior attainment of each student measured two years previously in standardized examinations.[5] Specifically we implement

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).

[3]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).

[4]We are in the process of linking this with post-secondary outcomes.

[5]We are currently developing this approach to focus on the subset of students with a similar propensity to be around a particular threshold. This involves three steps. First, estimating an ordered probit model of prior achievement on A-level grades for a previous cohort. Second, based on the estimated probabilities

a LR analysis with teacher-assessed rankings, using only the most marginal students at grade-subject boundaries conditional on prior attainment, to test for imbalance in student characteristics.

We use our methodology to compare the underlying share of certain characteristics — gender, ethnicity, Free School and Meals status (FSM) — across grade boundaries. 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, versus what there should be, suggesting generosity in favour of females at this grade boundary). The largest gaps occur at the D/C grade boundary, which is generally considered the pass fail threshold. Moreover, when considering the extent of this bias by subject, we find that teachers tend to favour the gender that is under-represented. Female students are 1.5 percentage points more likely to be ranked immediately above the threshold in maths, while they are 1.2 percentage points less likely in Psychology. Such sizable discontinuities across grade boundaries indicate biases in how teachers assigned these high-stakes grades in 2020.

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 using only grade boundaries with large numbers of students each side, and subject-boundary fixed effects. To address the concern that we are simply picking up a characteristic achievement gradient, we show 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 teachers may reward participation in class discussions, and punish tardiness, neither of which

of being assigned a grade, identify a subset of students in which the covariates in the treatment and control groups are similar. Third, within this subset use the propensity score to weight observations.

would be measured in a standardised examination grade. 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.

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 GCSE and A level exams, it was decided to replace the exams with teacher predicted

Figure 1: Timeline of educational decisions

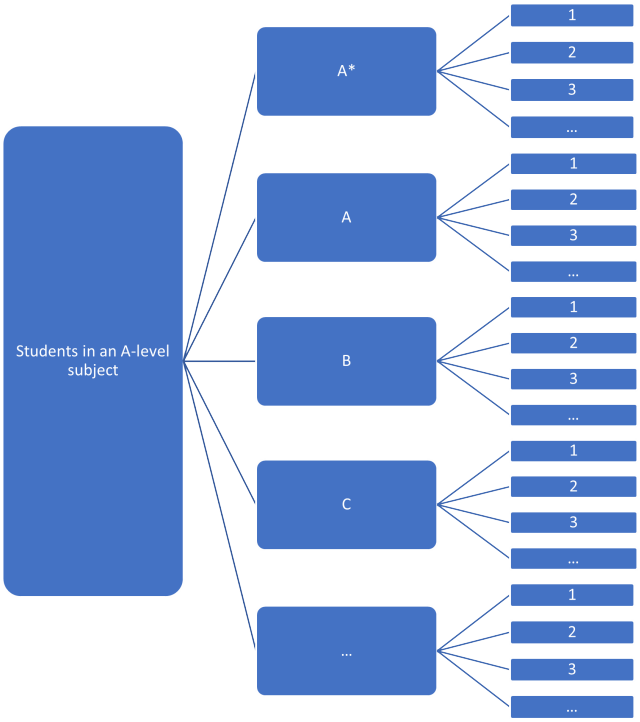
(a) Normal timeline

Year 11 (age 15/16)	Apply for sixth form
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
August – October	Clearing

(b) Pandemic affected timeline in 2020

Year 11 (age 15/16)	Apply for sixth form
End of year 11	Sit GCSEs
Year 13 (age 17/18)	Apply to universities & receive "conditional" offers
March 2020	COVID pandemic - all schools close
June 2020	A level exams cancelled
August 2020	A level results based on CAGs & confirm place at university
August – October 2020	Clearing with CAGs

Figure 2: School subject grading



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. In brief, every school was required to submit teacher assessed grades, known as Centre Assessment Grades (CAGs), and a rank order of their pupils within each grade (e.g. if the student was predicted to achieve an A grade in English, the teacher had to also estimate their rank within all students who received an A grade in English. The students grade and rank was to be provided for every subject). 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:

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

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, X is a cubic polynomial for prior achievement. 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.

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.

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

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.

