Centre for Education Policy & Equalising Opportunities

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 subjective 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 randomisation approach to estimate imbalance in student characteristics. Our findings reveal systematic bias favoring female and white 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.

REVISED VERSION: July 2026

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.

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 white students 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* Lindsey Macmillan* Richard Murphy† Gill Wyness‡

20th July 2026

Abstract

We propose a new approach to test for systematic biases in subjective 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 female and white 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

Subjective evaluations play a pivotal role in many high-stakes decisions, including hiring and promotion in the labor market (Taylor and Yildirim, 2011; Li and Kahn, 2018), bail and sentencing in the justice system (Ichino et al., 2003; Arnold et al., 2018; Cohen and Yang, 2019; Chyn et al., 2025), and assessments of pupil performance in education (Lavy, 2008; **?**). The prevailing concern with subjective assessment is its potential for bias. Yet there are inherent difficulties in measuring this bias, precisely due to the subjective nature of these evaluations.[1]

Despite these concerns, many countries (including Australia, Portugal, the UK, and the USA) are increasingly relying on subjective evaluation metrics.[2] 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). This removal of standardized test scores has been typically motivated by a desire for equity. However, the presence of bias in the subjective assessments that replace standardized tests is a clear threat to this goal.

But measuring bias in subjective assessment is inherently challenging. Existing studies measuring teacher bias either require primary data collection, limiting the sample size and representativeness (Hinnerich et al., 2011; Hanna and Linden, 2012; Alesina et al., 2018; Carlana, 2019), or rely on assumptions to compare achievements on blind and non-blind assessments (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 has neither of these limitations, and can be applied in many settings including outside of education,[3]

Our approach recovers average grading bias of teachers who perceive students to be of the same abilities. The simple intuition is that students adjacent to grade boundaries are likely to be similar in terms of perceived ability, and this similarity is increasing in the number of students being graded. We compare the average proportion of a student type ranked immediately above and below each grade boundary, and treat any discontinuous change in the representation of a student type as evidence of teacher bias. We apply this logic to a setting where high-stakes end of secondary education (A-level) exams were replaced with teacher assessments due to Covid-19. Instead of sitting standardized exams, students were assigned a letter grade by their teachers in each A-level subject they were taking

[1]Other concerns of subjective evaluations in education include, being less informative of future academic success (Chetty et al., 2023; Friedman et al., 2025), and limiting school accountability systems (Goodman, 2016)

[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.

(typically three). In addition to assigning grades to students, teachers were required to rank students within each letter grade. This provides a ranking of every A-level student by subject and letter grade in each secondary school.[4] We use these rankings to test for systematic biases in high-stakes teacher evaluations across all grade boundaries in England.

Our analysis uses The GRading and Admissions Dataset for England (GRADE) (Ofqual et al., 2025), which covers the population of students who received these subjective assessments including information on their rank, demographics, prior achievement, and applications to college. We find evidence of discontinuities around grade boundaries in favor of female and white students. Teachers grade students such that female students are 2.9 percentage points (5.5%) more likely to be immediately above a grade boundary than below it, and white students are 1.6 percentage points (2.6%) more likely to immediately above a grade boundary than below it.

Discerning whether bias is driven by taste based or statistical discrimination is notoriously difficult. However, we make some steps towards developing a better understanding by considering heterogeneity in bias across subjects. In subjects where prior performance is a strong predictor of future performance (e.g. mathematics) we would expect the magnitude of bias to be smaller if this was driven by statistical discrimination. Similarly, in subjects which historically have a strong positive correlation between type and achievement we would expect the bias to be more in favour of this type (e.g. males and mathematics). Finding neither to be true, we take as suggestive evidence that the bias we observe is not primarily driven by statistical discrimination. In contrast we find no empirical evidence contrary to taste based discrimination.

We show that these biases have direct consequences: being just over a grade boundary threshold increases the chance of being accepted at university by 2.1 percentage points. This is particularly important at the top of the grade distribution, with those just above the threshold 4.6 percentage points more likely to be accepted to their first choice course than those just below the boundary. We find evidence that suggests around 20 percent of the bias is due to 'altruism' by teachers trying to help students applying to university [5]. Students ranked just above a grade boundary are 3.0 percentage points more likely to have applied to university. However, when considering the sub-sample of students adjacent to

[4] The purpose of this was that the Department of Education was concerned about grade inflation. Having both the grade and rank within grade provided a complete ranking of students in each school-subject (ofqual, 2020). The original intention was for this ranking to be the basis of an algorithm that accounted for each school-subject's historical grade distribution and current composition, to compute an adjusted grade (House of Commons Education Committee, 2020b). The adjusted grades were awarded in August 2020. However, after a national outcry the adjustment was scrapped and students were awarded the teacher-assessed as their final grade (House of Commons Education Committee, 2020a).

[5] Dee et al. (2019) also find evidence of altruism among teachers through the manipulation of test scores to prevent students from facing sanctions involved with failing an exam.

grade boundaries who have both applied to college, we continue to find that females are 2.3 percentage points more likely to be given the higher grade.

Our approach is based on the local randomisation literature (Cattaneo et al., 2024) as the ranking of students provides discrete data points rather than the more typical continuous running variable used in the regression discontinuity or bunching literature. This allows us to restrict our main analysis to only students immediately adjacent to grade boundaries – the smallest possible bandwidth. These rankings will be based on teachers latent perception of student abilities. To recover our parameter of interest, average grading bias of teachers who perceive students to be of the same ability, we require this perceived performance to be continuous at the boundary. While this will be true in the theoretical limit, with an infinite class size, we show that in practice our estimates are robust to varying finite class sizes.

The key assumption is that individuals are as good as randomly assigned each side of a threshold. However, as we show that individuals are not randomly allocated across thresholds due to bias, we rely on a weaker assumption, that only perceived ability is as good as random across a threshold. With finite classes there are two main threats to this assumption. First, if there is a strong relationship between student attainment and student type e.g. females generally perform better than males (EPI, 2021). Second, if classes were sparsely populated then students adjacent to the grade boundary may not be similar in perceived ability. In either case differences in representation of a type across a grade boundary may represent bias, or real differences in perceived ability across a boundary.

We present four pieces of evidence to establish that our results are not driven by differences in perceived ability. First, we show that there are no discontinuities of student type in adjacent pairs away from these boundaries, implying that there is not a strong correlation between perceived ability and student type (see Figure 4). Second, we show our estimates are robust to controlling for blind measures of prior attainment, implying that these marginal students are similar in ability. Third, the heterogeneity in the bias across subjects is not consistent with the historical ability gradients by type e.g. there is a strong bias in favour of women in math, despite males being more present at higher ranks. Finally, and returning back to the hypothetical infinite class, we show that our estimates are robust to estimating only using increasingly larger classes, whereby the adjacent students would in expectation be more similar in terms of perceived ability.

Our novel approach to measuring bias makes several contributions to the existing literature. Previous studies 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) due to its ease of implementation. 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. For this to be recovering teacher bias requires three assumptions that have empirically been called into question; 1) that both sets of assessment are measuring the same ability (Hirnstein et al., 2023); 2) that both assessment types have the same mapping of ability to assessment performance by student type (Cai et al., 2019; Galasso and Profeta, 2024; Arenas and Calsamiglia, 2025); and that they have the same expected measurement error by student type (Zhu, 2024; Delaney and Devereux, 2025). Our results confirm much of the findings of this 'indirect approach' literature (Lavy, 2008; Lavy and Sand, 2015; Terrier, 2020; Burgess et al., 2022; Graetz and Karimi, 2022; Lavy and Megalokonomou, 2024), without requiring these assumptions relating two forms of assessment – since our approach only requires non-blind assessment.

Our paper is also closely related to research measuring the extent of test score manipulation by teachers and the consequences of manipulation, using a bunching approach. Diamond and Persson (2016) document manipulation in Sweden, by exploiting discontinuities in a (teacher assessed) test score distribution. By assuming the 'true' distribution is smooth, they identify where manipulation occurs and how far below the threshold students get bumped up. Dee et al. (2019) show manipulation in the New York Regents Examinations, and predict counterfactual distributions of achievement by fitting flexible polynomials in the region around the cutoff. Neither paper finds evidence that teachers selectively inflate based on student type (gender and ethnicity respectively). In contrast to these papers, the discrete nature of our ranking means that by focusing on only the most marginal students, we do not need to make assumptions regarding the entire distribution, or the range of potential manipulation.

Finally this paper provides some suggestive evidence on the nature of the bias. The patterns of the heterogeneity observed across subjects are not consistent with simple models of statistical discrimination, while the the constant bias in favour of White and female students across grade boundaries is consistent with taste based discrimination. This bias remains when only considering applicants to college, implying that the favoritism is not primarily driven by altruism. It is important to note that our approach only identifies grading bias among students that the teacher perceives to be similar in ability, in that respect one might consider it a lower bound. Our estimate will not capture any teacher bias that moves all of a student type throughout the perceived ability distribution.

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 defines our parameter of interest and sets out the empirical approach in more detail. Section 4 describes our dataset. We present and discuss our results in section 5, and demonstrate the robustness of our results in section 6. Section 7 concludes.

2 Background and context

2.1 Institutional context

In the last two years of secondary education, Years 12 13, the vast majority of students study towards A-levels, which are the main qualification for admittance to university.[6]. 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.[7] 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.

2.2 Changes due to the COVID-19 pandemic

The COVID-19 pandemic meant that the majority of schools closed in March 2020 and did not reopen for exams. In June it was decided that all standardised exams in the UK would be canceled, including A-levels (See figure 1b). Given the importance of A-level qualifications, the government decided to replace the exams with teacher assessed grades, so that students would obtain qualifications that were informative of their ability.

[6] 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.

[7] 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). Our estimates of teacher bias will indicate how biased these predicted grades are likely to be.

Figure 1: Timeline of educational decisions

(a) Normal timeline

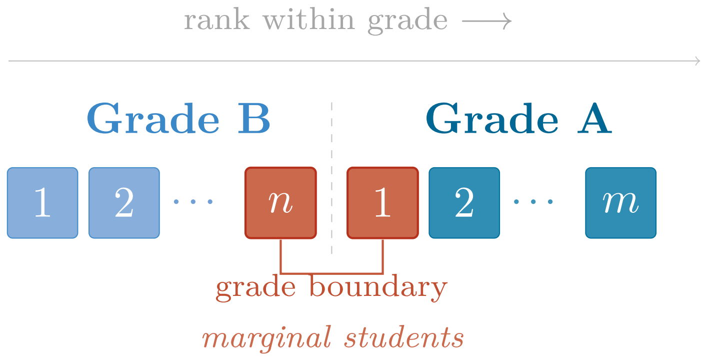
Year 13 (age 17/18)	Apply to universities & receive "conditional" offers
June of year 13	Take A level exams
August after year 13	Receive A level results & confirm place at university

(b) COVID-19 pandemic affected timeline in 2020

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

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.

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



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.[8] Although ultimately the student ranks within grade were never used, these rankings are strongly correlated with students prior performance and so we believe that teachers were subjectively ranking students according to their perceived abilities.

3 Empirical approach

In our setting student's A-level grades were entirely determined by the subjective evaluations of their teachers, who decided their grade and their rank within each grade. Our empirical strategy leverages these rankings, testing for imbalances in the concentration of student types across grade thresholds.

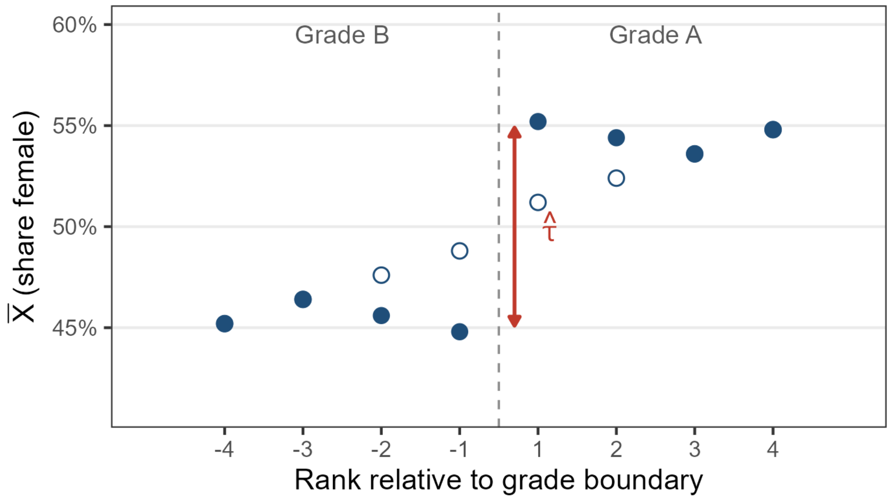
The intuition behind this approach is straightforward: without bias the concentration of student types around each grade boundary should be continuous. We illustrate this intuition in Figure 3, which shows a stylized example with the share of female students by distance from the grade boundary (where -1 equals the rank immediately below the grade threshold, and 1 equals the rank immediately above the grade threshold). The unfilled series represents a case of no teacher bias, the share of females is increasing in rank and there is no discontinuity at the boundary. The filled series shows what we would see in the case of bias — a discontinuity in the proportion of females ranked just above versus just below the grade cutoff. We define the difference in shares of student types of adjacent students either side of the boundary as τ, which we estimate directly.

