The merits of teacher assessment versus external exams to measure student achievement

Part I (one-pager)

Teaser: (100-120 characters)

Are teachers best placed to assess their students or are external exams more effective?

Keywords: teacher assessment, grading, external exams, students, teachers

Elevator pitch: (600 characters)

There is little to no consensus in the academic literature over whether centralised, standardised exams are better for students than teacher assessments. While a growing body of evidence from economics highlights bias in teacher assessments, educationalists and psychologists point to the harm caused by high-stakes exam-related stress and argue that exams and teacher assessments generally agree very closely. This lack of academic consensus is reflected in policy: a wide variety of assessment methods are used across (and even within) countries. Policymakers should be aware of the potential for inequalities in non-blind assessments and consider carefully the consequences of relying on a single method of assessment.

Key findings (pros and cons as bullet points, max 5 each)

Pros (of teacher assessment):

- Regular contact allows teachers to form a richer picture of students' abilities.
- Teachers can assess a much broader curriculum than even the best-designed standardised tests.
- Regular opportunities for assessment reduce the randomness inherent in sitting a limited number of exams.
- Teachers can (implicitly) account for student (dis)advantage in a way not possible for standardised tests.

Cons (of teacher assessment):

- There is significant variation in the assessments different teachers make even of identical pieces of work.
- There is also (likely) significant variation in the criteria used by teachers to assess their students.
- The variation mentioned above mean teacher assessments are likely to be biased against some groups of students.
- The variation in teacher assessments makes it difficult to compare scores across teachers and schools (and can harm students assessed by particularly strict teachers).

Graphical abstract

Author's main message (600 characters)

Studies from the US and Europe have revealed evidence of gaps between teacher assessed grades and those from externally marked exams that are correlated with students' characteristics, suggestive of potential bias in teacher assessments. Several mechanisms have been explored in the literature. Teachers may use factors other than ability such as behavior, or may favor types of students who have performed well in previous years, or those who are among the minority group (e.g. gender) in a field. In terms of consequences, a some studies shows that biased teacher assessments in certain subjects, such as maths, can impact pupil's progress, their choice of academic track, and even their degree choices.

Part II

Motivation (1,000 characters)

Recent years have seen a growing debate over the merits of teacher assessment versus externally marked exams. A leading example of this debate is the role of SATs versus GPA in US college admissions [1]. SATs were vilified as a "wealth test" because of a high correlation between SAT scores and parental income [2]. However, despite many selective colleges going 'test-optional' during the Covid-19 pandemic, one study suggests that while SATs might be unfair to disadvantaged students, the (current) alternatives are even more damaging for inequality [3].

The issue is not unique to the US. In the UK there was a recent government consultation over whether the current university admissions system – which relies on teacher-assessed "predicted" grades – should be scrapped in favor of a post-qualifications application system. The chief driver of this reform was concerns around the fairness and accuracy of predicted grades, 86% of which are inaccurate [4].

Another reason for standardised, external exams is that it helps to keep teachers and schools accountable. If the only measure of a student's performance is an assessment by their teacher, comparisons across teachers and schools become difficult at best, and impossible if teachers act in their own self-interest. Standardised exams provide a metric by which teachers' and schools' performances can be assessed, improving accountability. On the other side of the debate, education researchers and psychologists have raised concerns that high-stakes exams lead to significant test anxiety, harming students' wellbeing. These two aspects are beyond the scope of this article, instead, this article will focus on the existence of bias in teacher assessment, potential mechanisms behind this, and the consequences of mis-assessment.

Discussion of pros and cons (18,000 characters)

Proponents of teacher assessment suggest that the regular contact teachers have with their students allows them to form a "richer picture of what students know and can do than tests alone" [7] p. 78. Even the best-designed standardised tests are limited, both in the extent of abilities they can assess, and the time in which they implement these assessments [8].

Conversely, perhaps the chief criticism of teacher assessments is their variability, both in the scores they assign to students, and particularly in the criteria they consider in reaching these scores. Although variability in individual markers assigned scores is still an issue with external tests, blind-marking can ensure that markers only see the information they are supposed to use to grade the students work, and moderation and double-marking can reduce the variability.