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

Table 1: Descriptive statistics

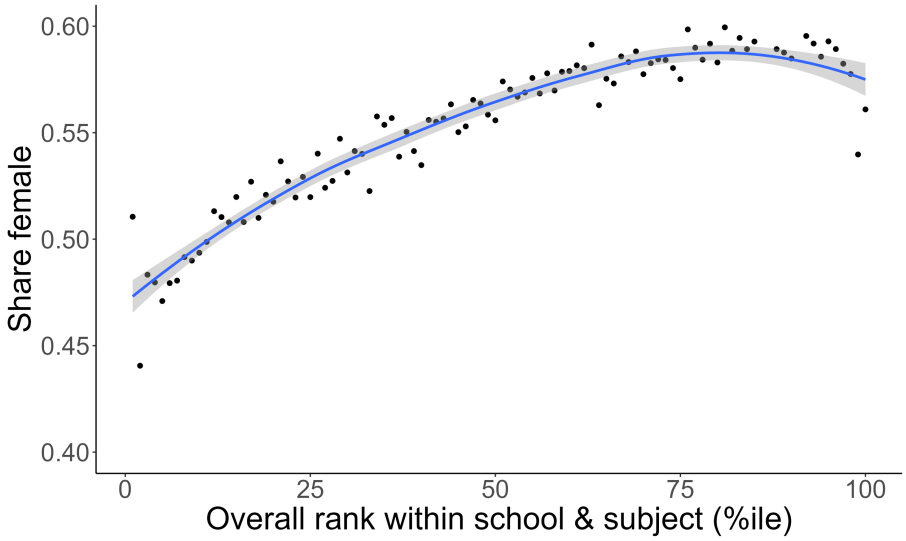
(a) Population share and numbers of students

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

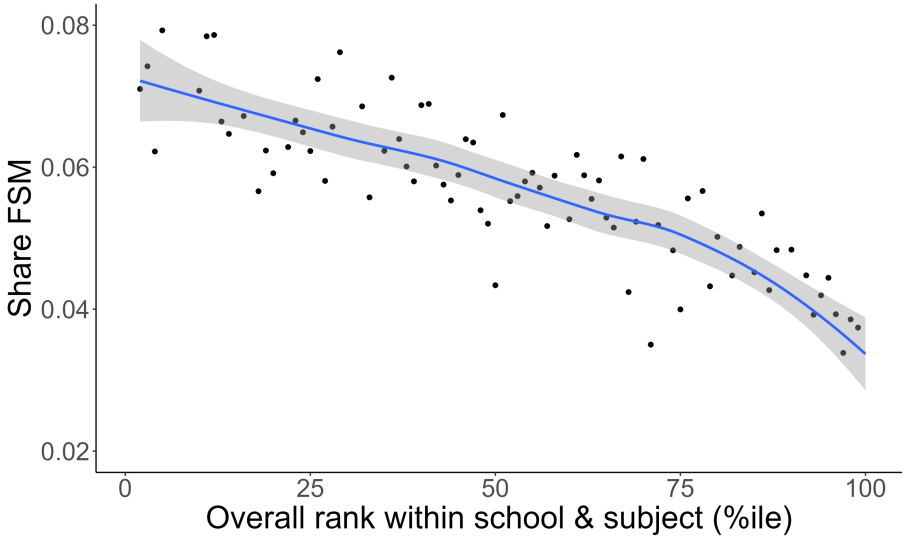
(b) Share and numbers of marginal observations (overall and by subject and grade boundary)

	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

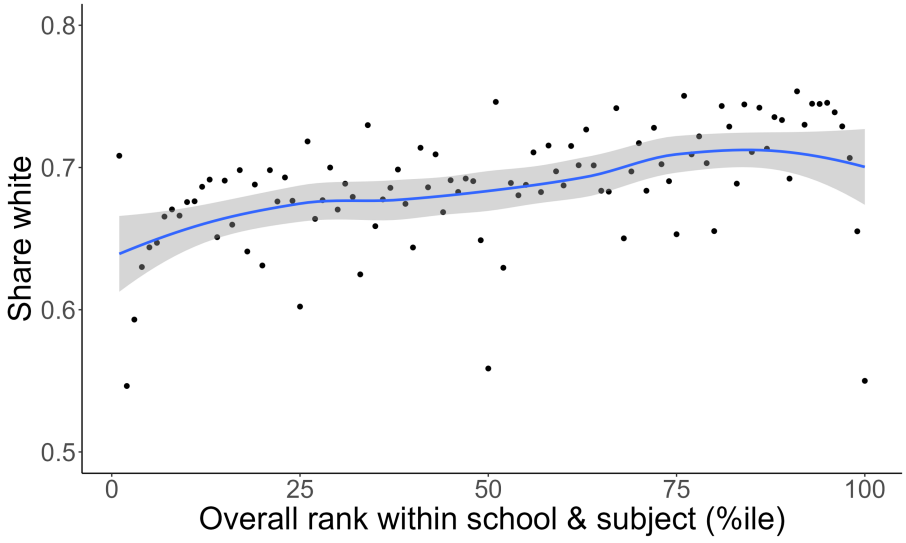
Figure 3: Characteristics by percentile (2020)