We then show that our estimate τ is proportional to the average grading bias of teachers Δ, defined as the difference in probability of receiving the higher grade across types of the same teacher-perceived ability. Moreover, we show that under our stringent sample restrictions (only immediately adjacent students, and only boundaries with students on either side) these two objects are equal.

However, as we only observe teacher-assigned grades and rankings and not teacher-perceived ability, we require some additional assumptions to ensure identification of Δ. Our empirical strategy compares the share of student type $X_i \in \{0,1\}$ (e.g. female, white)

[8] The DfE intended to use the rank information, along with student prior achievement, and previous school-subject achievement distributions to monotonically 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.[9] 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.

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.

immediately above and below each grade boundary b. Even with the unbiased series in Figure 3 there is a difference in $\bar{X}$ for adjacent students, our assumption will eliminate this difference in cases with no discrimination. This section provides a formal definition of teacher discrimination Δ, establishes its relationship to our estimated coefficient τ, and states the condition under which $\tau \neq 0$ identifies bias.

3.1 Relating our estimated coefficient τ to "bias" Δ

At each populated school-subject-grade boundary b_{jsg}, we focus on the two most marginal students: the student ranked last in the higher grade and the student ranked first in the lower grade. Let $D_i \in \{0,1\}$ indicate whether student i is assigned to the higher grade at their boundary ($D_i = 1$) or the lower grade ($D_i = 0$). As described in Section 2, both the grade assignment and the within-grade ranking are teacher-determined. The quantity D_i therefore reflects the joint outcome of both decisions, and may be a function of X_i if the teacher discriminates.

We define bias formally as the difference in probability of being assigned ($D_i = 1$) across students of different type $X_i \in \{0,1\}$ who have the same perceived performance δ. This writes as

$$\Delta = \Pr(D = 1 \mid X = 1, \delta) - \Pr(D = 1 \mid X = 0, \delta). \tag{1}$$

where $\Delta \neq 0$ reflects a dependence of D on X and implies bias with respect to X.

However, in practice we only observe a student's rank r relative to a grade boundary b_{jsg}. To recover an empirical equivalent we implement a Local Randomisation RDD (LR) balance test (Li et al., 2021), only using students immediately adjacent to the boundary, i.e. with $r_i = \pm 1$.

We operationalise the test by estimating τ by OLS via the following regression

$$X_i = \beta_0 + \tau D_{ib} + \varepsilon_{ib}, \quad \forall i \text{ s.t. } r_i = \pm 1 \tag{2}$$

where τ represents the percentage point difference in the share of a characteristic on one side of the boundary compared to the other. We are implicitly conditioning on perceived ability (δ) here by including only immediately boundary-adjacent students.[10] Therefore

$$\tau = \Pr(X = 1 \mid D = 1, \delta) - \Pr(X = 1 \mid D = 0, \delta). \tag{3}$$

We show via Bayes' rule in Appendix A, that

$$\tau = \frac{\Pr(X = 1 \mid \delta) \Pr(X = 0 \mid \delta)}{\Pr(D = 1 \mid \delta) \Pr(D = 0 \mid \delta)} \Delta. \tag{4}$$

This result establishes that τ is proportional to Δ as $\Pr(X = 1 \mid \delta), \Pr(D = 1 \mid \delta) \in (0, 1)$. Therefore our estimated τ is a valid sign test for Δ. Moreover, in our restricted sample of only immediately adjacent students to populated boundaries, we have:

$$\Pr(D = 1 \mid r_i = 1, \delta) = \Pr(D = 1 \mid r_i = 1, \delta) = \frac{1}{2} \tag{5}$$

at each boundary. In addition, only grade boundaries where adjacent students are of different types ($X_{RHS} \neq X_{LHS}$) contribute variation in X, so

$$\Pr(X = 1 \mid r_i = \pm 1) = \Pr(X = 0 \mid r_i = \pm 1) = \frac{1}{2} \tag{6}$$

at each boundary.

Therefore in our setting

$$\tau = \Delta. \tag{7}$$

3.2 Identification of τ and Δ

The substantive question is whether a nonzero Δ, the difference in the probability of receiving the higher grade by student type, can be interpreted as teacher discrimination, or whether it simply reflects underlying differences in ability between groups.

To address this, define the counterfactual grade assignment D_i^* as the assignment student

[10] In our application we impose an additional restriction using only students at "crowded" boundaries for our main estimates.

i would receive in the absence of bias, i.e. if the teacher graded students purely on ability $\Delta^* = 0$ The corresponding unbiased analogue of Δ is:

$$\Delta^* \equiv \Pr(D^* = 1 \mid X = 1, \delta) - \Pr(D^* = 1 \mid X = 0, \delta).$$

Assumption 1 (Local balance in potential outcomes)**.** *Conditional on perceived ability δ, the counterfactual assignment D_i^* is independent of student characteristic X_i for students adjacent to the boundary:*

$$\Delta^* = 0, \quad \forall i \textit{ such that } r_i = 1. \tag{8}$$

Assumption 1 states that two students with the same perceived ability who are ranked adjacent to a grade boundary have the same probability of receiving the higher grade in the absence of discrimination, regardless of their characteristic. Under the null of no discrimination, $D_i = D_i^*$ and hence $\Delta = \Delta^* = 0$; therefore any observed $\Delta \neq 0$ (equivalently, $\tau \neq 0$) arises from discrimination.

The assumption is motivated by a limit argument applied to the rank ± 1 pair specifically. Consider a school-subject group sj with a large number of students N_{js}, and one grade boundary. Teachers rank students on perceived ability and then draw a grade boundary within this ranking. As the number of students grows, the two students ranked immediately adjacent to the boundary converge in perceived ability δ: in the limit $N_{js} \to \infty$ they are observationally identical to the teacher. Under no discrimination, which side of the line a student falls on is then orthogonal to their characteristic, so $\Delta^* = 0$. The same logic can be applied to multiple grade boundaries with the number of students adjacent to both sides of any school subject grade boundary $n_{jsg} \to \infty$.

In practice, classes are small (typically $N_{js} = 30$), so the limit applies imperfectly. In our analysis we address this in two ways. First, we restrict our analysis to school-subject-grade boundaries with at least four students ranked in each adjacent grade $n_{jsg} \geq 4$ (approximately 70% of boundaries), which we refer to as the crowded sample. This is assigning zero weight to boundaries where adjacent students may differ in perceived ability. Second, for this subsample we estimate equation by weighted least squares (WLS), weighting each observation by the number of students in the school-subject N_{js}. Weighting by N_{js} is therefore a continuous analogue of the crowded-sample restriction (at least four students per adjacent grade), down-weighting boundaries where the identification assumption is least credible.[11]

[11] This will apply the same weights to all boundaries within a class. We use N_{js} rather than n_{jsg} for this weighting as n_{jsg} is endogenous to teacher bias. Results are robust to using n_{jsg}. We do not use N_{js} to define the crowded sample as we want to ensure there is a minimum n_{jsg} around every grade boundary.

Estimating τ requires variation in X around a grade boundary. Certain structural features limit our ability to detect bias. There is sorting into schools and subjects by student type; as an extreme example consider single-sex schools which will only have one gender type. In these situations our ability to measure bias will be attenuated. To account for this, in addition to weighting by N_{js}, we also weight each school-subject cell js the inverse Bernoulli variance of the type indicator, $\bar{X}_{js}(1-\bar{X}_{js})$. This means our final weights are

$$w_{js} = N_{js} \cdot \bar{X}_{js}(1-\bar{X}_{js}) \tag{9}$$

Intuitively, these weights reflect that we trust evidence of bias the most from cells where $\bar{X}_{js} = \frac{1}{2}$. Cells where $\bar{X} \to 0$ or 1 contain almost no variation in X and are down-weighted accordingly; in the limit they are dropped entirely, corresponding exactly to the "single-sex" and "all-FSM/non-FSM". We compute $\bar{X}_{js}$ over all students in N_{js} rather than for the n_{jsg} at each boundary. A teacher who discriminates may place boundaries where the favoured group clusters, making the boundary-level characteristic share a function of the bias being tested. The class level $\bar{X}_{js}$ is predetermined with respect to the teacher's grading decisions.

3.3 Interpretation of τ

Our estimand τ captures *marginal discrimination at grade boundaries*: teachers' tendency to disproportionately award the higher grade to students of a given type when the grading decision is close. A positive τ for female students, for instance, implies that teachers favour female students at grade boundaries, awarding them the higher grade more frequently than equally achieving male students.

An important consideration to bear in mind is that τ does not capture all forms of teacher discrimination. If teachers systematically rank students of a given characteristic higher — for example, because a teacher believes girls are more able and consistently places them above equally achieving boys — this *macro-bias* shifts the entire within-grade distribution and produces no discontinuity at the grade boundary. Our estimand identifies only the *micro-bias* at the decisive moment of grade assignment. If rankings are also biased, τ understates the full extent of discrimination and should be interpreted as a lower bound on total teacher bias.

The placebo evidence in Figure 4 is consistent with this interpretation: we find no systematic characteristic gradient at rank pairs away from grade boundaries, suggesting that within-grade rankings are, on average, relatively free of characteristic-based bias. The discrimination we identify is concentrated at the grade boundary itself.

4 Data

To exploit the ranking of students within grades we use the linked-administrative GRADE dataset (Ofqual et al., 2025). GRADE was set up to allow researchers to study the events of 2020 and combines several administrative datasets that have previously not been linked. GRADE contains pupil-level data on age 16 (GCSE) and age 18 (A-level) attainment from the Office for Qualifications (Ofqual), pupil and school characteristics from the Department for Education (DfE)'s National Pupil Database (NPD), and pupils' university applications (where applicable) from the University and College Admissions Service (UCAS). This includes students in both public and private schools. Crucially, the dataset contains the CAGs and teacher rankings awarded in 2020, for every student in each subject, grade and school. 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).

Our student types for which we determine teacher bias are: gender (female/male), ethnicity (White/non-White), and Free School Meal (FSM) eligible.[12] We treat FSM as an indicator for socio-economic status. In the UK, around 7% of A-level 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 total number 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, ethnicity, and FSM status in our analysis sample.[13] Around 56% of students are female, representing the greater tendency of female students to pursue A-levels.[14] Only 7% are FSM eligible and around 71% are White. The shares of these characteristics in the analysis 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, ethnicity, 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

[12] FSM results are presented in Appendix C. The FSM estimate is statistically insignificant under the crowding restriction, likely reflecting limited power for this smaller subgroup.

[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/

side of the grade boundary; (v) we apply a crowding restriction, keeping only observations from grade boundaries where there are at least 4 students on each side $n_{jsg} \geq 4$; (vi) our weighting procedure means we drop any observations in homogeneous school-subject combinations with respect to our dependent variables, gender, White ethnicity, and FSM. Finally in panel (c), we present shares and numbers of observations in our analysis sample, broken down by subject and grade boundary.

5 Results

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

5.1 Main Estimates

Table 2 presents weighted estimates averaged across all subjects and grade boundaries.

The table shows weighted estimates of τ, reflecting average differences in the proportion of each student type across all school-subject-grade boundaries, b_{sjg}. The coefficient of 0.029 in column 1 implies that, females are 2.9 percentage points (p.p.) (5.5%) more likely ranked on the Right Hand Side (RHS) of a boundary compared to the Left Hand Side (LHS). White students (column 2) are 1.6p.p.(2.6%) more likely to be given the higher grade. The FSM estimate (column 3) is negative but statistically insignificant. This may reflect lack of bias, limited statistical power, or the inability of teachers to identify FSM students. The remainder of the paper will focus on gender and ethnicity, but the equivalent set of results for FSM are reported in Appendix C.

Despite the use of the crowded subsample and weighting grade boundaries by n_{jsg}, a concern may remain that τ is capturing the underlying characteristic-achievement gradients e.g. females are increasingly represented at higher ranks. 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 $R = -1$ and $R = 1$ relative to the grade boundary — and 1.5 represents students ranked $R = 1, 2$ 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 2.7 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

Table 1: Descriptive statistics

(a) Numbers of students and population shares

Dependent variable (X):	Female	FSM eligible	White
Share	0.558	0.065	0.714
Number of students	237,037	210,013	210,013

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

	No. of obs.	% female	% FSM	% White
w/ info. on all X's	961,105	0.56	0.06	0.70
– missing GCSE	926,563	0.56	0.06	0.71
– ranked 2+ from GB	198,051	0.55	0.07	0.72
– unbalanced GBs	172,338	0.55	0.06	0.74
– GBs < 4 both sides	47,432	0.56	0.06	0.71
– single-sex cells	43,564	0.54	–	–
– all FSM/non-FSM cells	32,774	–	0.08	–
– all White/non-White cells	42,836	–	–	0.70

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

	Female		FSM		White	
	N	Share	N	Share	N	Share
A/A*	27,912	0.58	16,556	0.09	25,530	0.73
B/A	37,962	0.57	22,188	0.12	33,024	0.70
C/B	38,968	0.54	23,038	0.14	33,010	0.68
D/C	28,364	0.49	17,728	0.14	24,100	0.66
E/D	12,674	0.45	8,584	0.14	11,154	0.63
Biology	3,870	0.62	6,416	0.11	9,152	0.65
Business Studies	2,356	0.38	3,794	0.10	5,614	0.70
Chemistry	2,914	0.53	5,600	0.11	8,042	0.59
Economics	1,848	0.28	3,182	0.12	4,940	0.61
English Literature	2,586	0.77	5,690	0.13	7,456	0.69
History	2,958	0.56	5,306	0.12	7,362	0.74
Mathematics	5,204	0.37	7,904	0.09	10,324	0.66
Physics	2,112	0.20	4,166	0.11	7,176	0.67
Psychology	4,902	0.73	8,044	0.11	9,738	0.71
Sociology	2,802	0.75	6,014	0.13	7,000	0.67

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, FSM eligible, or White. 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. The crowding restriction (row 5) keeps only observations from grade boundaries with at least 4 students on each side. Rows 6–7 show the estimation samples after additionally dropping school-subject cells that are homogeneous with respect to FSM or White ethnicity. Panel (c) breaks these totals down by grade boundary and subject (top 10 most popular).

