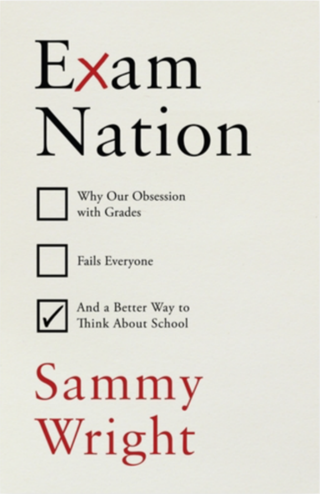
Systematic gaps in teacher judgement: A new approach

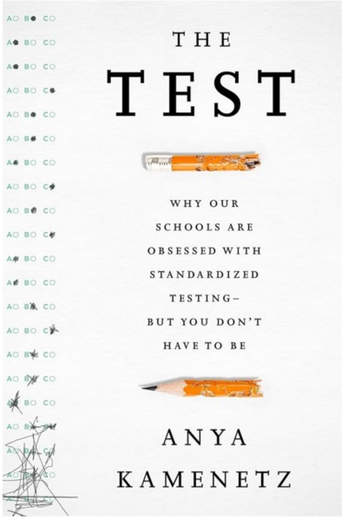
Oliver Cassagneau-Francis[1] Lindsey Macmillan[1] **Rich Murphy**[2,3] Gill Wyness[1,3]

[1]UCL Centre for Education Policy and Equalising Opportunities
[2]UT Austin, NBER, IZA, CESifo
[3]Centre for Economic Performance, LSE

NBER Education Meeting
May 1st 2025

Motivation - Changing Student Assessments





Motivation - Changing Student Assessments

Over the last 30 years International trend towards standardized testing (Lingard et al., 2016)

Trend reversed recently with many countries are rethinking their reliance on standardized testing

Especially as part of the university application process

Motivation - Changing Student Assessments

Over the last 30 years International trend towards standardized testing (Lingard et al., 2016)
Trend reversed recently with many countries are rethinking their reliance on standardized testing
Especially as part of the university application process

- UK: Considering replacing use of continuous teacher assessment
- Portugal: Discussing abolishing final school exams
- USA: Stopping and gradual re-use of SATs in university applications
- AUS: Returning to exams after use of AI to write essays

Motivation - Changing Student Assessments

Concerns with Standardized Testing

- Increase Student Anxiety (Holbein and Ladd, 2017)
- Perverse Incentives - Teaching to the test, Narrowing of Curriculum (Barlevy and Neal, 2012)

Standardized Testing has undesirable features of a performance metric

- Narrow - Do not adequately capture skills, personality, motivations (Kautz et al., 2014)
- Noisy - Dependent on performance on single day (Rimfeld et al., 2019)
- Uninformative (Allensworth and Clark, 2020; Geiser and Santelices, 2007)
- Biased - Gender (Cai et al., 2019; Galasso and Profeta, 2024) and Culture (Lemann, 2024)

Motivation - Changing Student Assessments

Teacher Assessment also has undesirable features of a performance metric

- Incomparable
- Uninformative (Chetty et al., 2023; Friedman et al., 2025)
- Biased (Lavy, 2008; Lavy and Sand, 2015; Carlana, 2019; Terrier, 2020; Avitzour et al., 2020; Burgess et al., 2022)

Establishing the extent of bias of teachers when assigning grades is critical

This Paper

Research question:

Do teachers favour some students over others when awarding high-stakes grades?

Teacher bias in relation to student

- Gender
- Ethnicity
- Social Economic Status

Related literature

There is a large literature measuring teacher biases

Direct Measures of Bias

- Grading the same test twice (Hinnerich et al., 2011)
- Randomise student characteristics (Hanna and Linden, 2012)
- Implement Implicit Association Tests (IAT) (Carlana, 2019; Alesina et al., 2018)

Concerns: Limited populations, External validity, Applicability

Indirect Measure of Bias

- Compare the differences in achievement gaps with 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)

Concerns: Strong Assumptions, Specific Settings

Indirect Measure of Bias Approach

Indirect Approach

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

Where Y is student performance, and X is student type
β is a combination of factors e.g. ability, effort, test-format, bias
Need to account for factors not related to bias, to the extent that they correlate with X

Lavy (2008) compare similar assessments, which are Blind (B) to X and Non-Blind (NB)

$$Y_{iNB} = \nu + \gamma X_i + \sigma_{NB} Assessment(ability, effort, format)_i + \zeta_i$$

$$Y_{iB} = \mu + \sigma_B Assessment(ability, effort, format)_i + \psi_i$$

If assume $\sigma_{NB} = \sigma_B$, then

$$Y_{iB} - Y_{iNB} = \pi + \gamma X_i + \chi_i$$

Indirect Measure of Bias Approach

What comes with $\sigma_{NB} = \sigma_B$ assumption?

The NB and B assessments:

- Have same mapping of ability to Y_i
- Elicit the same performance
- Have same measurement error

for each X

Indirect Measure of Bias Approach

For example:
Assume teachers are not biased
But teachers still may appear biased when comparing verbal NB assessments to written B assessments

This could be due to

- Females may have better verbal skills than males (Hirnstein et al., 2023)

Indirect Measure of Bias Approach

For example:
Assume teachers are not biased
But teachers still may appear biased when comparing verbal NB assessments to written B assessments

This could be due to

- Females may have better verbal skills than males (Hirnstein et al., 2023)
- Females may perform worse in standardized assessments (Cai et al., 2019; Galasso and Profeta, 2024; Arenas and Calsamiglia, 2025)

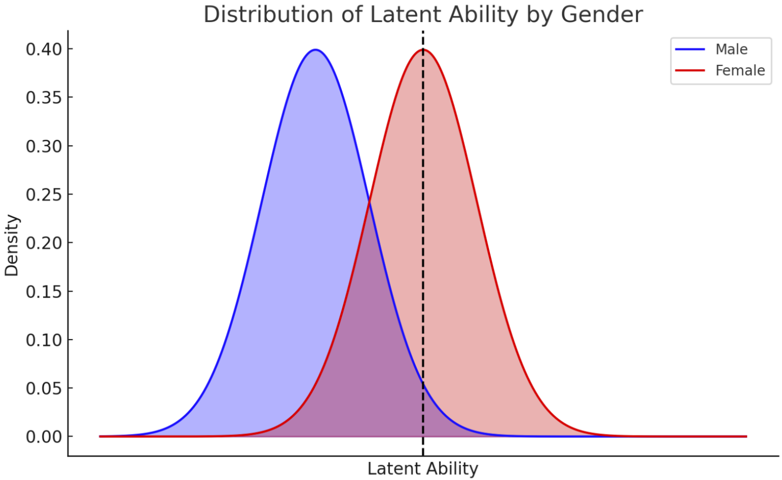
Indirect Measure of Bias Approach

For example:
Assume teachers are not biased
But teachers still may appear biased when comparing verbal NB assessments to written B assessments

This could be due to

- Females may have better verbal skills than males (Hirnstein et al., 2023)
- Females may perform worse in standardized assessments (Cai et al., 2019; Galasso and Profeta, 2024; Arenas and Calsamiglia, 2025)
- Females high achieving in B, less likely driven by measurement error(Zhu, 2024; Delaney and Devereux, 2025)

Indirect Measure of Bias Approach



Indirect Measure of Bias Approach

For example:
Assume teachers are not biased
But teachers still may appear biased when comparing verbal NB assessments to written B assessments

This could be due to

- Females may have better verbal skills than males (Hirnstein et al., 2023)
- Females may perform worse in standardized assessments (Cai et al., 2019; Galasso and Profeta, 2024; Arenas and Calsamiglia, 2025)
- Females high achieving in B less likely driven by measurement error(Zhu, 2024; Delaney and Devereux, 2025)

This makes it hard to determine the magnitude or the direction of the bias
This approach relies on the assumption linking Y_{iB} and Y_{iNB}, namely $\sigma_{NB} = \sigma_B$
What if there was no Y_{iB}?

This Paper

Question: *Do teachers favour some students over others when awarding high-stakes grades?*

Setting: During COVID teachers in UK graded and ranked all students in each school-subject

Method: Measure X of adjacent students next to every school-subject-grade boundary. Compare concentration of student type $\overline{X}_{RHS}$ to $\overline{X}_{LHS}$

Results: Teachers are biased in favor of white (1.2ppt, 1.9%), female (2.3ppt, 4.5%), non-FSM students (0.9ppt, 4.6%). The extent of the bias varies by grade boundaries and subject.