(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.[6]

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 2a-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 0.01 implies that across all schools and all subject-grade boundaries females are 1 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 (1.4 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. The inclusion of subject indicator variables makes little difference to the estimated parameters. In contrast the inclusion of the grade boundary fixed effects increase the magnitude implying there are considerable differences in the magnitude of the effect across grade boundaries.

Column 5 conditions on prior achievement, to account for expected distance from the grade boundary. While the magnitude of each coefficients has reduced, they remain significant, and similar in magnitude to the raw estimates.

These estimates are an average across all subject, but teachers bias may vary by subject. Figure 4 presents the gender discontinuity conditional on prior attainment for the ten most popular A-level subjects. Indeed we can see that the bias in favour of female students varies by subject. Females are favoured in mathematics, biology, business studies and physics. In contrast males are favoured in psychology, history, English literature and sociology. However, only one of these estimates is statistically significant at the 5% level (thin red lines), with two more significant at the 10% level (thick red lines). Females are favoured in mathematics and biology, with jumps of 1.5 and 1.2 percentage points, while males are favoured in psychology, with a bias 1.2 percentage points. Figure 5 presents these estimates against the proportion of female students studying each subject. There

[6] 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 (τ)	0.011***	0.010***	0.025*	0.024***	0.008***
	(0.002)	(0.003)	(0.007)	(0.002)	(0.002)
Mean GCSE					0.019***
					(0.001)
Constant	0.530***				0.401***
	(0.001)				(0.006)
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 (τ)	−0.005***	−0.005***	−0.009**	−0.010***	−0.002†
	(0.002)	(0.003)	(0.001)	(0.001)	(0.001)
Mean GCSE					0.017***
					(0.000)
Constant	0.075***				0.191***
	(0.001)				(0.003)
Subject FEs		✓			
GB FEs			✓		
Subject×GB				✓	

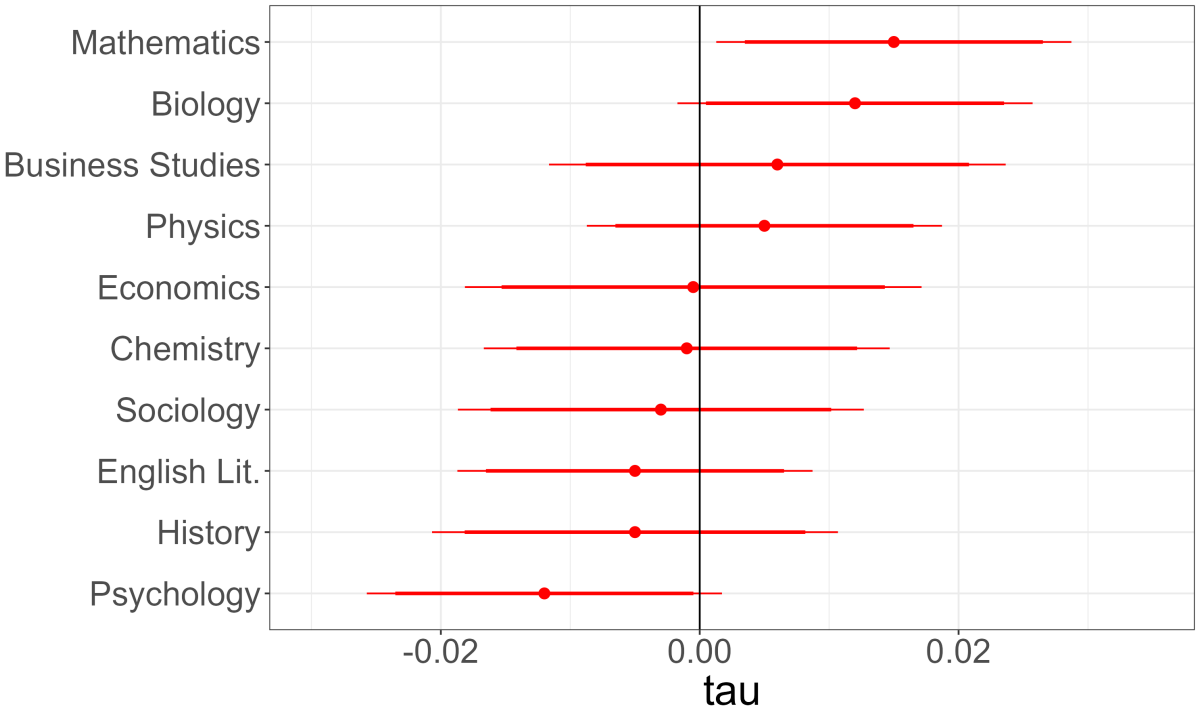
Notes: $^{\dagger}p < .1$, $^{*}p < .05$, $^{**}p < .01$, $^{**}p < .001$. $N = 228,570$ for all specifications. Columns (1)-(5) all estimate variations of equation (1) with no controls in column (1), adding subject and GB FEs in (2) and (3), subject-GB FEs in (4) and controlling for mean GCSEs (our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