Table 2: Discontinuous change (τ) in student types at grade boundaries

Dependent variable (X_i):	Female (1)	White (2)	FSM (3)
τ	0.029	0.016	-0.006
	(0.005)	(0.005)	(0.004)
Constant	0.524	0.619	0.117
	(0.003)	(0.003)	(0.002)
N	43,564	42,834	32,772

Notes: This table presents the results of our main specification (equation 3.2) pooling observations across all subjects and grade boundaries, using the crowded sample (at least three students each side of every grade boundary). We estimate a weighted OLS weighting observations by how close the share of X is to 50% within a school-subject. Standard errors are in parentheses. The FSM estimate is negative but not statistically significant under the crowding restriction; see Appendix C.

gradients. Although not statistically significant the negative estimates below the cut off reflect that there are fewer females ranked immediately below a boundary compared to further into the lower letter grade, potentially reflecting where the positive mass of women at 0 are coming from, as illustrated in figure 3.

The second panel of figure 4 shows estimates for White students. The estimate at zero is the estimate from Table 2, reflecting that White students are more likely to be given the higher grade. The estimates at ranks away from the threshold are statistically indistinguishable from zero, confirming that the effect is concentrated at the grade boundary. Teachers are setting grade boundaries to ensure that marginal white students are 1.5pp more likely to receive the higher grade.[15]

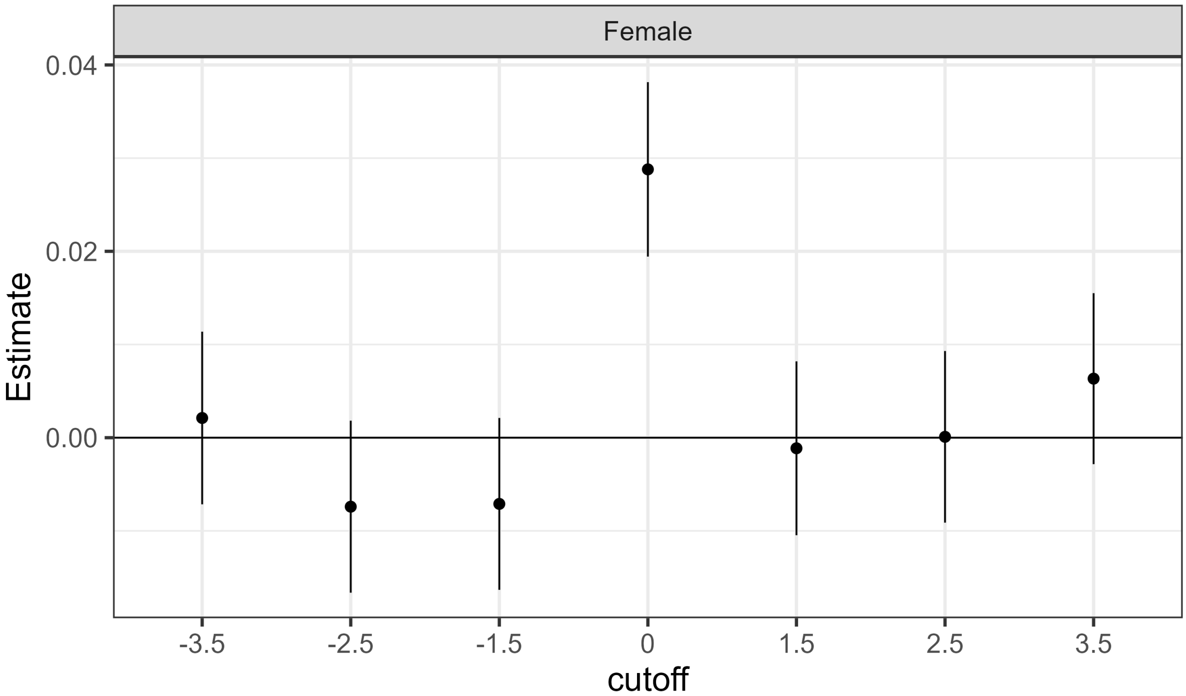
5.2 Discussion of mechanisms

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).

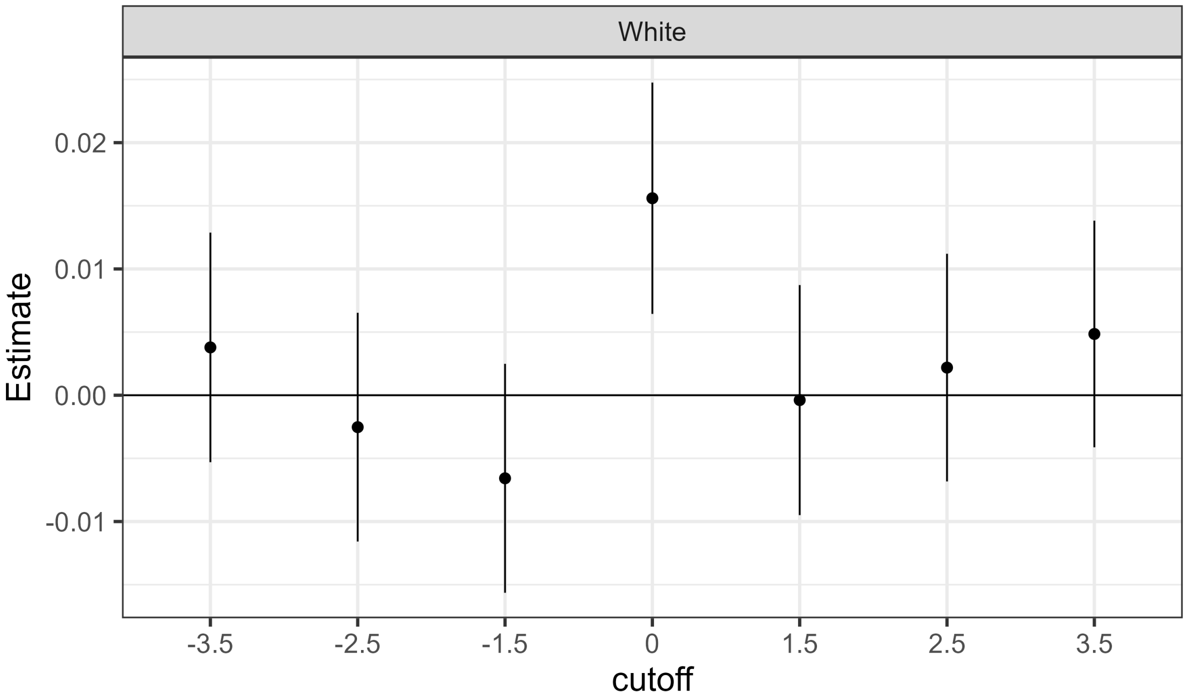
[15] FSM estimates are presented in Appendix C. The FSM point estimate is close to zero and imprecisely estimated under the crowding restriction, likely reflecting limited statistical power given the small share of FSM students in the crowded boundary sample.

Figure 4: Placebo tests by distance from Grade Boundary

(a) Female



(b) White



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.

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.

We present the extent of bias by grade boundary in Figure 5. Each estimate represents the change in the share female (Panel A) or White (Panel B) for adjacent students at each grade boundary.

For females we observe clear and consistent 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 95% confidence intervals for the E/D boundary are considerably wider as few teachers assigned a grade of E to students during the first few months of Covid-19. A similar pattern occurs for White students across all grade boundaries, albeit with only the C/B boundary being statistically significant at traditional levels. The homogeneity of these estimates across grade boundaries is consistent with a taste based discrimination in favour of female and White students.

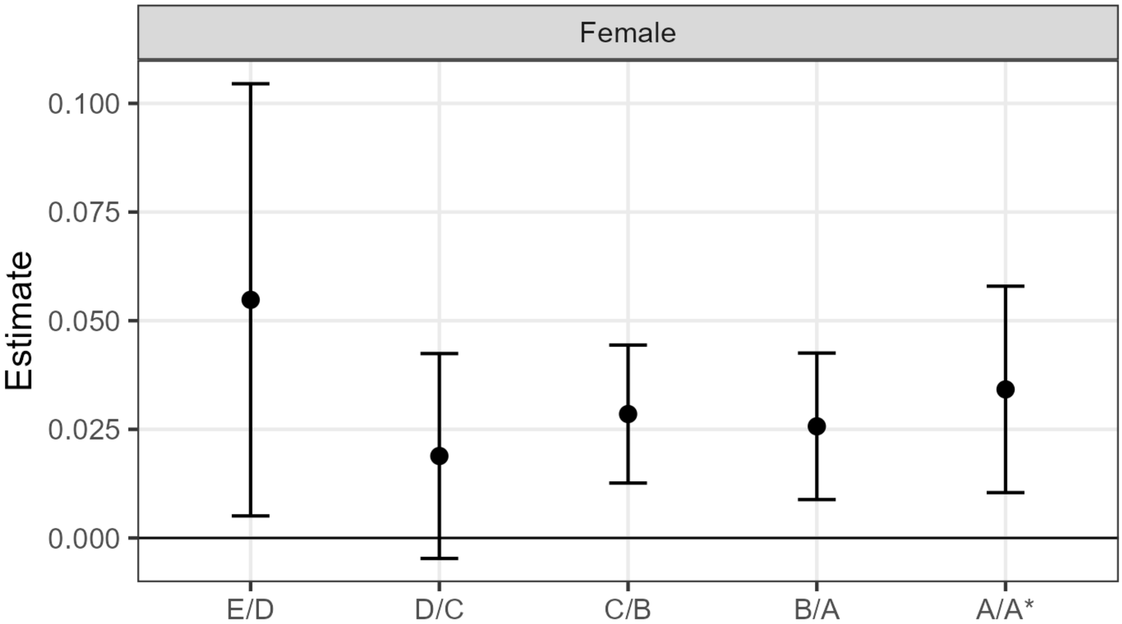
Estimates presented so far are an average across all subjects, but teachers bias may vary by subject (Breda and Hillion, 2016). Figure 6 shows there is considerable variation in τ by A-level subject. There is no subject for which males or non-White students experience a positive bias. The bias in favor of females is the largest in English, Mathematics and Biology, while for White students the bias is largest for Art, English and Computing. With the exception of English, there appears to be little correlation in the degree of bias across types of students.

This considerable variation in τ across subjects could be driven by teachers with different preferences sorting into different subjects. Similarly this heterogeneity in bias across subjects could also be driven by statistical discrimination if the was variation in teachers beliefs about student abilities by type across subjects. To explore the potential for statistical discrimination we quantify how informative prior measures of achievement are, for each subject. The intuition is that for subjects with low informativeness, teachers may rely more on stereotypes or group averages and in subjects where prior test scores are highly informative then there is less room for statistical discrimination as an explanation. Conversely, a key feature of simple taste-based discrimination is that it is unresponsive to information. 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 the A-level cohort prior to our main sample (2019), which was assessed externally at both age 16 and age 18.