Consequences:
Teacher assessments are biased
We show that these grades impact post secondary enrollment

Contributions

1. New approach
   - Do not need to make assumption linking B and NB assessments
   - Applicable to many settings
   - Alternate approach validates existing estimates
2. Nature of the measurement
   - Direct measures of subjective assessment
   - Impact of bias in high stakes setting
3. Heterogeneity of effects
   - Measure bias for gender, ethnicity and SES
   - Measure bias by student ability
   - Measure bias by subject

Institutional Setting

English Secondary School Qualification Timetable

Year 11	Compulsory Age 16 Standardized exams (GCSE)
Fall of Year 13	Apply to universities
June of Year 13	Take Age 18 Standardized exams (A level)
August after Year 13	Receive exam results & confirm place at university

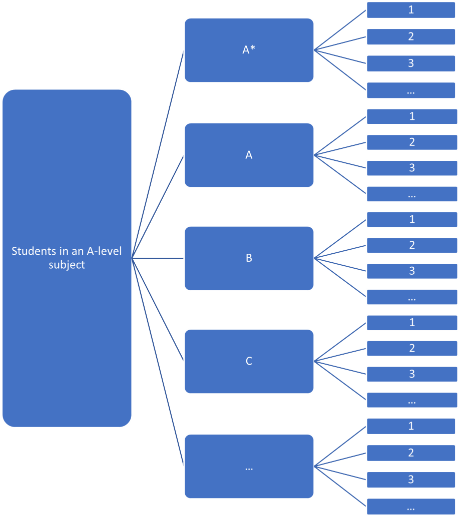
COVID-19 pandemic disruption

Year 11	Compulsory Age 16 Standardized exams (GCSE)
Fall of Year 13	Apply to universities
March 2020	COVID pandemic - all schools close
June 2020	Age 18 exams cancelled, Teacher Assessment
August 2020	Teacher grades used & confirm place at university

Institutional Setting

How were grades awarded?

Ofqual guidance: "Exam boards will ask exam centres to generate, for each subject":



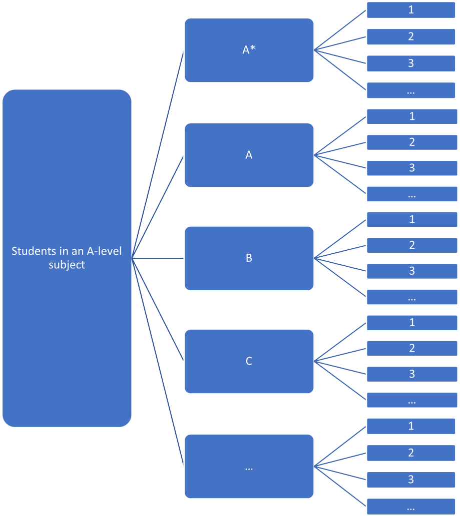
*"the grade that each student is most likely to have achieved if they had sat their exams"

How were grades awarded?

Ofqual guidance: “Exam boards will ask exam centres to generate, for each subject”:

- “centre assessment grades* for their students”

*“the grade that each student is most likely to have achieved if they had sat their exams”

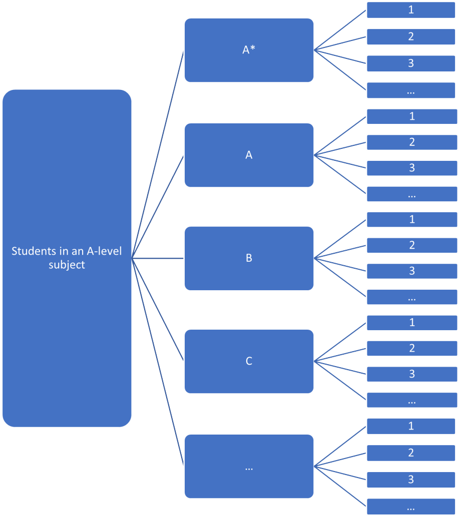


How were grades awarded?

Ofqual guidance: “Exam boards will ask exam centres to generate, for each subject”:

- “centre assessment grades* for their students”
- “and then to rank order the students within each of those grades.”

*“the grade that each student is most likely to have achieved if they had sat their exams”



Institutional Setting

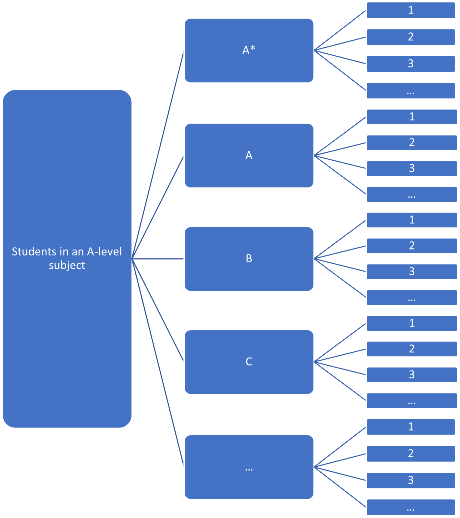
How were grades awarded?

Ofqual guidance: “Exam boards will ask exam centres to generate, for each subject”:

- “centre assessment grades* for their students”
- “and then to rank order the students within each of those grades.”

  → for most grades in all schools and subjects we have top and bottom ranked students

*“the grade that each student is most likely to have achieved if they had sat their exams”



How were grades awarded?

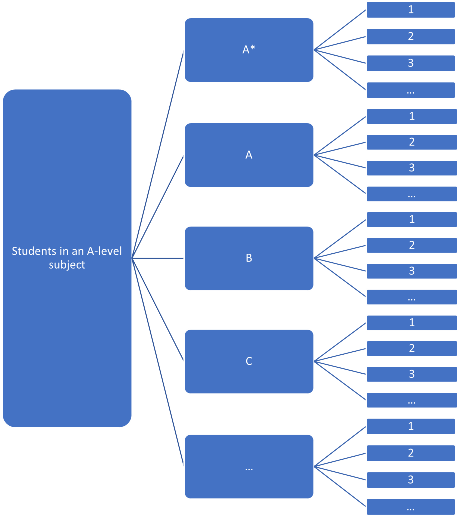
Ofqual guidance: "Exam boards will ask exam centres to generate, for each subject":

- "centre assessment grades*"
- "and then to rank order the students within each of those grades."

  $\rightarrow$ for most grades in all schools and subjects we have top and bottom ranked students

*"the grade that each student is most likely to have achieved if they had sat their exams"