(c) Ethnicity

Dep. var.: White	(1)	(2)	(3)	(4)	(5)
RHS of GB (τ)	0.008***	0.007**	0.016*	0.014***	0.006**
	(0.002)	(0.003)	(0.004)	(0.002)	(0.002)
Mean GCSE					0.014***
					(0.001)
Constant	0.746***				0.649***
	(0.001)				(0.005)
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 (τ) by subject

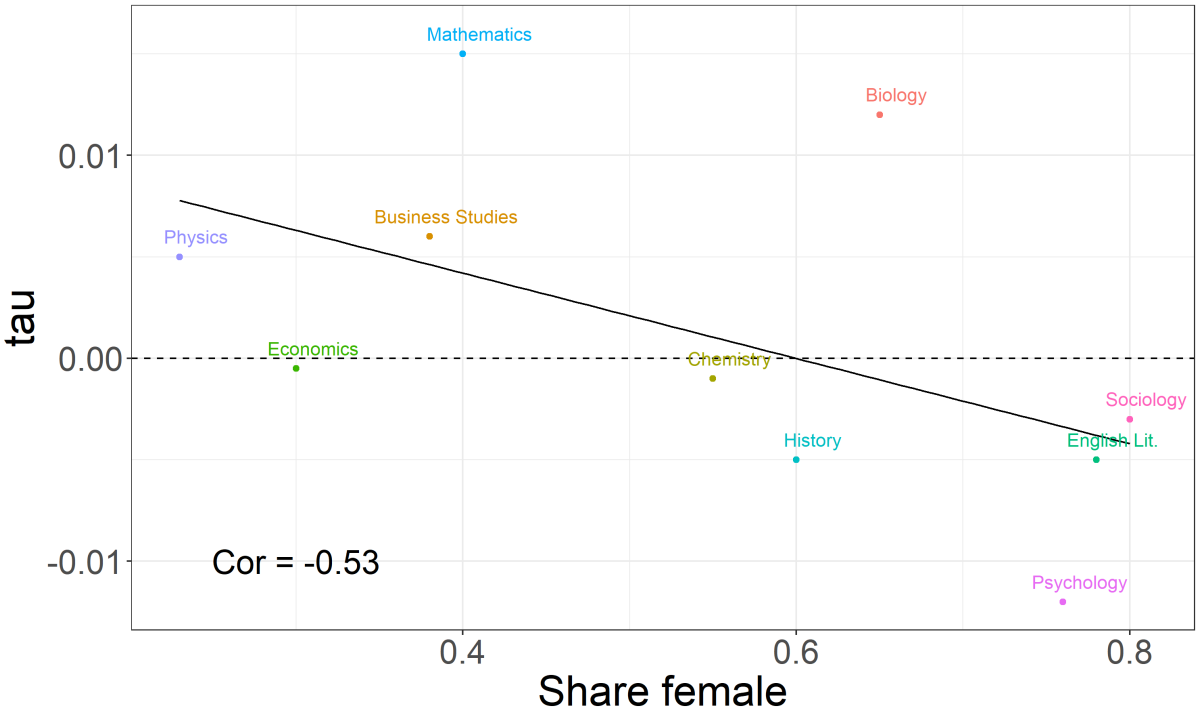


Notes: The red points in the above figure represents estimates of τ obtained by estimating equation (1) separately for each of the top 10 subjects at A-level. The thick and thin red lineranges represent 90% and 95% confidence intervals.

is a negative correlation between the two, indicating that the higher the concentration of a gender in a subject the higher the bias against that gender. This is consistent with Breda and Hillion (2016), who using the difference in performance between non-blind and blind exams find a negative correlation between the degree of male-dominance in STEM fields and the pro-male bias, suggesting examiners are favouring the minority gender in a subject.