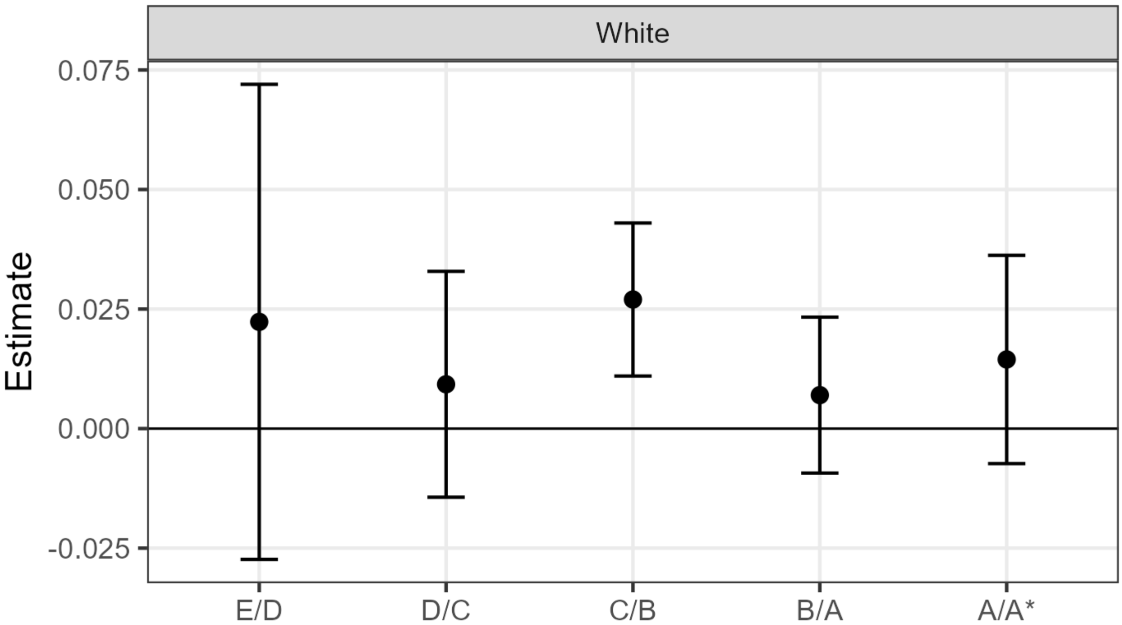
In Figure 7 we plot $|\tau|$, the absolute value of our estimates by subject, against these R-squared measures (our rationale for using the absolute value of τ is that we are interested

Figure 5: Estimated bias (τ) by grade boundary

(a) Female



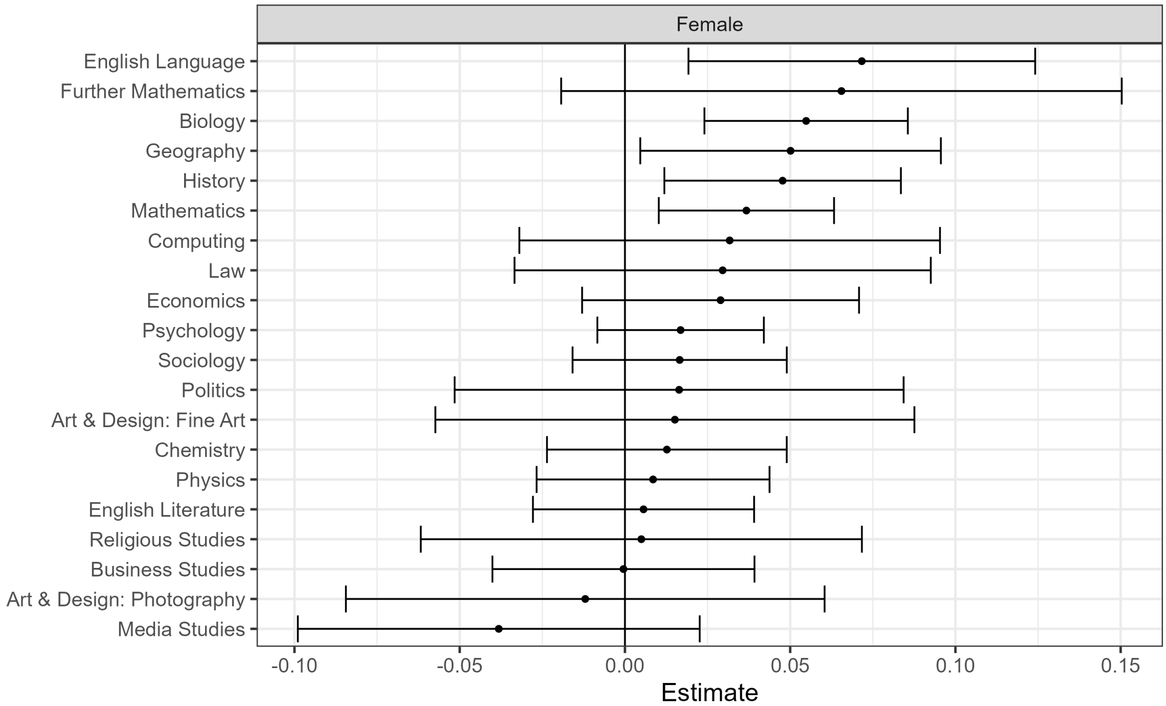
(b) White



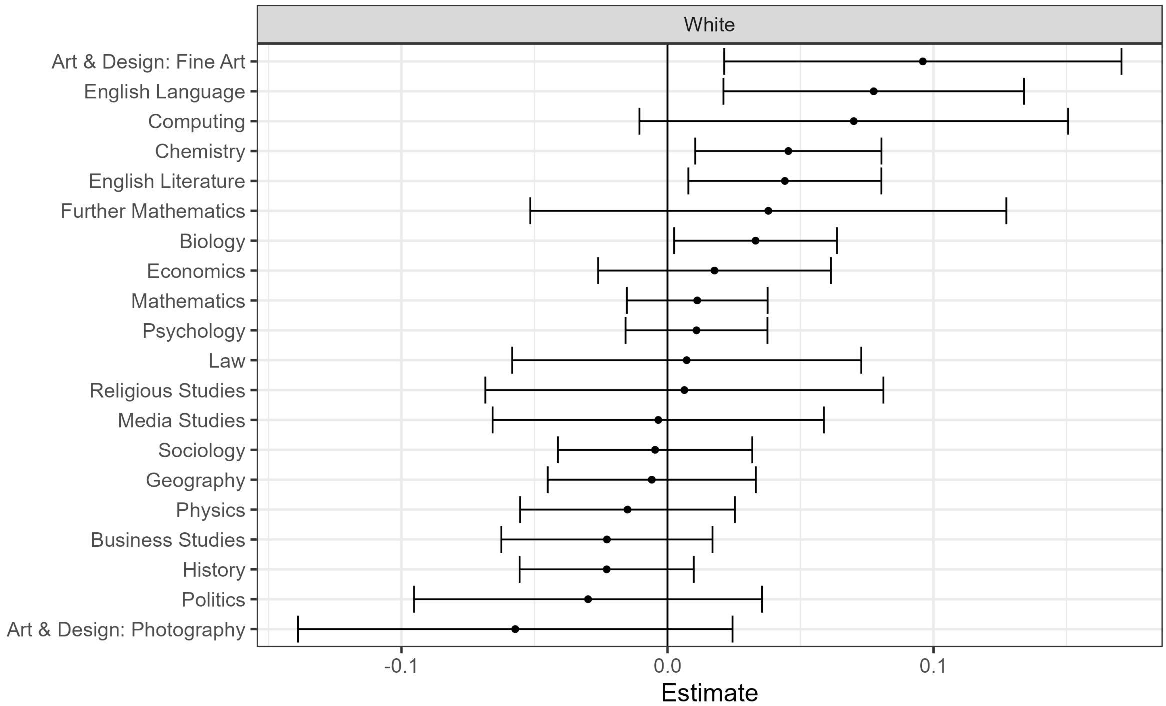
Notes: The points show the estimated pro-female (a) / pro-White (b) bias (τ) at different grade boundaries. The extending bars represent 95% confidence intervals. All estimates are conditional on prior attainment and weighted by how close the share of X is to 50% and the number of students within a school-subject.

Figure 6: Stacking GBs by subject

(a) Female



(b) White



Notes: The points show the estimated 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.

in the extent of bias, not the direction). Subjects with the most discretion (lowest R-squared) are Media Studies, English, and History, while Maths, Biology and Chemistry have the least. We observe a weak non-statistically significant negative correlation across subjects between R-squared and absolute bias for both ethnicity and gender (-0.16, -0.15 respectively). The lack of a significant relationship between information and bias is consistent with taste based discrimination.[16]

The other aspect required for statistical discrimination to be consistent with the empirical results would be for students types to experience greater positive bias in subjects where they historically are higher achieving. In Figure 8 we plot our by-subject bias estimates against achievement-type gradients in those subjects, measured by the proportion of a type $\bar{X}$ against within school-subject percentile (the underlying data for these gradients are shown in appendix B.4). Statistical discrimination would predict a strong positive gradient, however we observe a weakly negative gradient for gender, and an effectively flat gradient for ethnicity. For example for mathematics, females experience a positive bias despite having a negative achievement gradient. Again this is implying that the nature of the teacher bias is not correlated with information.

Considering the variation in τ holistically (i.e. in favour of females and White students), there is weak evidence that the heterogeneity across subjects is due to statistical discrimination and some evidence against it. The homogeneity across grade boundaries is consistent with taste based discrimination.

5.3 Impact on Future Outcomes

A-level grades are a key determinant of which degree courses students are accepted to. We now establish the consequences of biases by demonstrating how small differences in rankings impact future university outcomes. Now, instead of estimating the imbalance around the thresholds, we use our approach to estimate the impact of receiving a higher grade on university application outcomes. The estimates should be interpreted as the consequence of receiving a higher grade among students that were perceived to be similar in ability.

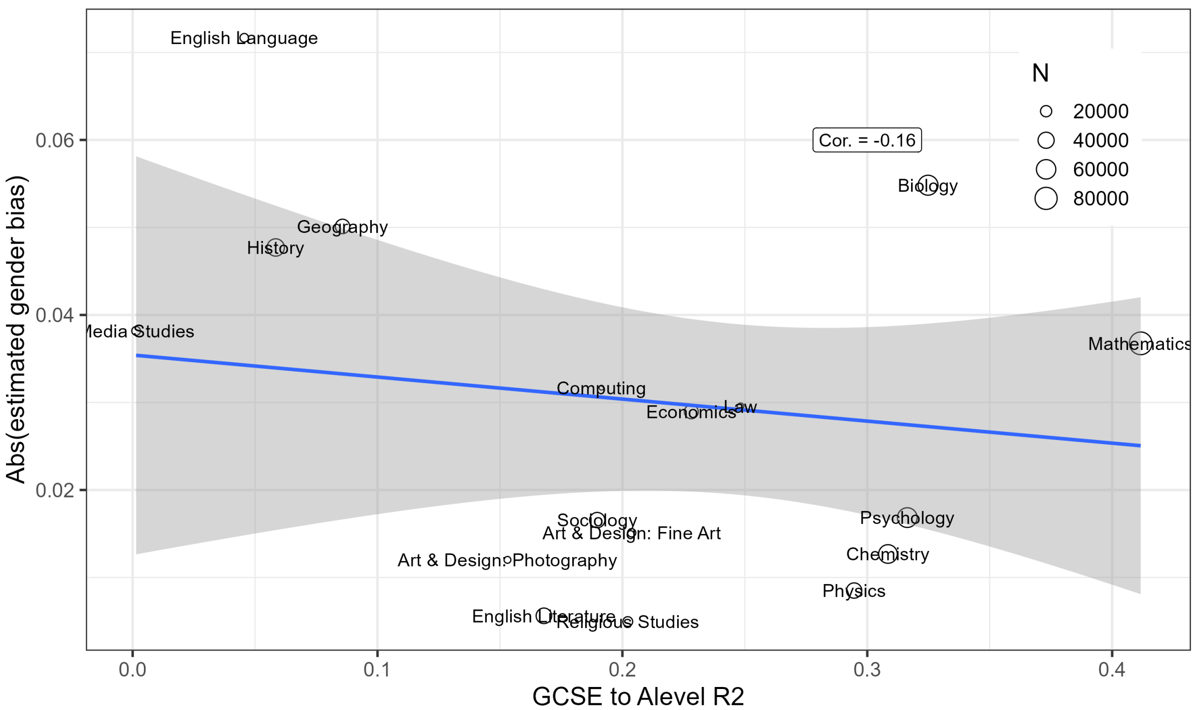
Table 3 shows the impact for those immediately above a grade boundary, on being accepted to any university degree conditional on applying (Panel a), and being accepted to their first ("firm") choice course conditional on being accepted to any university degree.[17]

[16] 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.

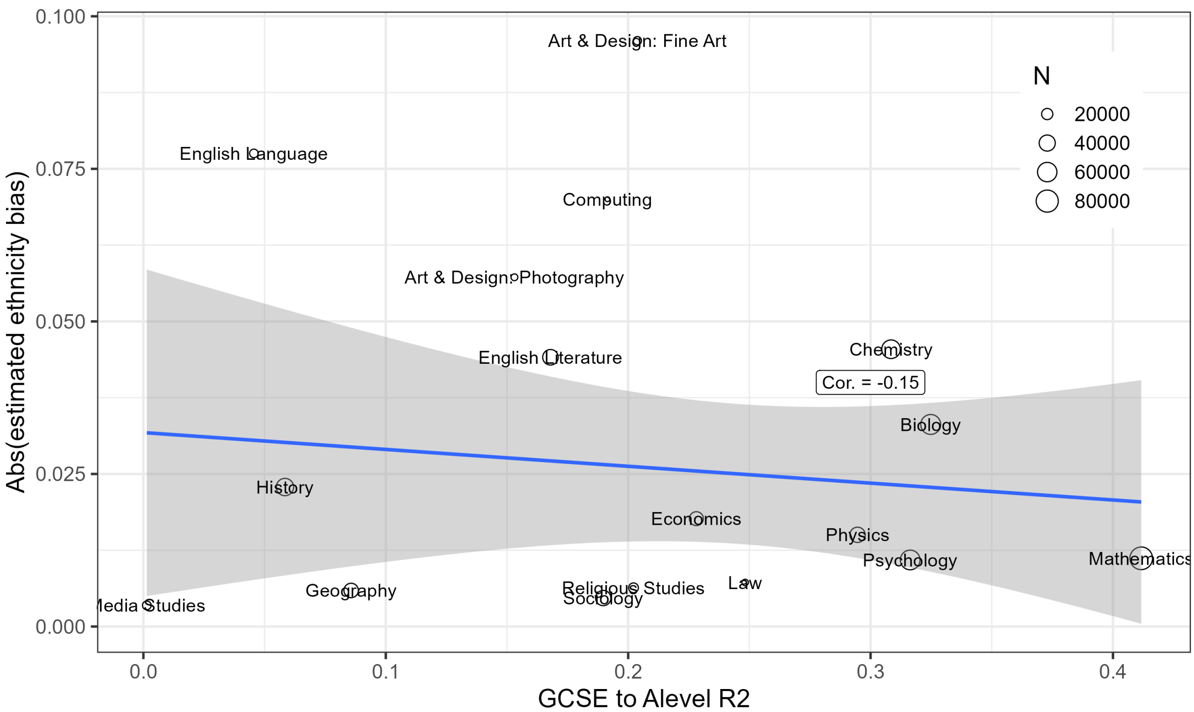
[17] As students in the UK generally apply and receive offers from university before completing the exams that these offers will hinge upon, they choose a "firm" and an "insurance" choice from their offers. The "firm" choice is their first choice that they will attend if they achieve the required grades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