Despite general agreement that teacher assessments and external exams measure different underlying traits in general, the majority of the literature assumes that teacher assessed grades and standardised tests are attempting to measure the same thing, i.e. teachers (aim to) use the same criteria that are measured by standardised tests. A move towards standards-based grading (SBG) has helped teachers align more closely with external tests. In SBG, student grades are only judged against specific criteria, or standards, and other judgements such as behavior are reported separately. However, even under such an assumption, the extent of the

Figure 1 - Excerpt from Starch and Elliott (1912, p. 451)

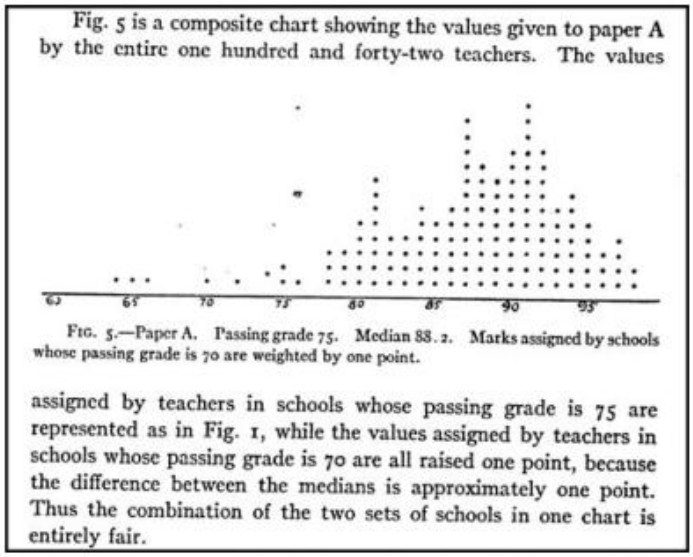


Fig. 5 is a composite chart showing the values given to paper A by the entire one hundred and forty-two teachers. The values

FIG. 5.—Paper A. Passing grade 75. Median 88.2. Marks assigned by schools whose passing grade is 70 are weighted by one point.

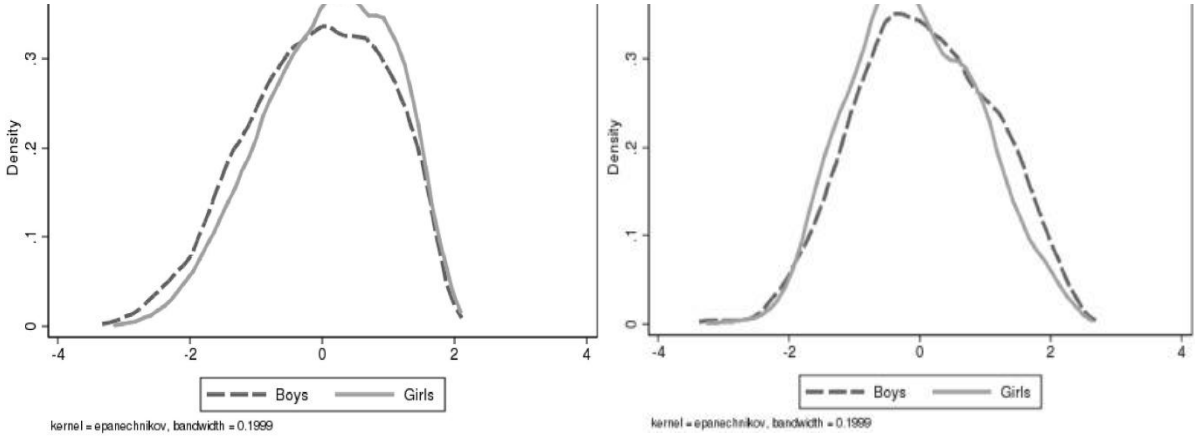
assigned by teachers in schools whose passing grade is 75 are represented as in Fig. 1, while the values assigned by teachers in schools whose passing grade is 70 are all raised one point, because the difference between the medians is approximately one point. Thus the combination of the two sets of schools in one chart is entirely fair.

variability in teacher assessments is large. Figure 1 shows the results of an experiment from 1912 in which 142 teachers were asked to mark the same paper [9]. The variation in scores is striking and highlights the need for standardisation of marks even when exams are marked blind – something that is only possible with externally-set, standardised exams. Another study replicated this experiment with high-school teachers in a single US district [10]. The scores from 73 teachers who graded the same paper on a 0-100 scale ranged from 50 to 96. This variability is likely to be problematic when grades are particularly high stakes.