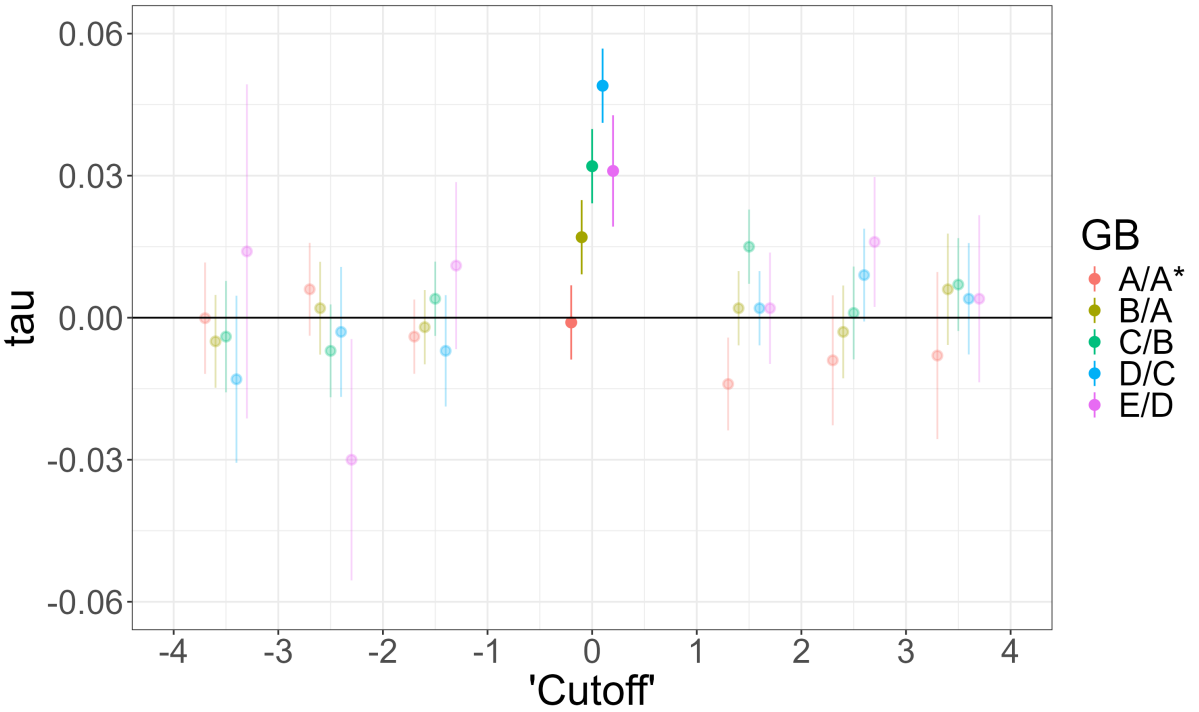
In place of considering the differences across subjects, Figure 6 presents the extent of female bias across grade boundaries, and discontinuities across ranks away from the grade boundary, conditional on prior achievement. Here the x-axis represents the rank-pairs distance from the grade-boundary, e.g. 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 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.

Figure 5: Estimated pro-female bias (τ) versus share female by subject



Notes: The coloured points show the share of female students (x-axis) and estimated pro-female bias (τ, y-axis) in each of the top 10 subjects from figure 4. A solid line of best fit is also displayed along with the correlation of -0.53.

Figure 6: Estimated pro-female bias (τ) by grade boundary (with placebo tests)



Notes: The coloured points show the estimated pro-female bias (τ) 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.

Placebo tests. The estimates at non-zero cutoffs in figure 6 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.

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

Bergbauer, A. B., E. A. Hanushek, and L. Woessmann (2021): "Testing," *Journal of Human Resources*, publisher: University of Wisconsin Press Section: Articles.

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

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*.

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," .

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.