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

(a) Female



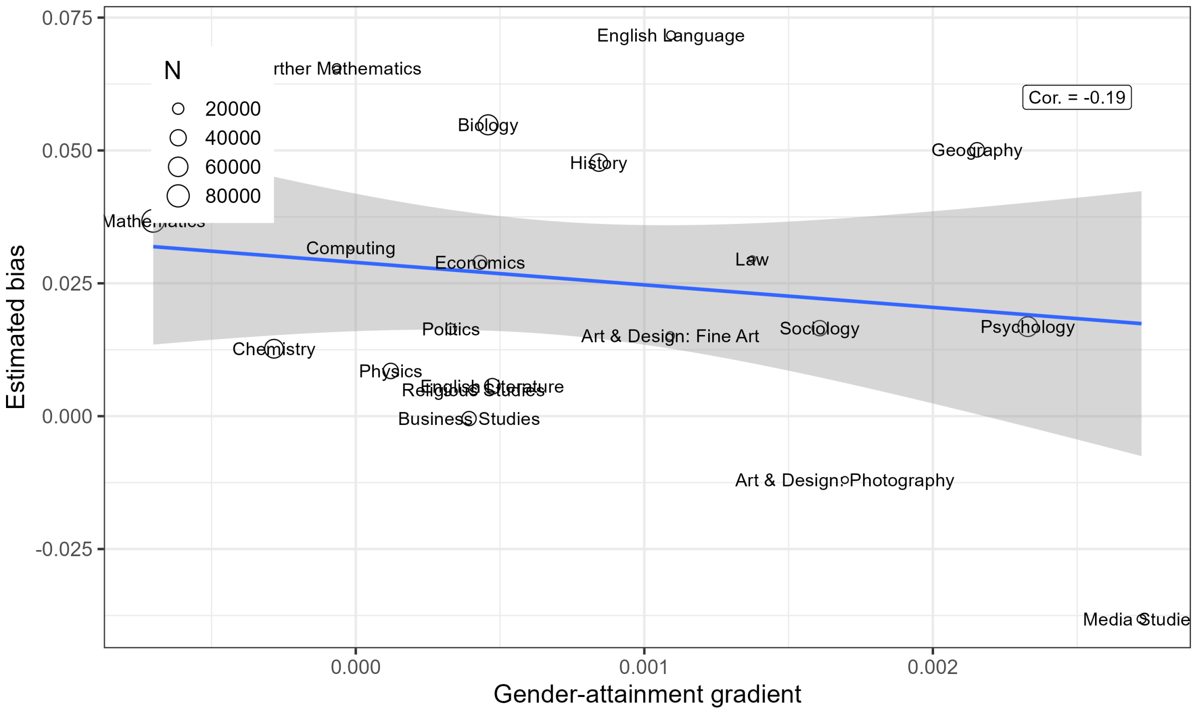
(b) White



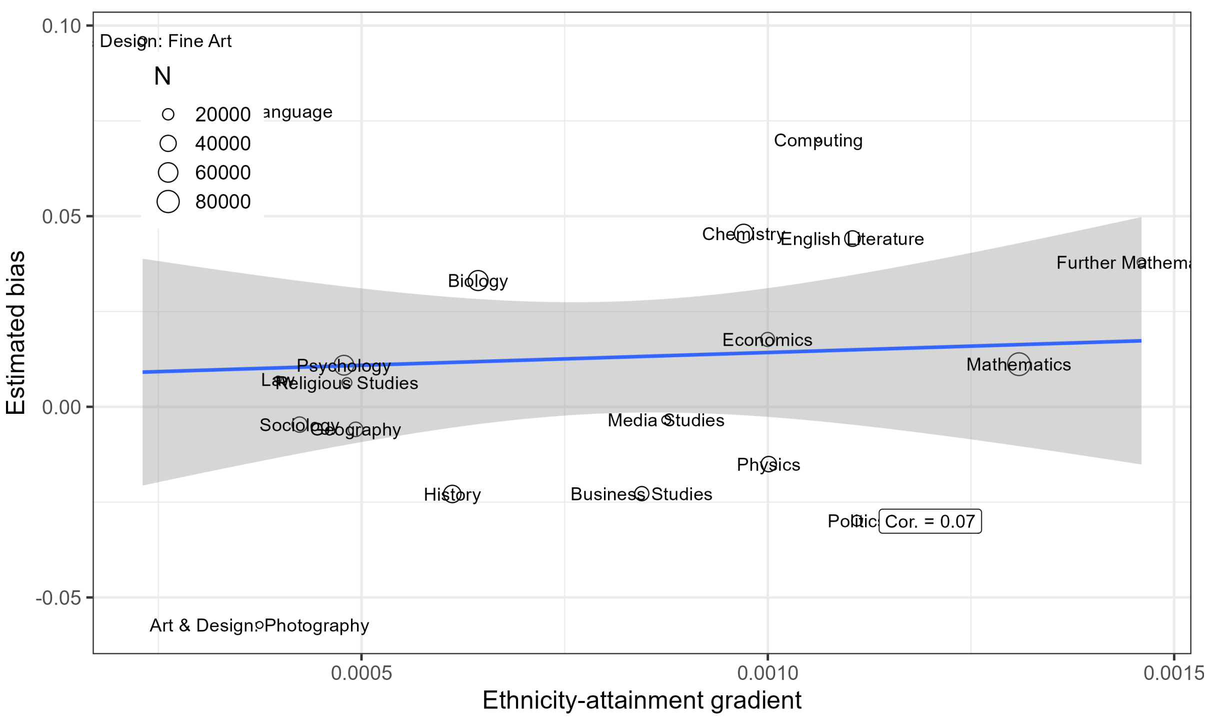
Notes: These panels show scatter plots of prior attainment informativeness on the absolute value of estimated bias by subject, weighted by the size of the subject (N). We measure prior attainment informativeness as the R^2 from a regression of age 18 attainment on age 16 attainment for the 2019 cohort (assessed externally). 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 pro-White bias estimates.

Figure 8: Bias estimates (τ) versus attainment-type gradient by subject

(a) Female



(b) White



Notes: These panels show scatter plots of attainment-type gradients on the absolute value of estimated bias by subject, weighted by the size of the subject (N). We estimate an attainment-type gradient for each subject from the 2019 cohort (assessed externally), by calculating the share of a given type at each percentile of the attainment distribution (see appendix B. 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 pro-White bias estimates.

Table 3: Consequences of achieving a higher grade on:

(a) being accepted anywhere | applying

Grade boundary:	All (1)	A*/A (2)	A/B (3)	B/C (4)	C/D (5)	D/E (6)
τ	0.021	0.010	0.036	0.034	−0.023	0.034
	(0.005)	(0.009)	(0.008)	(0.009)	(0.015)	(0.035)
Constant	0.724	0.831	0.751	0.669	0.647	0.604
N	32,624	6,412	11,164	10,398	3,870	772

(b) being accepted at "firm" choice | accepted

Grade boundary:	All (1)	A*/A (2)	A/B (3)	B/C (4)	C/D (5)	D/E (6)
τ	0.010	0.046	−0.003	−0.023	−0.007	0.127
	(0.007)	(0.011)	(0.011)	(0.012)	(0.021)	(0.048)
Constant	0.592	0.750	0.617	0.486	0.433	0.347
N	22,656	5,458	8,236	6,404	2,144	406

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.

Focusing first on the pooled estimates in column one, students with a higher grade in one of their three A-level courses are around 2.1p.p. more likely to be accepted onto any degree conditional on applying. The pooled effect on being accepted at the firm choice (panel b, column 1) is close to zero and statistically insignificant, masking substantial heterogeneity across grade boundaries.

The remaining columns break down these pooled estimates by grade boundaries, to capture effects throughout the grade distribution. We see that obtaining an A* rather than an A, does not impact gaining a place at university (Panel a), but increases the chance of being accepted at their firm (first) choice by 4.6p.p. (Panel b). For being accepted anywhere, the largest effects occur at the A/B and B/C boundaries (3.6p.p. and 3.4p.p. respectively), with negligible effects at the top boundary where the overwhelming majority of applicants are accepted regardless.

5.4 Teacher Altruism

Teachers assigned grades and ranks to students around five months *after* the students had applied to college. Given the high-stakes nature of the qualifications being awarded, a

choice.

Table 4: Teacher altruism towards university applicants as a mechanism

(a) applying to university

Grade boundary:	All (1)	A*/A (2)	A/B (3)	B/C (4)	C/D (5)	D/E (6)
τ	0.030	0.009	0.030	0.040	0.046	−0.003
	(0.004)	(0.007)	(0.006)	(0.007)	(0.011)	(0.024)
Constant	0.778	0.901	0.841	0.746	0.662	0.647
N	46,146	7,464	14,086	15,882	7,124	1,582

(b) estimated bias (τ) among university applicants

Dependent variable (X_i):	Female (1)	White (2)	FSM (3)
τ	0.023	0.019	−0.005
	(0.006)	(0.006)	(0.004)
Constant	0.540	0.590	0.117
	(0.004)	(0.004)	(0.003)
N	30,306	30,272	23,158

Notes: Panel (a) presents estimates of the effect of being ranked one above a grade boundary versus one below (+1 vs −1) on the probability of having applied to university, for the pooled sample (column 1) and by grade boundary (columns 2–6). Panel (b) re-estimates the main specification (equation 3.2) pooling across all subjects and grade boundaries, restricted to students who applied to university. Standard errors are in parentheses.

potential explanation for the biases we observe could be teachers bumping students over grade boundaries if they had applied to college. To test this directly Panel a of Table 4 shows our main specification, with the dependent variable being a indicator for whether the student had applied to university. We find that students who were awarded the higher grade were 3 percentage points more likely to have applied to university, compared to a baseline of 77 percent of students immediately below a threshold applying to university. This implies that the teachers are partly motivated when assigning grades by the application status of the student. This is consistent with Dee et al. (2019) who find evidence of altruism among teachers through the manipulation of test scores to prevent students from failing an exam.

To determine how much of our bias estimates could be attributed to such altruism we re-estimate our main specification on the sub-sample of students adjacent to grade boundaries who have both applied to university. Panel b of Table 4 shows that we continue to find statistically significant bias in favor of female and White students even among college applicants. The probability of being female increases by 2.3 percentage points across the grade boundary, which is around 20 percent lower than the 2.9 percentage point difference found in the full sample. This implies that around 20 percent of the gender bias is due to 'altruism'. For ethnicity, White students are 1.9 percentage points more likely to be given the higher grade in these edge cases, compared to 1.6 in the main sample. This implies that the teacher bias in favour of White students is stronger among the university applicant population. Note however, for both student types the difference in the estimates between the full and university applicant populations are not significantly different from each other. The implication is that teacher motivations to help students attend college do not explain the biases in favour of female and White students.

6 Robustness

The key assumption for our approach is that counterfactual assignment D_i^* is independent of student characteristic X_i for students adjacent to the boundary. For example the female achievement distribution is to the right of the male distribution (see Figures B.2 and B.4 in the appendix), then these achievement gaps will be correlated with student characteristics, and τ could be a combination of bias and achievement differences. We provide three distinct pieces of evidence to establish robustness of our main estimates are not driven by imbalance in ability across the threshold.

6.1 Consistency with Historical Achievement by Student Type

The simplest approach to examine whether the discontinuities in student types could be driven by differences in student abilities is to ascertain that the extent of bias for a type across subjects correlates by their performance across subjects. Figure 8 plots the extent of bias against type-achievement gradients. The intuition is that, if we were merely recovering differential abilities across thresholds we would expect to see a positive relationship between extent of bias for a student type and the historical achievement of that type. However, there is a weak negative correlation between bias and achievement across subjects by gender, and it is effectively flat by ethnicity. While not definitive proof that individual students around boundaries are balanced in potential outcomes, this provides suggestive evidence that τ is recovering the extent of teacher bias, rather than ability.

6.2 Conditioning on Prior Attainment

Rather than using the overall relationship between achievement by type, Table 5 examines the sensitivity of our main estimates to conditioning on individual prior achievement (T_i). We use students performance in national standardized and externally graded GCSE qualifications taken at age 16. Specifically we use mean standardized score across all GCSEs, which is more detailed than letter grades. Conditioning on prior achievement reduces the female estimate from 0.029 to 0.027 and the White estimate from 0.016 to

Table 5: Robustness — Conditioning on prior achievement

	Female		White	
	(1)	(2)	(3)	(4)
τ	0.029	0.027	0.016	0.015
	(0.005)	(0.005)	(0.005)	(0.005)
Constant	0.524	0.505	0.619	0.613
	(0.003)	(0.004)	(0.003)	(0.004)
T_i		✓		✓
N	43,564	43,563	42,834	42,833

Notes: This table reproduces the raw estimates from Table 2 (odd columns) alongside estimates conditional on mean standardised GCSE mark (T_i, even columns), using the crowded sample. Standard errors in parentheses.