Another concern is that when teachers possess additional information about the pupil, they will use this in their assessment. Several authors have compared students' performances in standardised tests to their teacher-assessed grades under the assumption that, on average across the population, standardised tests provide a "true" measure of their attainment. Their findings suggest that teachers appear to use information other than "pure attainment" (that is attainment measured by a standardised test) to assess students, and that the differences

between teacher-assessed and external grades vary systematically with student characteristics, such as race and gender. The next paragraphs discuss some of the evidence in more detail, evidence which generally relies on similar methods[1] and a key assumption: that non-biased teacher assessments (non-blind) and external exams (blind or quasi-blind) are directly comparable and hence should be the same *on average*.

1 Distribution of non-blind (left) and blind (right) test scores in mathematics (France, Grade 6). Boys outperform girls in blind test scores (dotted line is to the right of the solid line), while the reverse is true for non-blind scores, suggesting non-blind assessments favor girls. [Source: Terrier (2020)]



Notes: Test scores are standardized within test (blind or non-blind) and subject so the mean equals 0 and the variance equals 1 (Terrier, 2020, p. 4).

Many studies focus on differences by gender. Most studies find that boys and girls with the same standardised test scores are graded differently, with girls being awarded more generous grades by their teachers. Girls were shown to be favored in all subjects at primary schools in the US [11], while this was only true in mathematics in a study of middle schools in France [23]. Another study exploits evidence from high-school matriculation exams in Israel, where students are assessed using both external blind exams and within-school exams and finds that female students are favored in teacher assessments [12]. Analysing Swedish data, one author also finds a bias against male students when comparing the teacher-assessed "School-Leaving Certificate" to results from standardised tests [13]. And yet another study finds that boys do relatively better in externally blind-marked central exit exams in Norway than in teacher assessments [14].

Although a large part of the literature has focused on gender, there is also evidence that teacher assessments vary with other characteristics too. For example, one study finds that teachers grade students differently based on their ethnicity [15]. Similarly, another study reports that teachers in Brazil award lower maths assessments to black students than to their white peers with the same test scores [16]. Two authors run a field experiment in India, finding that teacher assessments vary by the *caste* of the exam taker, and suggest this is due to statistical discrimination on the part of the teacher [17]. In Italy, one study finds that teachers award immigrant students lower grades than to their native peers of similar (tested) ability [18].

The literature on teacher bias generally assumes that teacher-assessed grades follow standards-based grading (SBG). This is a conceptual approach that allows for both an easily comparable summary attainment grade and the recognition of other factors, where grades are based on certain standards of achievement, and other factors such as effort and behavior are reported separately. Hence teacher-assed grades that follow SBG are measuring the same underlying ability (or achievement) as standardised tests. For example, many of the papers discussed in this article argue that teachers are told to use standards-based grading (see e.g. [12] and [13]), with one study providing a link to "online [materials] to support teachers in 'aligning their judgments systematically with national standards'." [15, p. 540].