Teachers would be aware that grades and ranks would be used in an algorithm

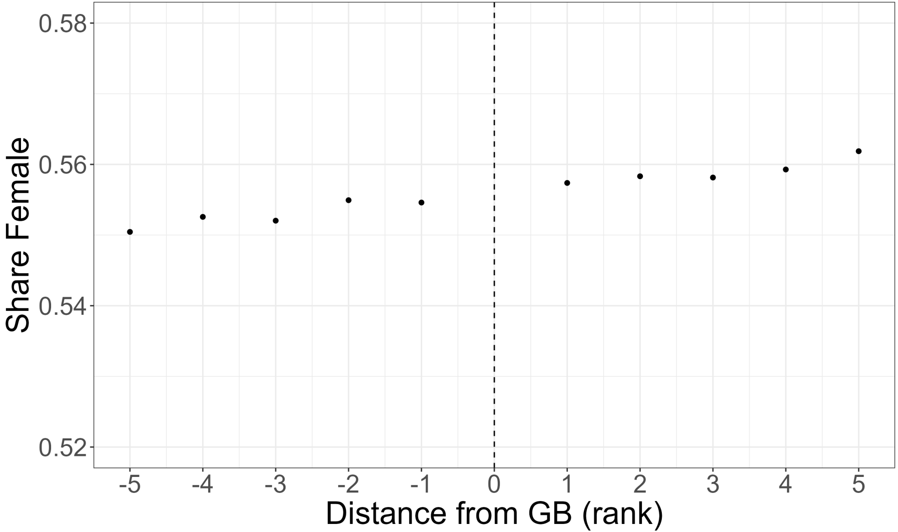


Empirical approach + Data

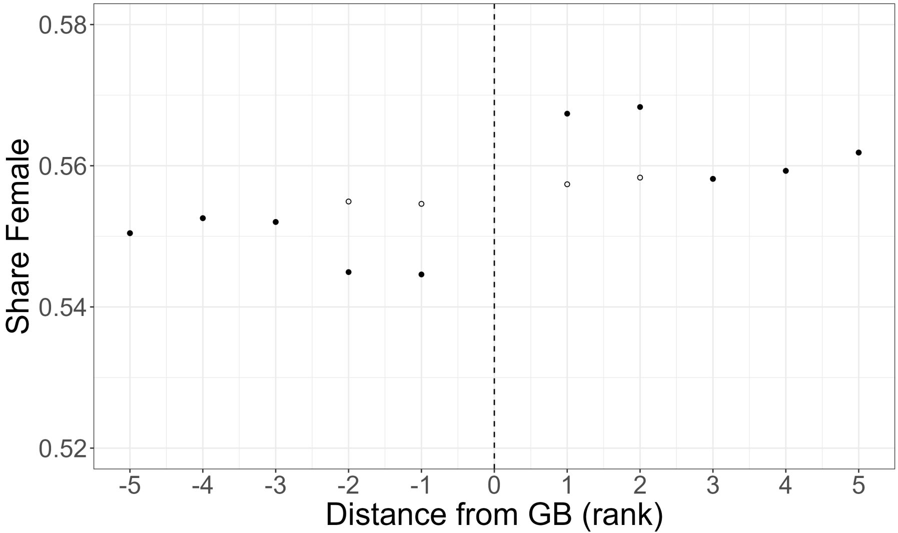
Empirical Approach - Intuition

- Exploit the ranking of students within grades
  - Identify the most marginal students
  - Each boundary will have marginal students
  - Do not need to extrapolate to less marginal students
- Concentration of student characteristics $(\overline{X})$ *should be* continuous
- Change in concentration either side of school-subject-grade boundary evidence of bias

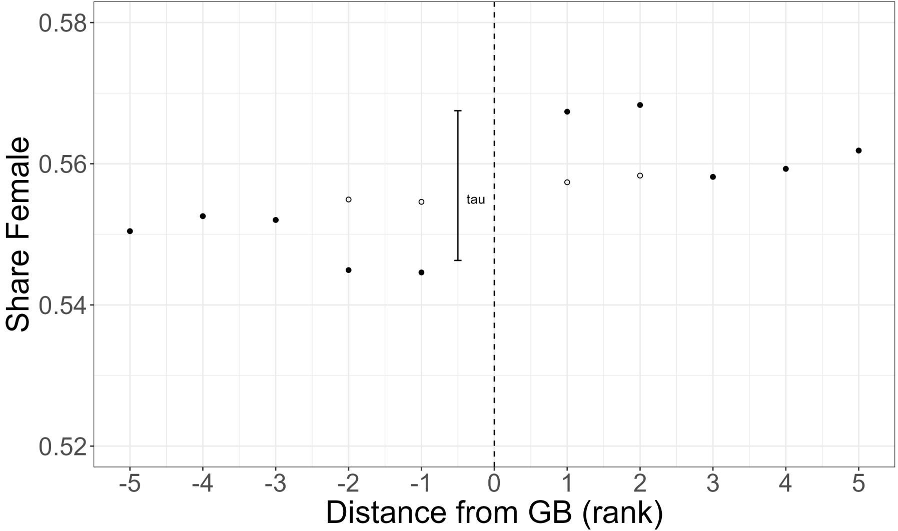
Stylized setup



Stylized setup



Stylized setup



Empirical approach:Local randomisation

Implement Local Randomization approach (Li et al 2021) ▸ details

- Assumption: observations adjacent to cutoff is “as-good-as-random”.

  $\rightarrow$ have enough data close to cutoff to choose smallest window (-1, 1) around boundary *b* in subject *s* and in school *j*
- Akin to RDD “robustness test”
- Easy to operationalise as a regression:

$$X_i = \beta_0 + \tau D_{ijsb} \; (+\beta_1 T_i) + \varepsilon_{ijsb}$$

where

- X_i is indicator for student characteristic
- D_{ijsb} is an indicator for being on the RHS of GB *b*, in subject *s* in school *j*
- T_i prior attainment — average marks across all age 16 exams

GRADE data

- Bespoke admin data linking ofqual, DfE, + UCAS, for first time
- Population of students who would take age 16 and 18 examinations in England in 2019 and 2020
- Ranking for every student in each subject and school
- Private schools not required to report FSM/ethnicity

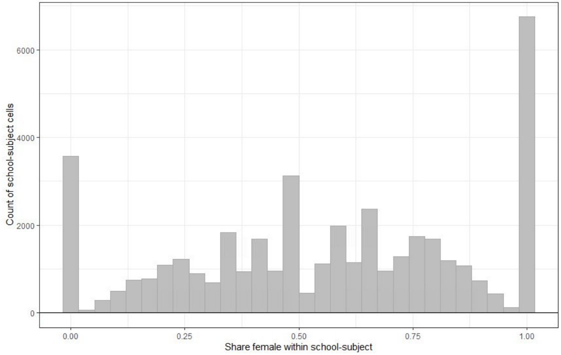
Academic year	2020
Female	.55
N	222,643
FSM eligible	.07
White	.73
N	198,328

GRADE data

Ability to detect bias limited by

- Lack of diversity in Post-16 education
- Student sorting to schools
- Student sorting to subjects
- Single sex schools

In these situations does $\tau = 0$ mean no bias?
Will attenuate measurement of bias
We weight each school-subject-boundary observation by $W_{js} = \overline{X}_{js} \cdot (1 - \overline{X}_{js})$



Female White FSM

Results

Overview of results

- Present bias with regards to
  - Female
  - White
  - FSM
- In terms of
  - Bias overall
  - Bias by grade Boundary (GB)
  - Bias by subject

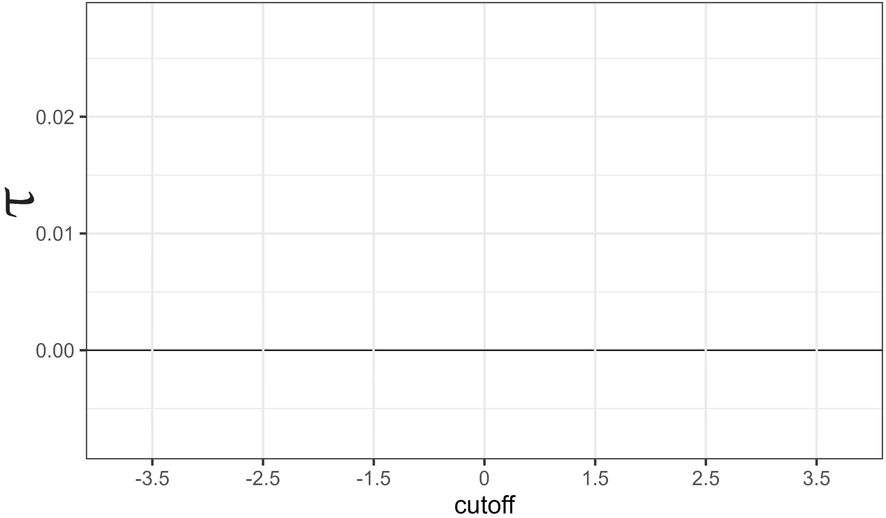
Overall Bias
(stacking GBs and Subjects)

Overall bias

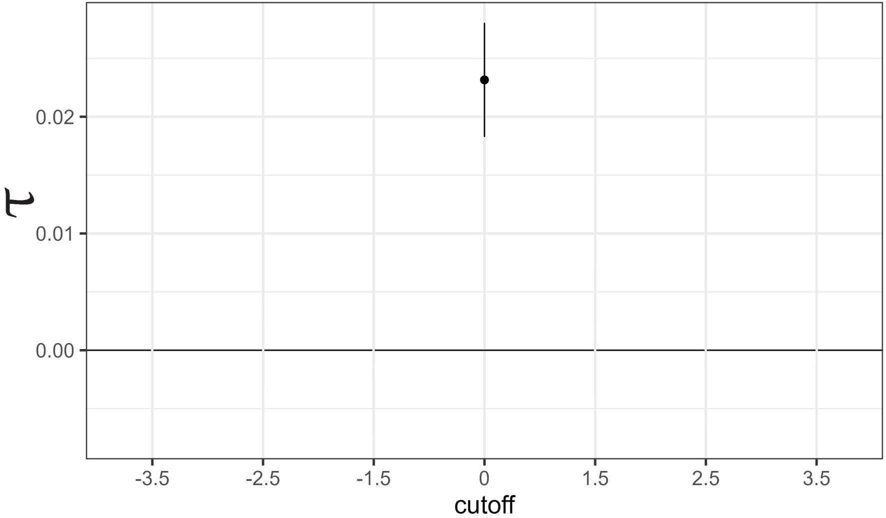
(X_i):	Female		White		FSM	
RHS of cutoff (τ)	0.032	0.023	0.019	0.012	-0.017	-0.009
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Constant	0.508	0.479	0.628	0.605	0.194	0.213
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
T_i		✓		✓		✓
N	161,982	161,982	126,818	126,818	88,100	88,100