0.015. Both remain highly significant and not significantly different in magnitude. These results demonstrate that characteristic-achievement gradients are not the primary driver of our estimates, and that our approach of weighting observations by the likely crowding in the class is sufficient to identify the parameter of interest.

However, simply conditioning on prior achievement may be unsatisfying for two reasons. First, it may be the case that age 16 scores 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 an alternative which makes fewer assumptions.

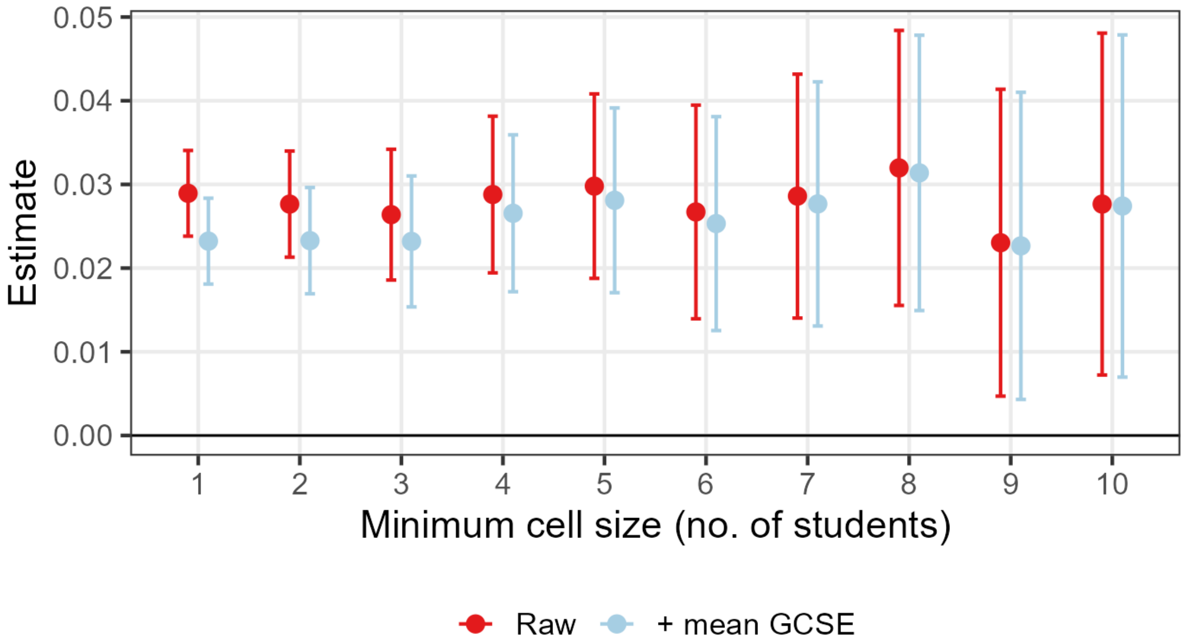
6.3 Crowding

Our third robustness test returns to our hypothetical example motivating the estimation approach, that in the limit, classes with infinite students will have identical students each side of the boundary. Our main specification already applies some "crowding" restrictions, including only well-populated boundaries and weighting by school-subject size. These apply the intuition that in settings with more students in adjoining grades the more likely it is that the marginal students are of similar abilities. Here we extend this analysis, estimating τ in sub-samples including only boundaries with increasing numbers of students in adjacent grades n_{jsg}. If the estimate of bias is stable in increasing n_{jsg} this is evidence that our main specification with a restriction of at least four students either side $N_{js} \geq 4$ is sufficient. This requires no functional form assumptions linking past and present achievement.

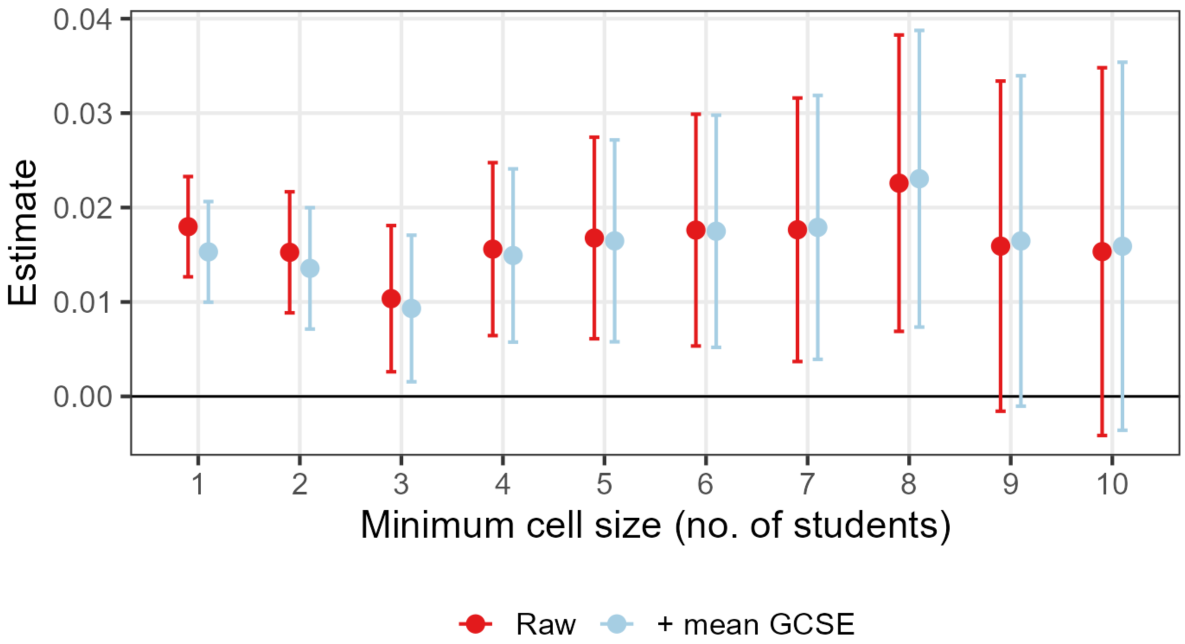
The results of this exercise can be found in Figure 9. Our baseline estimates are those with at least four students either side $n_{jsg} \geq 4$. For female students, the estimates are remarkably stable even for $n_{jsb} \leq 4$. As the number of adjacent students increases to 4 the raw estimates converge on the conditional estimates, after that point both the raw and conditional estimates are very similar and follow the same pattern. For White students, both the raw and conditional estimates are similarly stable across all cell-size restrictions. This coefficient stability implies that τ is not being driven by differences in potential outcomes. FSM results from this analysis are presented in Appendix C.

Figure 9: Crowding analysis — increasing number of students in adjacent grades

(a) Female: τ



(b) White: τ



Notes: These panels show the results of a crowding analysis exercise where we restrict our sample to students in larger and larger n_{jsg} (i.e. number of students adjacent to the grade boundary). 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 White ethnicity in (b). FSM results are in the appendix, figure C.4.

7 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 – in large classes 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 and White students increases discontinuously as we pass over a grade boundary, implying teacher bias in favour of both groups. Our results suggest teachers are awarding students grades based on their characteristics, rather than solely on the basis of their academic performance, or their future accademic plans. Our results imply that 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 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 the difference-in-difference approach that is popular in the literature (Lavy, 2008; Lavy and Sand, 2015) will not recover the parameter of interest. By contrast our local randomisation approach only has one dimension through which students are measured, teacher perception of ability, 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.

The empirical approach 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 many agents/objects in a system with thresholds. The approach is ideal when there is subjective ranking with many assessed units, but the same intuition can be applied to any setting 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.

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.

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.

Dee, T. S., W. Dobbie, B. A. Jacob, and J. Rockoff (2019): "The causes and consequences of test score manipulation: Evidence from the New York regents examinations," *American Economic Journal: Applied Economics*, 11, 382–423.

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.

Diamond, R. and P. Persson (2016): "The long-term consequences of teacher discretion in grading of high-stakes tests," Tech. rep., National Bureau of Economic Research.

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

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.

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 education & 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," .

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.

A Application of Bayes Rule

Our empirical strategy compares the share of student characteristic $X_i \in \{0,1\}$ (e.g. female, FSM-eligible) immediately above and below each grade boundary. This appendix establishes the formal relationship between our estimated coefficient τ and an economically interpretable measure of teacher discrimination Δ.

Setup. At each populated school-subject-grade boundary b_{jsg}, we focus on the two most marginal students: the student ranked last in the higher grade and the student ranked first in the lower grade. Let $D_i \in \{0,1\}$ indicate whether student i is assigned to the higher grade at their boundary ($D_i = 1$) or the lower grade ($D_i = 0$). As described in Section 2, both the grade assignment and the within-grade ranking are teacher-determined. The quantity D_i therefore reflects the joint outcome of both decisions, and may be a function of X_i if the teacher discriminates.

A.1 Relating our estimated coefficient τ to "bias" Δ

Define

$$\tau = \Pr(X = 1 \mid D = 1) - \Pr(X = 1 \mid D = 0), \tag{10}$$

$$\Delta = \Pr(D = 1 \mid X = 1) - \Pr(D = 1 \mid X = 0). \tag{11}$$

Step 1: Apply Bayes' rule.

$$\Pr(X = 1 \mid D = 1) = \frac{\Pr(D = 1 \mid X = 1)\Pr(X = 1)}{\Pr(D = 1)}, \tag{12}$$

$$\Pr(X = 1 \mid D = 0) = \frac{\Pr(D = 0 \mid X = 1)\Pr(X = 1)}{\Pr(D = 0)}. \tag{13}$$

Factoring out $\Pr(X = 1)$:

$$\tau = \Pr(X = 1)\left[\frac{\Pr(D = 1 \mid X = 1)}{\Pr(D = 1)} - \frac{\Pr(D = 0 \mid X = 1)}{\Pr(D = 0)}\right]. \tag{14}$$

Step 2: Simplify the bracket. Let $q = \Pr(D = 1 \mid X = 1)$ and substitute $\Pr(D = 0 \mid X = 1) = 1 - q$:

$$\tau = \Pr(X = 1) \cdot \frac{q(1-p) - (1-q)p}{p(1-p)} = \Pr(X = 1) \cdot \frac{q - p}{p(1-p)}. \tag{15}$$

Step 3: Expand $q - p$ using the law of total probability. Since $p = q\Pr(X = 1) + \Pr(D = 1 \mid X = 0)\Pr(X = 0)$:

$$\begin{aligned} q - p &= q - q\Pr(X = 1) - \Pr(D = 1 \mid X = 0)\Pr(X = 0) \\ &= q\Pr(X = 0) - \Pr(D = 1 \mid X = 0)\Pr(X = 0) \\ &= \Pr(X = 0)\,\Delta. \end{aligned} \tag{16}$$

Step 4: Substitute back.

$$\boxed{\tau = \frac{\Pr(X = 1)\Pr(X = 0)}{\Pr(D = 1)\Pr(D = 0)}\,\Delta.} \tag{17}$$

Since $\Pr(X = 1), \Pr(X = 0), \Pr(D = 1), \Pr(D = 0) \in (0, 1)$, the prefactor is strictly positive, so $\tau = 0 \iff \Delta = 0$. The result holds analogously conditional on perceived ability δ, replacing all unconditional probabilities with their δ-conditional counterparts.

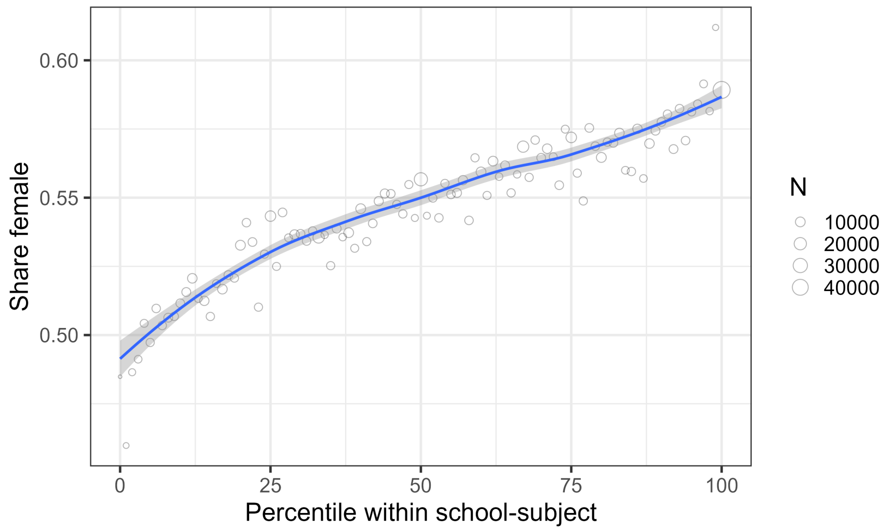
B Attainment gradients

Figures B.2a – B.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.

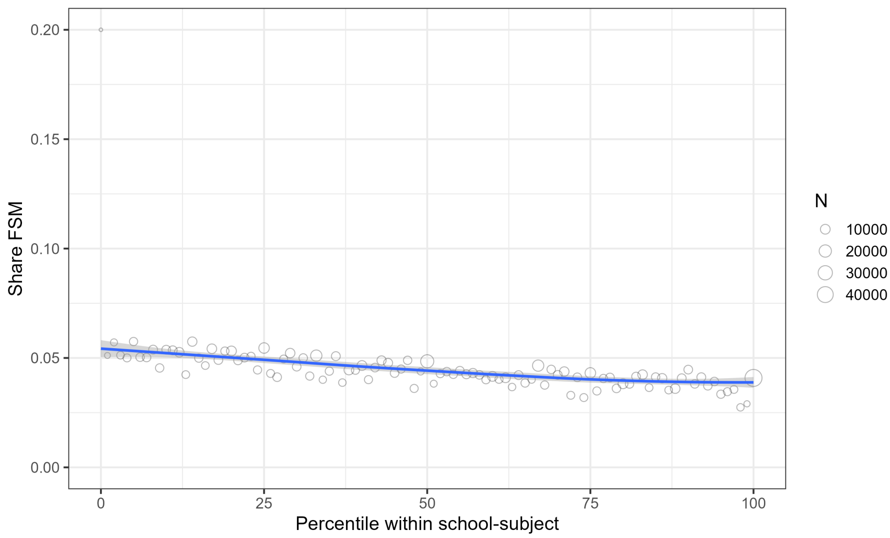
We repeat the exercise for each of the top 20 subjects in figure B.2.

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

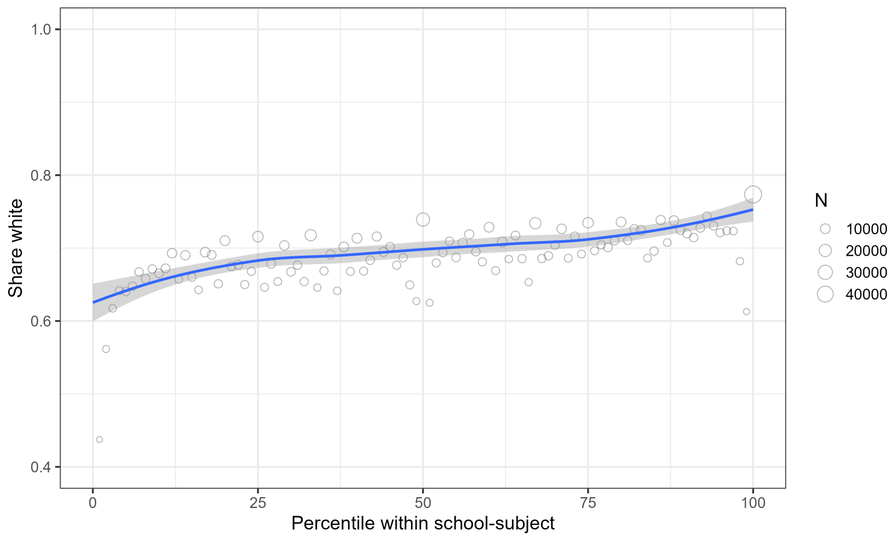
(a) Female



(b) FSM



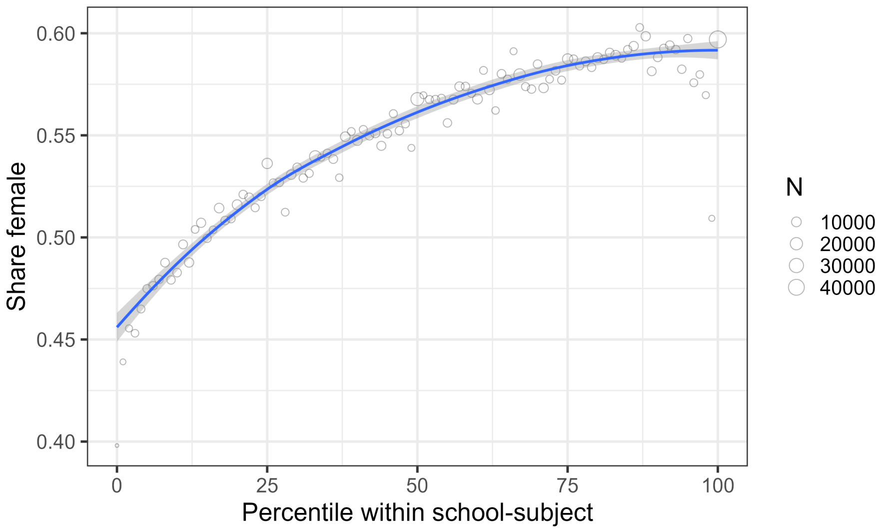
(c) White



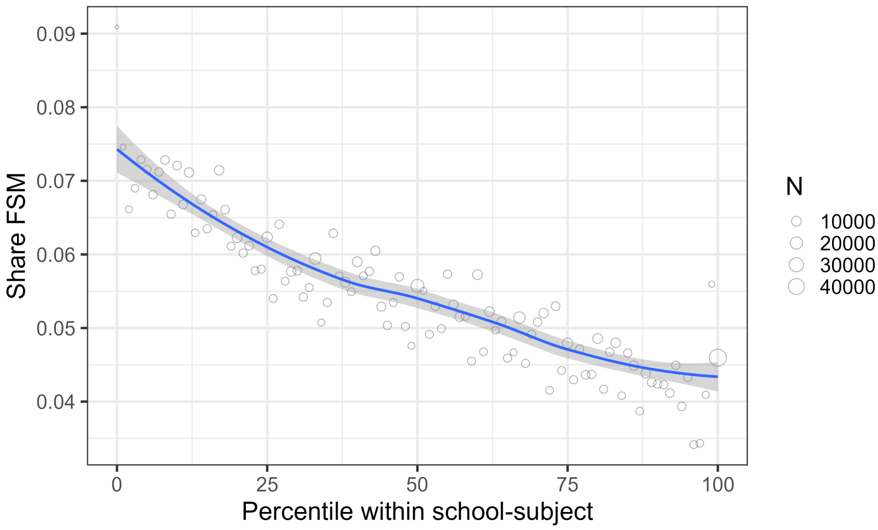
Notes: These plots show the share of characteristic X among all students for each percentile in a school and subject. Then a line of best fit is drawn using local regressions.

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

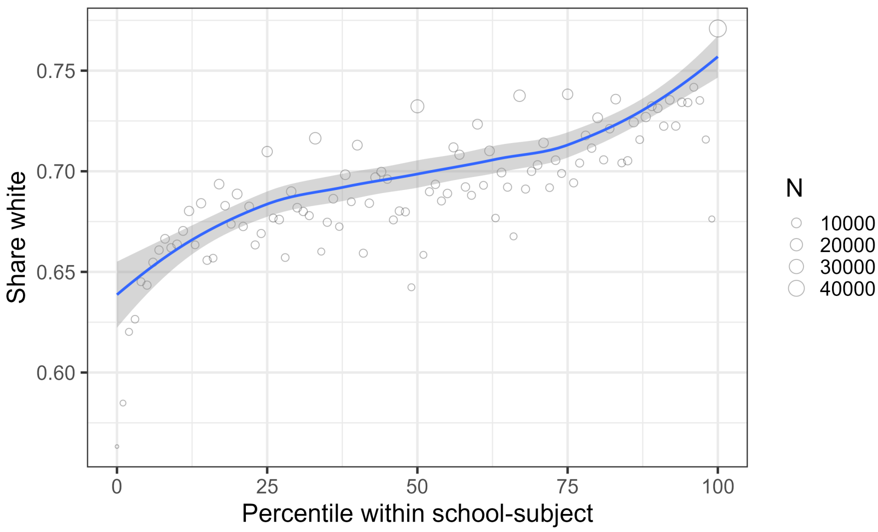
(a) Female



(b) FSM



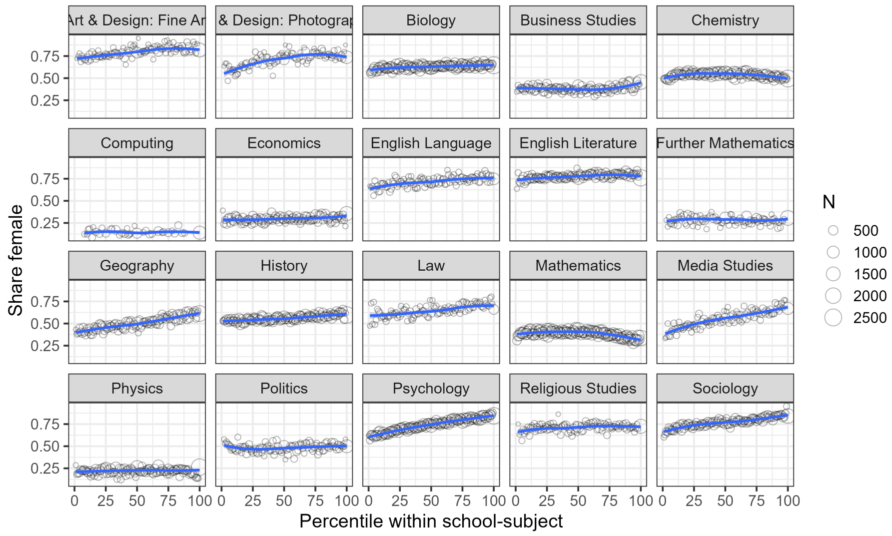
(c) White



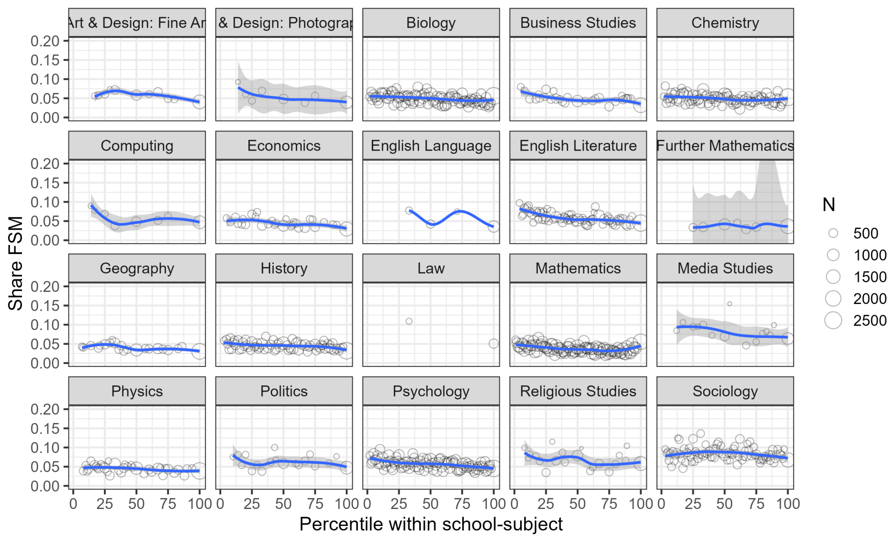
Notes: These plots show the share of characteristic X among all students for each percentile in a school and subject. Then a line of best fit is drawn using local regressions.