Stack Subjects by GB (+ placebo test)

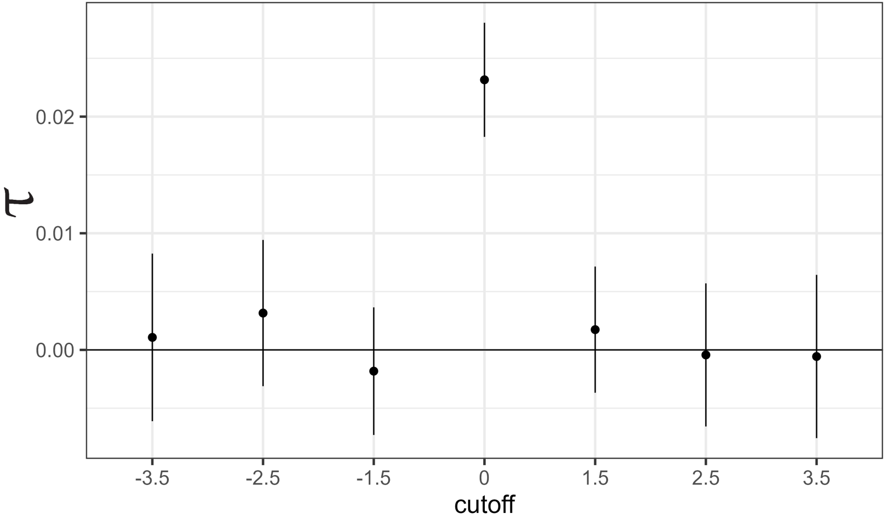
Stacking Subjects and GB by distance from GB– female



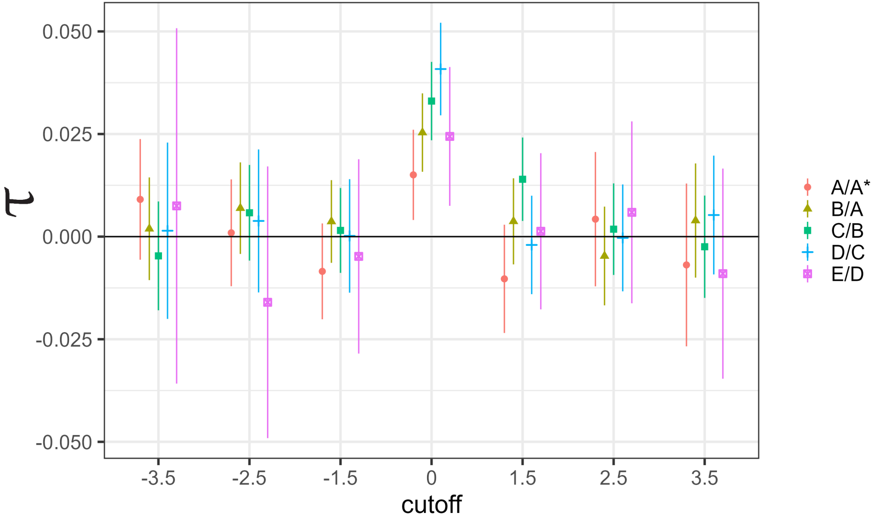
Stacking Subjects and GB by distance from GB– female



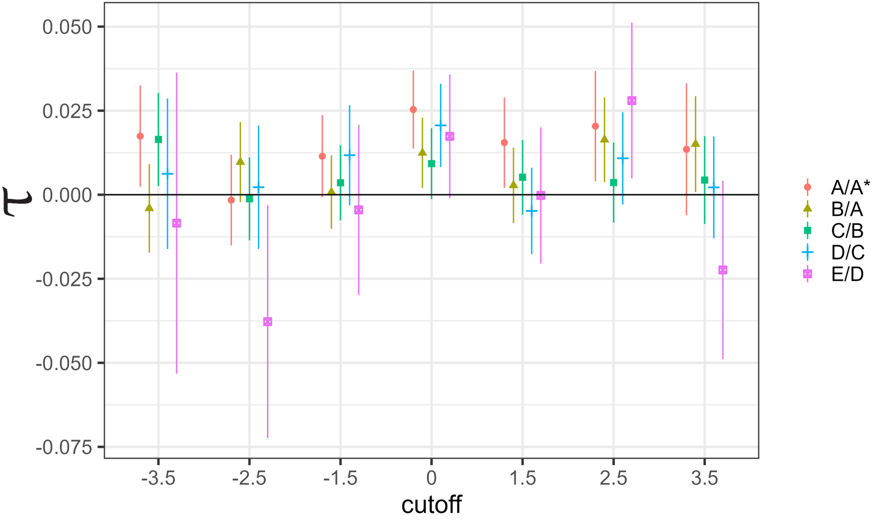
Stacking Subjects and GB by distance from GB– female



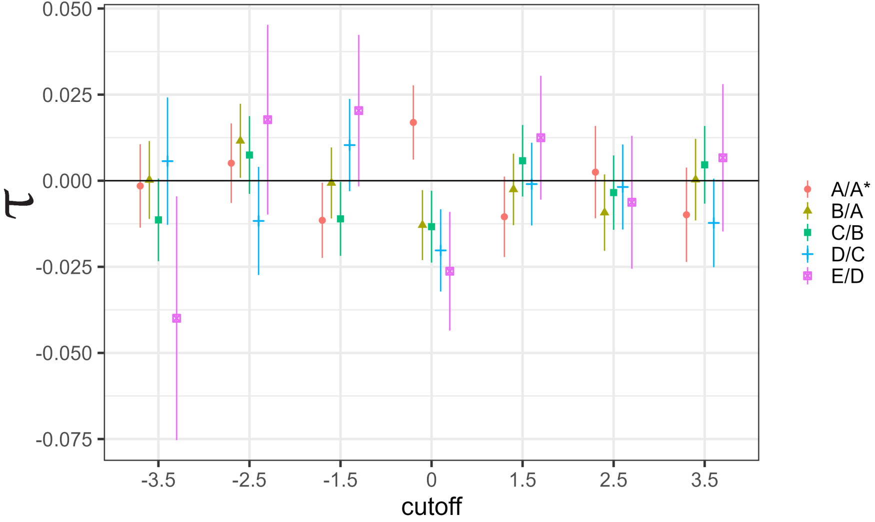
Subjects stacked by GB – Female



Subjects stacked by GB – White

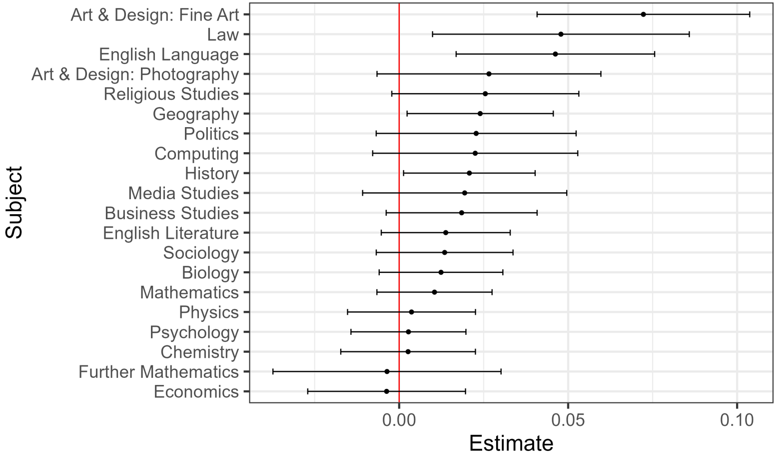


Subjects stacked by GB – FSM



Stack GBs by Subject

Stacked GBs by subject — Female



▸ Shares ▸ White ▸ FSM

Robustness

Robustness

Potential Problem

- Grade boundaries are endogenous

Robustness

Potential Problem

- Grade boundaries are endogenous
- Teachers likely put them where there are clear achievement differences

Robustness

Potential Problem

- Grade boundaries are endogenous
- Teachers likely put them where there are clear achievement differences
- This in and of itself is fine

Robustness

Potential Problem

- Grade boundaries are endogenous
- Teachers likely put them where there are clear achievement differences
- This in and of itself is fine
- The problem is the female ability distribution is to the right of the male

Robustness

Potential Problem

- Grade boundaries are endogenous
- Teachers likely put them where there are clear achievement differences
- This in and of itself is fine
- The problem is the female ability distribution is to the right of the male
    - attainment gradients: overall by subject

Robustness

Potential Problem

- Grade boundaries are endogenous
- Teachers likely put them where there are clear achievement differences
- This in and of itself is fine
- The problem is the female ability distribution is to the right of the male
  - attainment gradients: overall by subject
- Achievement gaps correlated with gender

Robustness

Potential Problem

- Grade boundaries are endogenous
- Teachers likely put them where there are clear achievement differences
- This in and of itself is fine
- The problem is the female ability distribution is to the right of the male
  - attainment gradients: overall by subject
- Achievement gaps correlated with gender
- So τ will be a combination of gender bias and achievement differences

Robustness

Require students adjacent to boundaries to have same ability

- Standard
  - Conditioning
  - Age 16 Qualifications
- Novel
  - Latent Achievement
  - Crowding

Robustness - Conditioning

Require students adjacent to boundaries to have same ability

- Standard - Conditioning
  - Condition on age 16 test scores T_i
  - Condition on age 16 test scores in respective subjects T_{is}
- Concerns
  - Age 16 are a bad predictor of future achievement
  - Assumes specific functional form specification

Robustness - Conditioning

	Main Sample		Same Subject Sample			
Female (τ)	0.032	0.023	0.035	0.029	0.024	0.024
	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)
N	161,982	161,982	76,064	76,064	76,064	76,064
White (τ)	0.019	0.012	0.022	0.015	0.014	0.013
	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)
N	126,818	126,818	61,846	61,846	61,846	61,846
FSM (τ)	-0.017	-0.009	-0.018	-0.011**	-0.008	-0.007
	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)
N	88,100	88,100	41,519	41,519	41,519	41,519
T_i		✓			✓	✓
T_{is}				✓		✓

Outcomes Age 16

Teacher assigned grades (and rankings) replace Age 16 standardised testing

Apply same approach, conditional on Age 11 standardised Test Scores
T_{is}: Maths, Reading, Writing

	Female	White	FSM
τ	0.014	0.010	-0.002
	(0.001)	(0.001)	(0.001)
$\overline{X}$	0.502	0.741	0.140
T_{is}	✓	✓	✓
N	595,186	595,186	595,186

Robustness - Latent Achievement

Rank information is ordinal Create a metric that is cardinal - Latent Achievement

- Step 1 Construct Latent Achievement

Robustness - Latent Achievement

Rank information is ordinal Create a metric that is cardinal - Latent Achievement

- Step 1 Construct Latent Achievement
  - Estimate each students probability,ϕ_{α}, of achieving grade α

Robustness - Latent Achievement

Rank information is ordinal Create a metric that is cardinal - Latent Achievement

- Step 1 Construct Latent Achievement
  - Estimate each students probability,ϕ_{α}, of achieving grade α
  - Use prior 2019 cohort to establish the mapping between GCSE and Alevels

Robustness - Latent Achievement

Rank information is ordinal Create a metric that is cardinal - Latent Achievement

- Step 1 Construct Latent Achievement
  - Estimate each students probability, ϕ_α, of achieving grade α
  - Use prior 2019 cohort to establish the mapping between GCSE and Alevels
- Step 2 Establish common support

Robustness - Latent Achievement

Rank information is ordinal Create a metric that is cardinal - Latent Achievement

- Step 1 Construct Latent Achievement
  - Estimate each students probability, ϕ_{α}, of achieving grade α
  - Use prior 2019 cohort to establish the mapping between GCSE and Alevels
- Step 2 Establish common support
  - Adjacent students similar ϕ, distributions

Robustness - Latent Achievement

Rank information is ordinal Create a metric that is cardinal - Latent Achievement

- Step 1 Construct Latent Achievement
  - Estimate each students probability,ϕ_{α}, of achieving grade α
  - Use prior 2019 cohort to establish the mapping between GCSE and Alevels
- Step 2 Establish common support
  - Adjacent students similar ϕ, distributions
- Step 3 Re-Estimate τ

Robustness - Latent Achievement

Rank information is ordinal Create a metric that is cardinal - Latent Achievement

- Step 1 Construct Latent Achievement
  - Estimate each students probability,ϕ_α, of achieving grade α
  - Use prior 2019 cohort to establish the mapping between GCSE and Alevels
- Step 2 Establish common support
  - Adjacent students similar ϕ, distributions
- Step 3 Re-Estimate τ
  - Double restriction on propensity score and adjacency

Robustness - Latent Achievement

Rank information is ordinal Create a metric that is cardinal - Latent Achievement

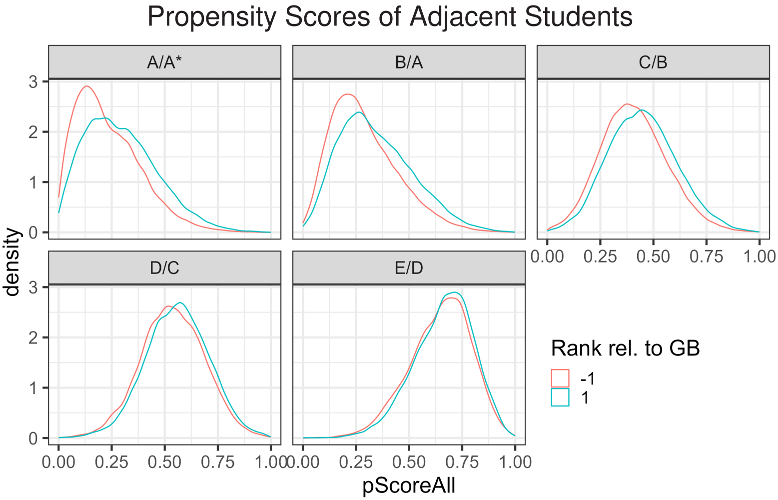
- Step 1 Construct Latent Achievement
  - Estimate each students probability,ϕ_α, of achieving grade α
  - Use prior 2019 cohort to establish the mapping between GCSE and Alevels
- Step 2 Establish common support
  - Adjacent students similar ϕ, distributions
- Step 3 Re-Estimate τ
  - Double restriction on propensity score and adjacency
  - On increasingly similar Adjacent students up to $\phi_\alpha(0.45, 0.55)$

Robustness - Latent Achievement

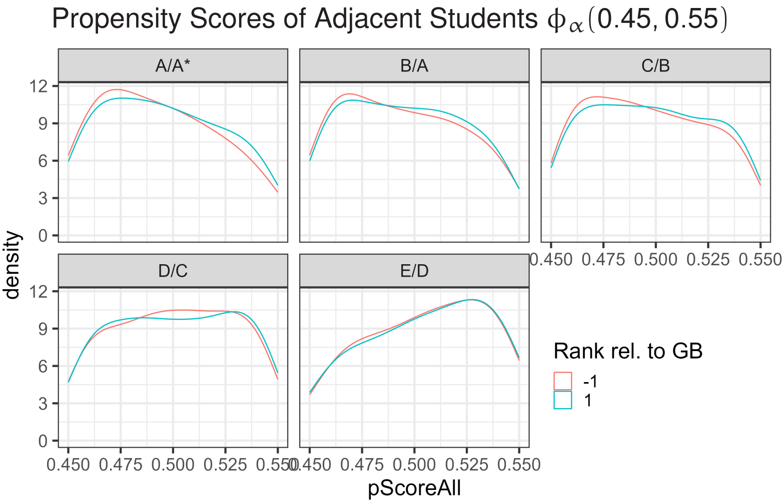
Rank information is ordinal Create a metric that is cardinal - Latent Achievement

- Step 1 Construct Latent Achievement
  - Estimate each students probability, ϕ_α, of achieving grade α
  - Use prior 2019 cohort to establish the mapping between GCSE and Alevels
- Step 2 Establish common support
  - Adjacent students similar ϕ, distributions
- Step 3 Re-Estimate τ
  - Double restriction on propensity score and adjacency
  - On increasingly similar Adjacent students up to $\phi_\alpha(0.45, 0.55)$
- Requires fewer functional form assumptions