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

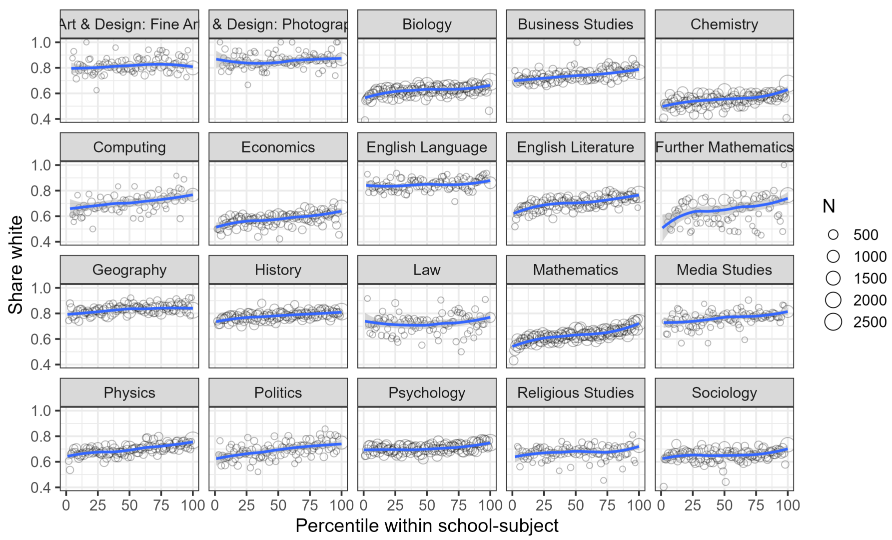
(a) Female



(b) FSM



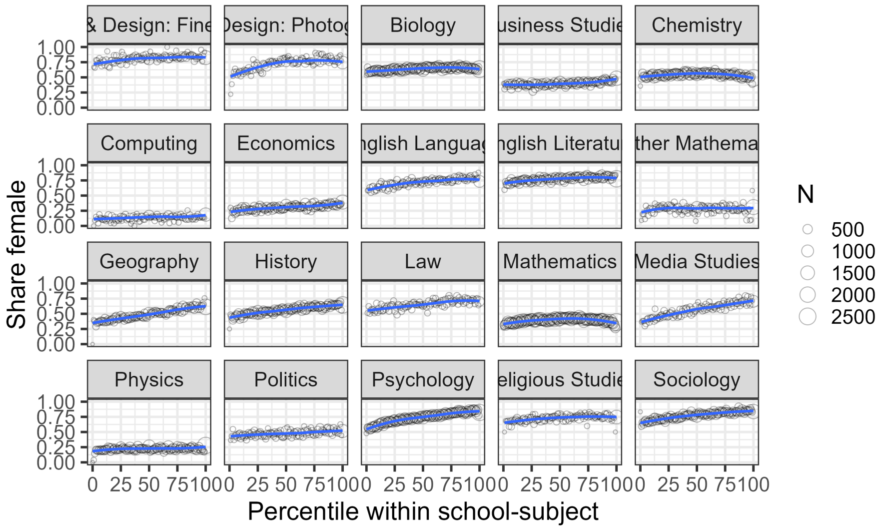
(c) White



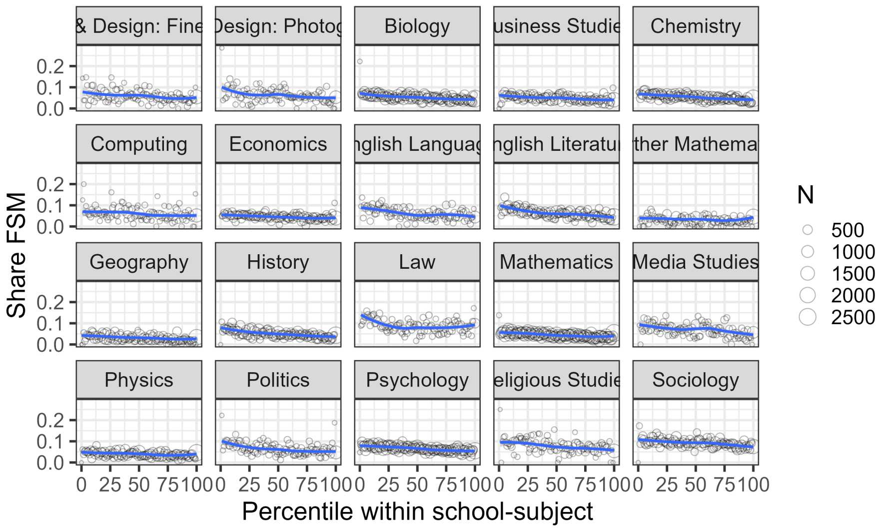
Notes: These plots show the share of characteristic X among all students for each percentile in a school and subject. Then a line of best fit is drawn using local regressions.

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

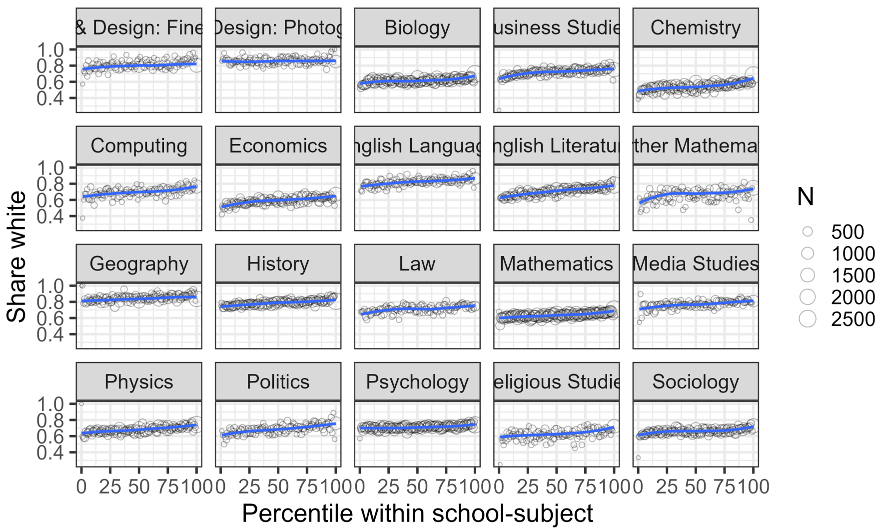
(a) Female



(b) FSM



(c) White

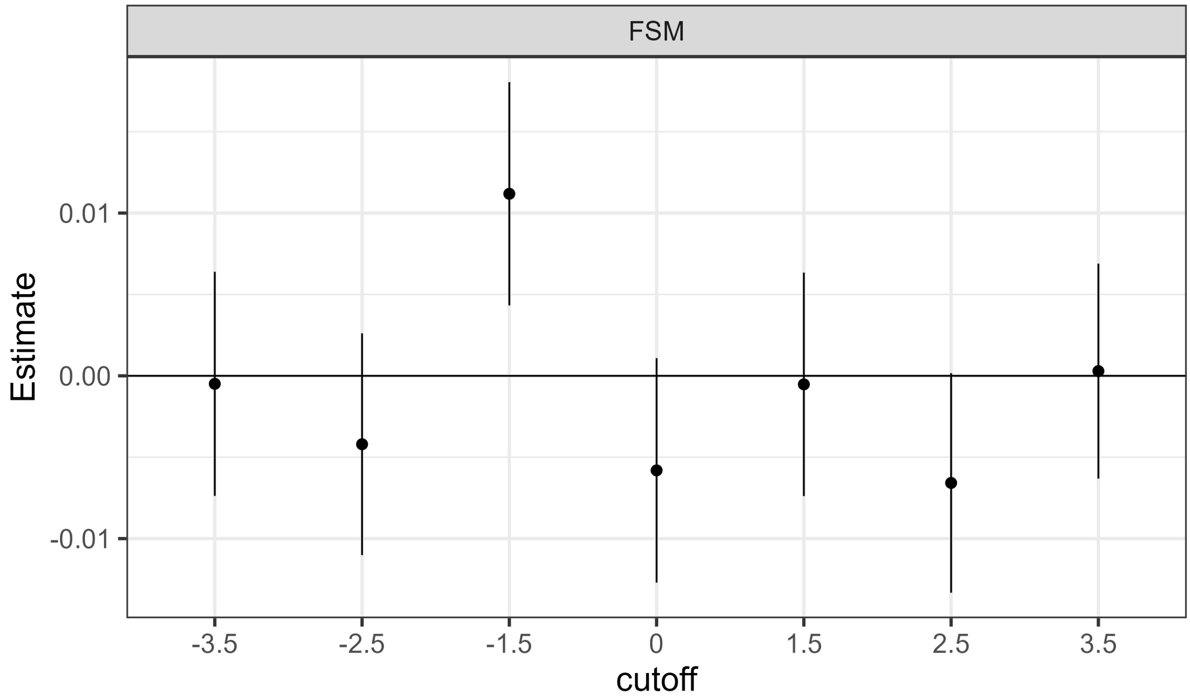


Notes: These plots show the share of characteristic X among all students for each percentile in a school and subject. Then a line of best fit is drawn using local regressions.

C Results for FSM students

Under the crowding restriction (at least 3 students on each side of every grade boundary), the FSM estimate is close to zero and statistically insignificant ($\hat{\tau} = -0.003$, SE $= 0.004$, $N = 32{,}772$). This likely reflects limited statistical power: the small share of FSM students in the crowded boundary sample reduces effective cell sizes for this subgroup. The estimates below document the full pattern of FSM results for reference.

Figure C.1: Placebo tests by distance from grade boundary: FSM



Notes: This figure presents FSM estimates alongside "placebo" estimates for pairs of adjacently ranked observations that do not straddle grade boundaries. The extending bars represent 95% confidence intervals. All estimates are conditional on prior attainment and weighted by how close the share of FSM is to 50% within a school-subject.

Figure C.2: Estimates by subject: FSM

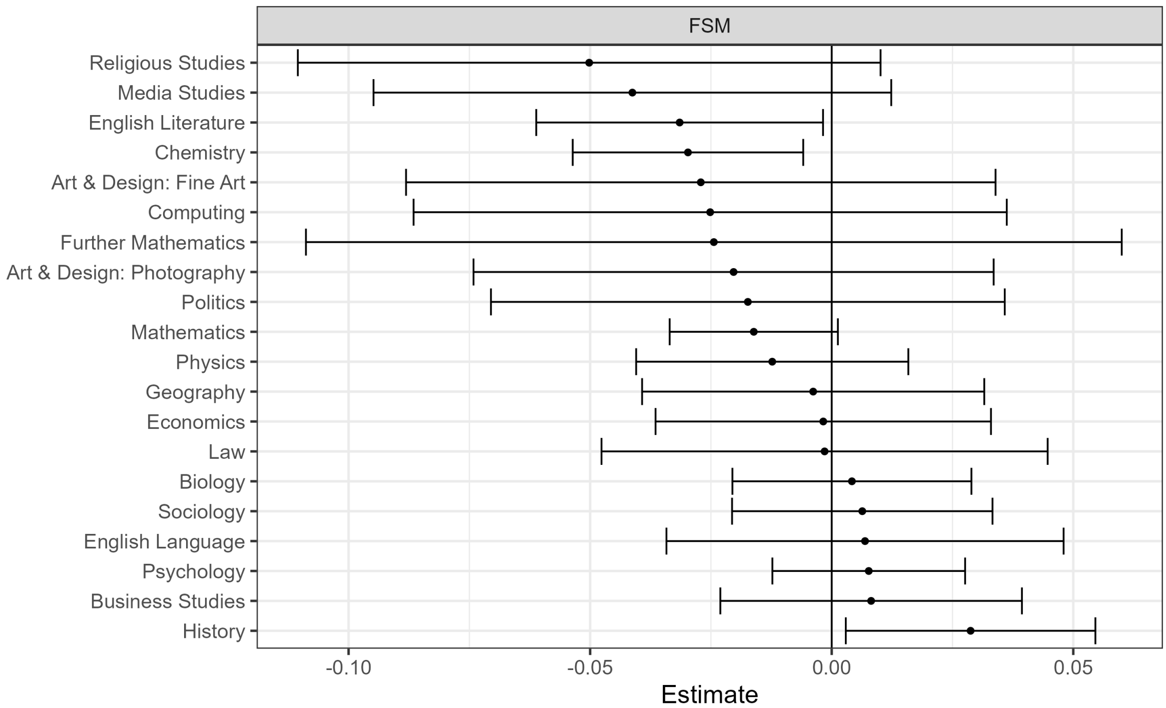


Figure C.3: FSM τ: latent achievement bandwidth analysis

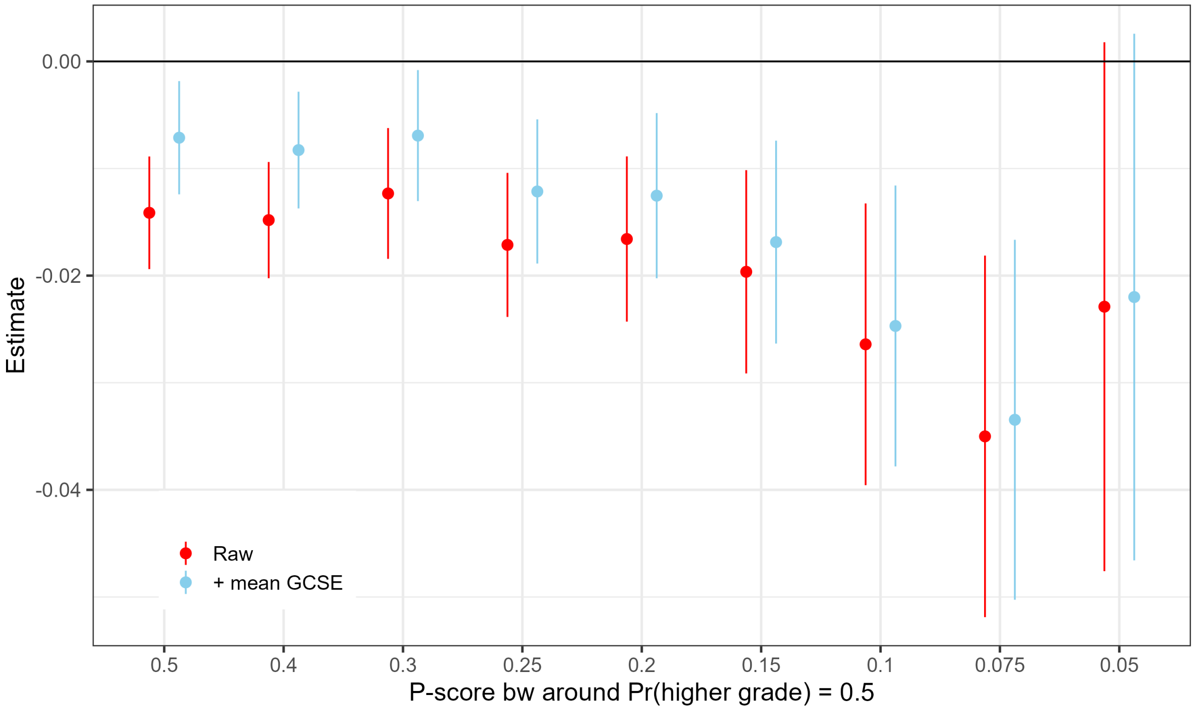


Figure C.4: FSM: τ versus minimum adjacent students