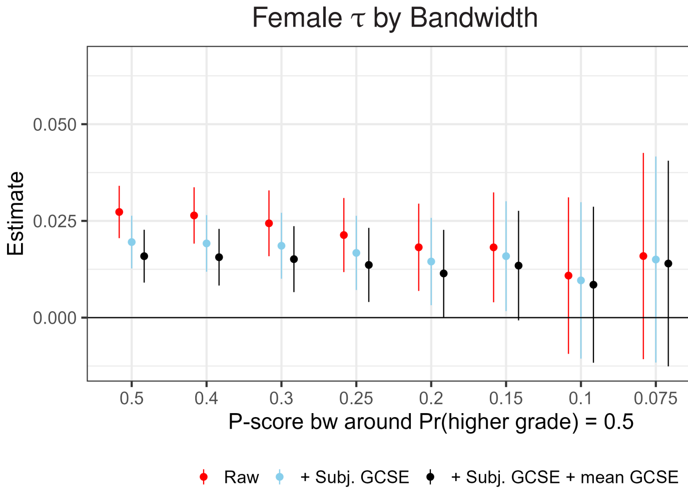
Robustness - Latent Achievement



Robustness - Latent Achievement



Robustness - Latent Achievement: Female τ



White FSM

Robustness- Crowding

Less-Parametric approach

- Main specification contains all GBs with at least one adjacent student
- However, if only single student of each grade unlikely they are of similar abilities
- Similarity of students increasing in number in adjoining grades
    - Re-estimate τ with sub-samples with increasing numbers of adjacent students

Robustness- Crowding

Less-Parametric approach

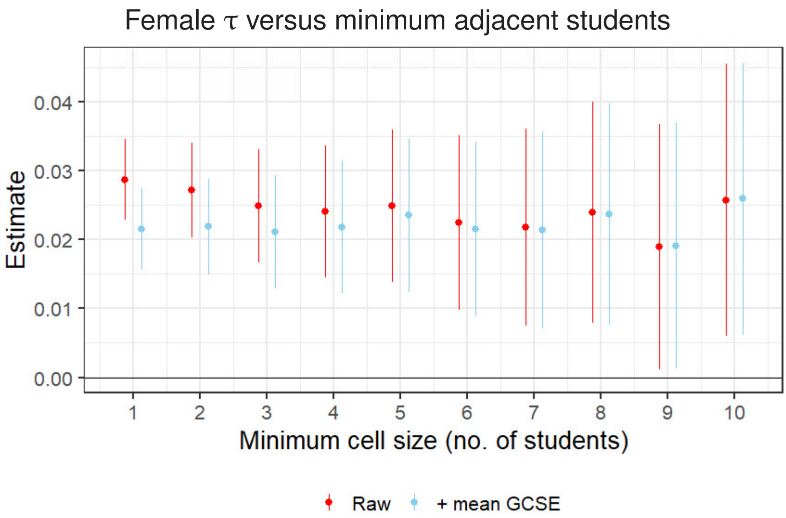
- Main specification contains all GBs with at least one adjacent student
- However, if only single student of each grade unlikely they are of similar abilities
- Similarity of students increasing in number in adjoining grades
  - Re-estimate τ with sub-samples with increasing numbers of adjacent students
  - Coefficient stability implies that τ not driven by ability

Robustness- Crowding

Less-Parametric approach

- Main specification contains all GBs with at least one adjacent student
- However, if only single student of each grade unlikely they are of similar abilities
- Similarity of students increasing in number in adjoining grades
    - Re-estimate τ with sub-samples with increasing numbers of adjacent students
    - Coefficient stability implies that τ not driven by ability
    - Does not require functional form assumptions between past and present achievement

Robustness - Crowding



White FSM

Consequences of Higher Grade

Teacher assigned grades have direct consequences on student outcomes
A-levels are the key qualifications for post-secondary

	Accepted Anywhere	Accepted First Choice	Accepted Insurance
τ			
$\overline{Y}$	0.745	0.641	0.315
N	131,174	87,298	12,125

Consequences of Higher Grade

Teacher assigned grades have direct consequences on student outcomes
A-levels are the key qualifications for post-secondary

	Accepted Anywhere	Accepted First Choice	Accepted Insurance
τ	0.021		
	(0.002)		
$\overline{Y}$	0.745	0.641	0.315
N	131,174	87,298	12,125

Consequences of Higher Grade

Teacher assigned grades have direct consequences on student outcomes
A-levels are the key qualifications for post-secondary

	Accepted Anywhere	Accepted First Choice	Accepted Insurance
τ	0.021	0.019	
	(0.002)	(0.003)	
$\overline{Y}$	0.745	0.641	0.315
N	131,174	87,298	12,125

Consequences of Higher Grade

Teacher assigned grades have direct consequences on student outcomes

A-levels are the key qualifications for post-secondary

	Accepted Anywhere	Accepted First Choice	Accepted Insurance
τ	0.021	0.019	-0.038
	(0.002)	(0.003)	(0.008)
$\overline{Y}$	0.745	0.641	0.315
N	131,174	87,298	12,125

Summary

- Provide new evidence that teachers are biased in high-stakes situations
- Validate existing estimates with new approach
- Teacher assessment exacerbating existing inequalities
- Critical to consider if moving away from external assessment
- Teacher assessment bias currently impacts
  - Coursework
  - Predicted Grades
  - GPA

Thanks!

References I

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

Allensworth, E. M. and K. Clark (2020). High school gpas and act scores as predictors of college completion: Examining assumptions about consistency across high schools. *Educational Researcher 49*(3), 198–211.

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

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

Barlevy, G. and D. Neal (2012). Pay for percentile. *American Economic Review 102*(5), 1805–1831.

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*(2), 249–263.

References II

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

Cattaneo, M. D., N. Idrobo, and R. Titiunik (2024, March). 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, July). Diversifying Society's Leaders? The Determinants and Causal Effects of Admission to Highly Selective Private Colleges.

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*.

Friedman, J. N., B. Sacerdote, D. O. Staiger, and M. Tine (2025). Standardized test scores and academic performance at ivy-plus colleges. Technical report, 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*(664), 3461–3475.

References III

Geiser, S. and M. V. Santelices (2007). Validity of high-school grades in predicting student success beyond the freshman year: High-school record vs. standardized tests as indicators of four-year college outcomes.

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*(4), 146–168.

Hinnerich, B. T., E. Höglin, and M. Johannesson (2011). Are boys discriminated in swedish high schools? *Economics of Education review 30*(4), 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*(1), 67–90.

Holbein, J. B. and H. F. Ladd (2017). Accountability pressure: Regression discontinuity estimates of how no child left behind influenced student behavior. *Economics of Education Review 58*, 55–67.

References IV

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

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

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

Lavy, V. and E. Sand (2015). On the origins of gender human capital gaps: Short and long term consequences of teachers. *stereotypical biases* (w20909)). Publisher: National Bureau of Economic Research.

Lemann, N. (2024). *Higher Admissions: The Rise, Decline, and Return of Standardized Testing*. Princeton University Press.

Lingard, B., G. Thompson, and S. Sellar (2016). National testing in schools. *Oxon, United Kingdom: Routledge*.

References V

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*(12), 1278–1288.

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

Zhu, M. (2024). New findings on racial bias in teachers' evaluations of student achievement. Technical report, IZA Discussion Papers.

Local randomisation details

RDD robustness test using "local randomisation" (Cattaneo et al., 2024)

- RD compares marginally treated to marginally untreated
- In practice: few observations close to cutoff

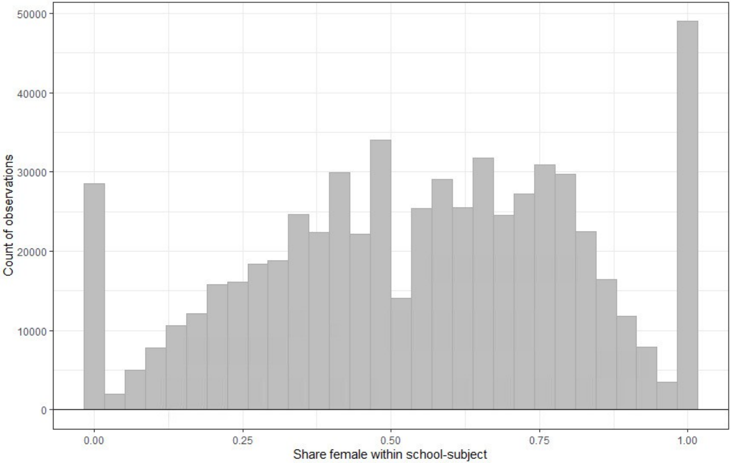
Types of RDD:

- Traditional (global): use all data to project to cutoff
- Local linear: use local data to project to cutoff
- Local randomisation: compare means of marginal students
- Better suited to discrete running variable

  $\rightarrow$ Not often used as lack of data close enough to cutoff
- we have over 200,000 observations at cutoff (ranked 1st or last)

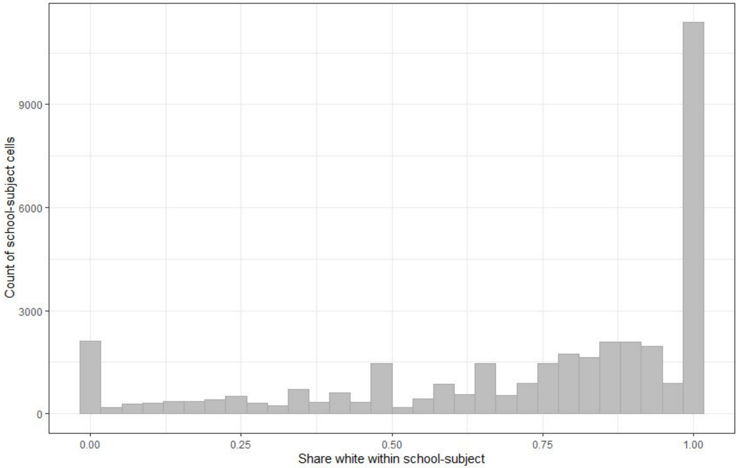
back

Female Share - Student Distribution



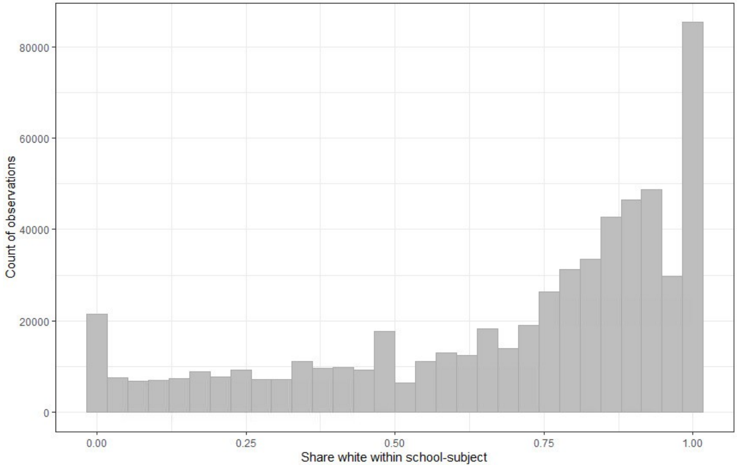
back

White Share - Boundary Distribution



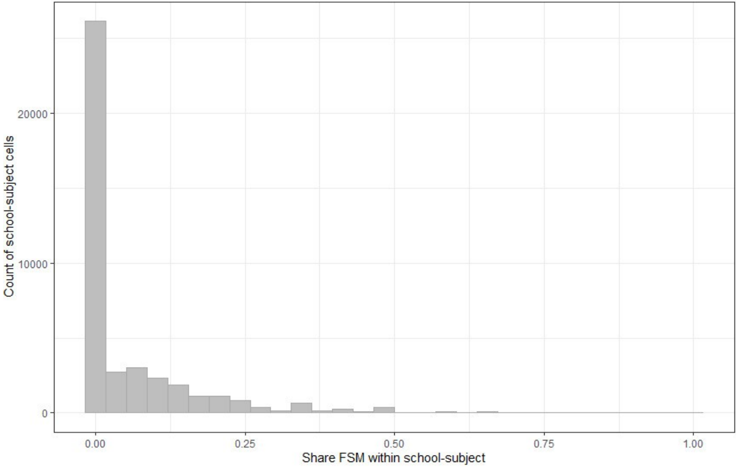
back

White Share - Student Distribution



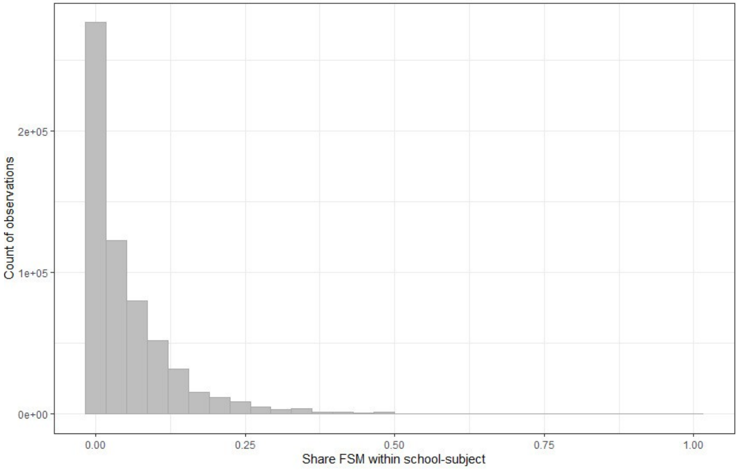
back

FSM Share - Boundary Distribution



back

FSM Share - Student Distribution



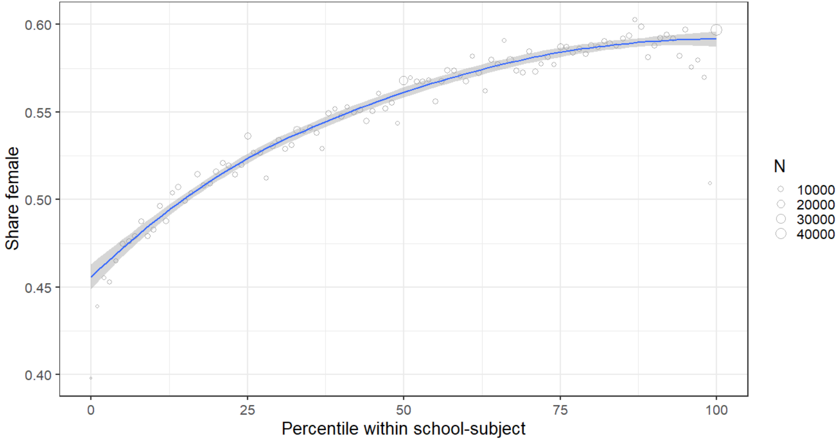
back

Top ten subjects (2020 A-level entries)

	2019	2020
Mathematics	84,552	82,774
Psychology	62,060	60,511
Biology	63,689	56,575
Chemistry	54,950	49,158
History	47,120	39,811
English Literature	37,214	36,985
Sociology	35,864	35,421
Physics	36,068	33,663
Business Studies	30,545	31,743
Economics	29,798	29,372

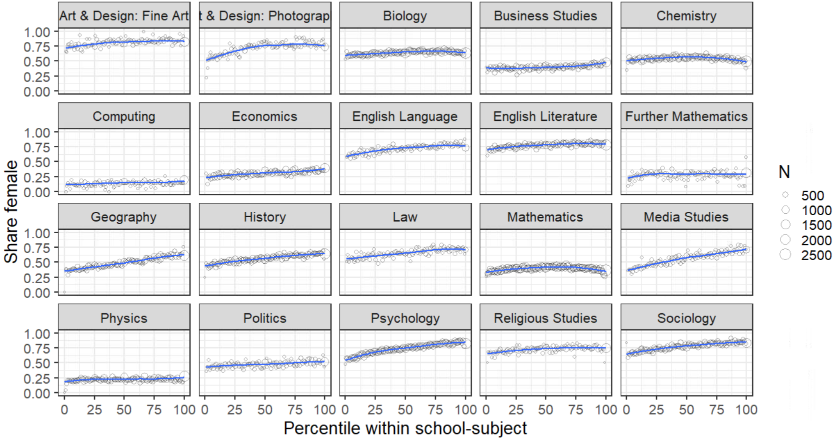
back

Attainment-gender gradients



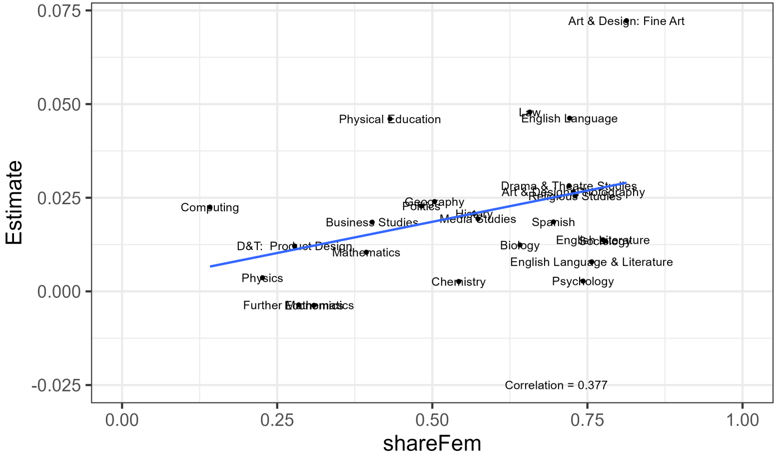
◂ back

Attainment-gender gradients by subject



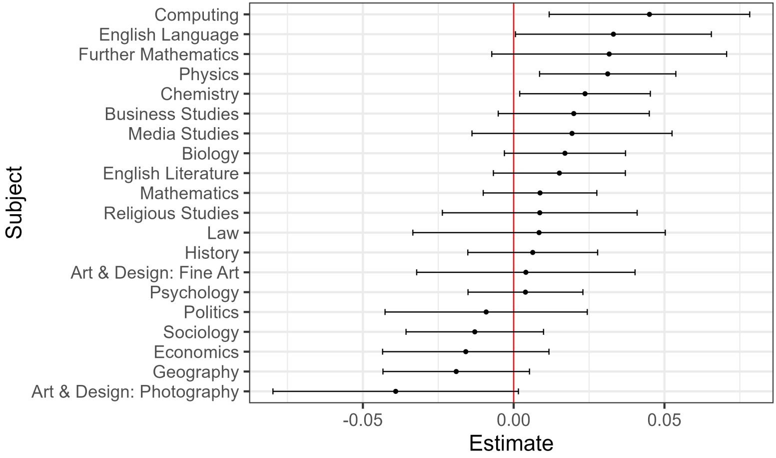
back

Female τ by subject by Proportion Female



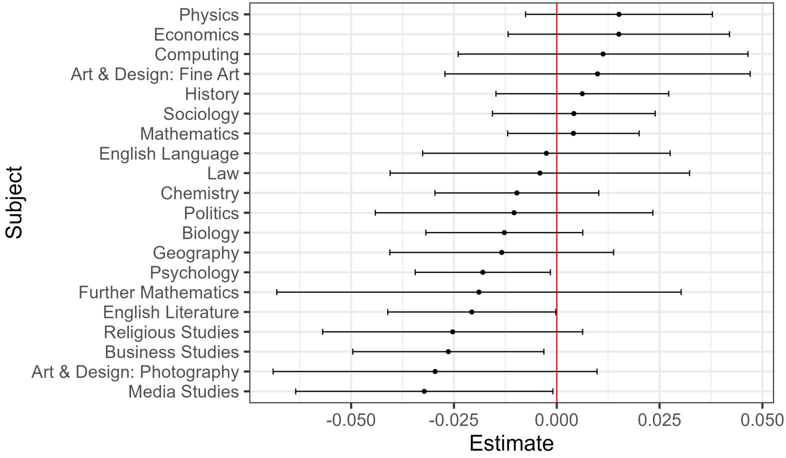
Back

Stacked GBs by subject — White



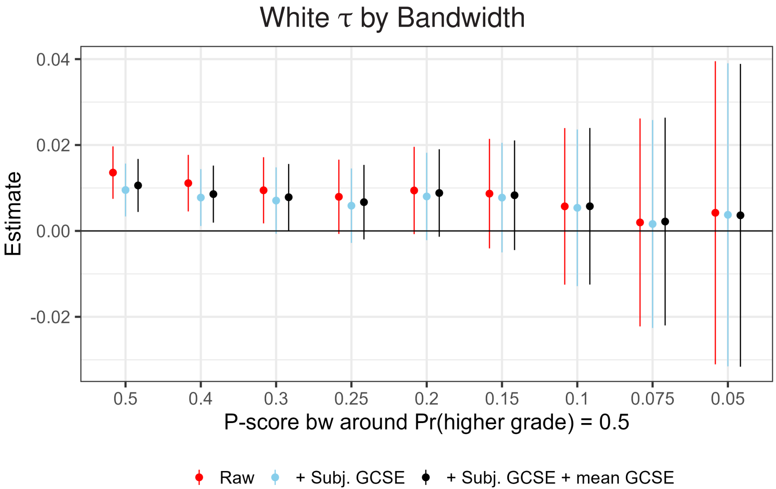
Back

Stacked GBs by subject — FSM



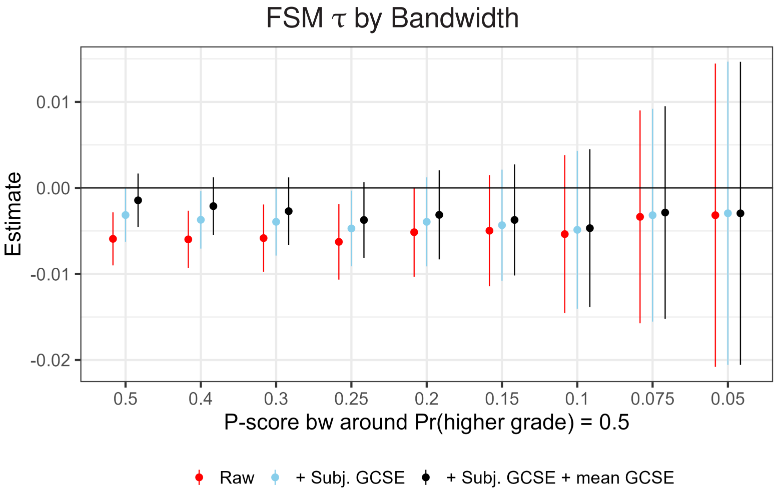
Back

Robustness - Latent Achievement: White τ



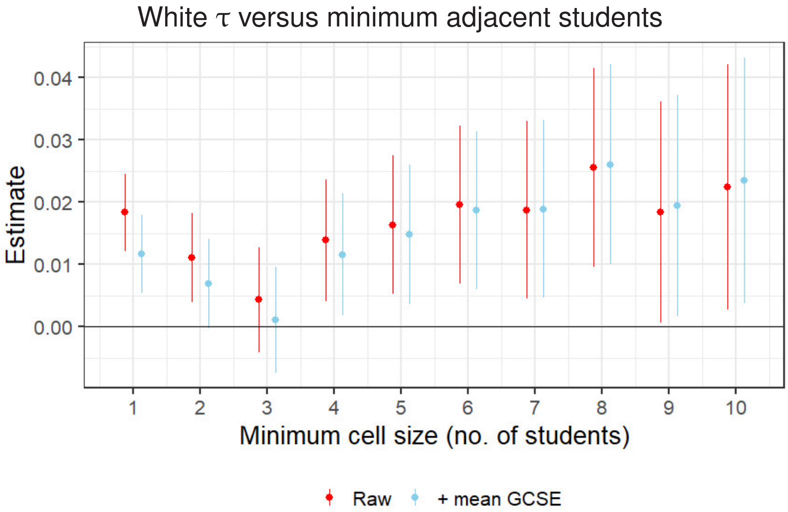
Back

Robustness - Latent Achievement: FSM τ



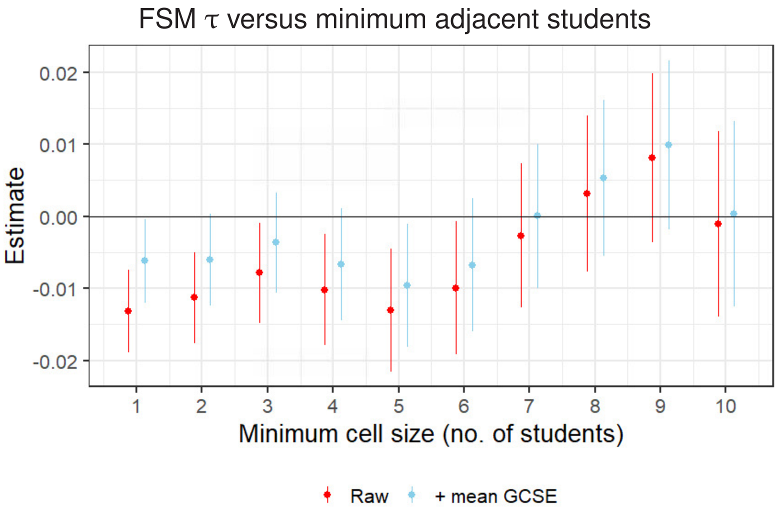
Back

Robustness - Crowding



Back

Robustness - Crowding



Back