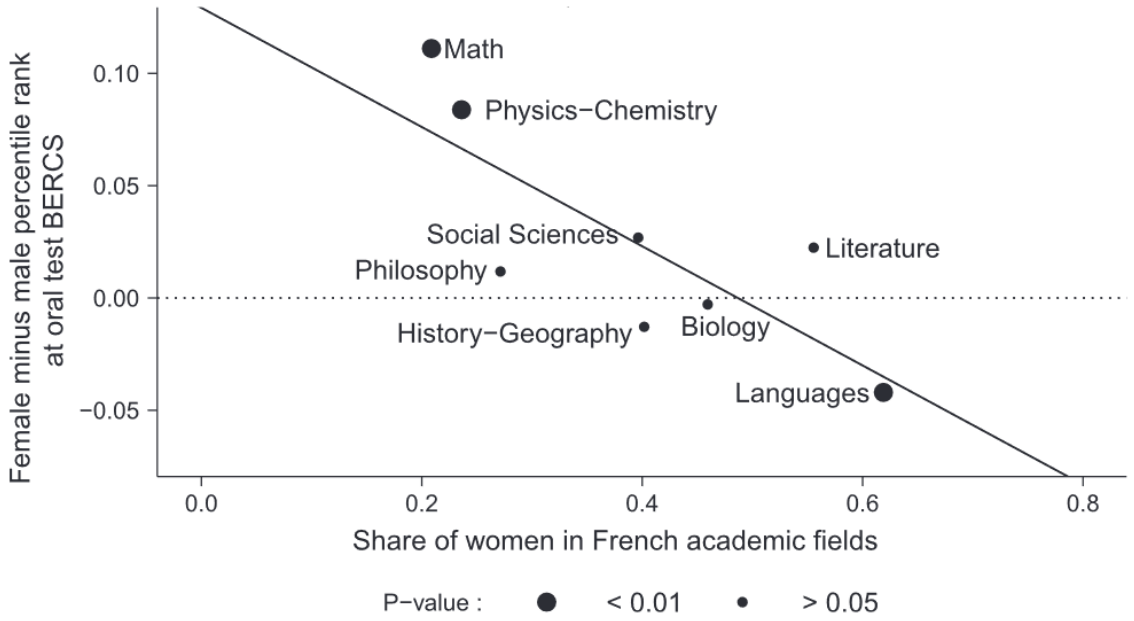
Mechanisms

Many of the studies mentioned above have also explored the possible mechanisms behind their findings. One study assesses whether student behavior has an impact on teacher assessments, finding that teachers inflate the scores of better-behaved students [19]. They find that this explains a large part of the gap between genders in their results. The authors of another study also point to the same mechanism for the gender gap, finding that when non-cognitive skills are accounted for, there is no difference in how teachers grade boys versus girls [11]. The authors of yet another study find that their results are driven by student-teacher interactions during coursework which favor girls in teacher-assessed grades [14].

Conversely to [11], one study rules out differences in behavior as a driver of differences across genders and asserts that they are due to differences in *teacher* behavior, i.e. to gender-based discrimination [12]. Others also test several mechanisms, finding that a stereotype model fits best (although they do not rule out the other models completely) [15]. Their analysis suggests that groups who had performed well in previous years were favored in teacher assessments – i.e. a teacher will categorise students and create prototypes or exemplars to make conscious or unconscious judgments about future students of the same group. Yet another study also provides support for the role of teacher stereotypes, finding that teachers' immigrant-native bias is reduced when they are informed about their own stereotypes [18].

Another possible mechanism uncovered is suggested in two studies, that find a correlation between the degree of male-dominance in STEM fields and the pro-female bias in non-blind oral exams relative to gender-blind written exams, suggesting examiners are favoring those among the minority gender in a field [20], [21]. Figure 2 shows the correlation between the share of females in a field, and the relative performance of males versus females by field, in a

Figure 2 -- Female advantage or disadvantage on a test which is identical across fields [Source: Breda and Hillion (2016)]



Notes: The difference between women's and men's average rank on the oral test "Behave as an ethical and responsible civil servant" in the different subject-specific medium-level exams (y axis). The size of each point indicates the extent to which it is different from 0 (P value from Student's t test). Any fields' extent of (non-)male-domination measured by the share of women academics in each field (x axis) (Breda and Hillion, 2016, p. 476).

test that is common to all fields.

While there is considerable evidence and consensus that teacher assessments are biased towards certain groups, there is much less evidence (and consensus) on the causes of these biases. Further research is required to better understand the causes of these biases and to understand the contexts in which they occur. Policymakers can aid this research by ensuring the data required to perform these analyses is available to researchers. For example, to understand whether bias in teacher assessments is correlated with characteristics of teachers themselves, researchers need access to data not only on students but also on the characteristics of their teachers, data which is often unavailable.

Consequences of teacher mis-assessment

But do these divergences between teacher assessment and external exams matter for students' later outcomes? Evidence on this question is limited, but a small number of studies have attempted to answer this question, comparing outcomes of those who were over- or under-assessed by their teachers (again by comparison with external exams). One study suggests

pupils might be placed in a lower secondary school "set" due to underassessment, harming their future outcomes and motivation [15]. Another study finds that underassessed children are less likely to enrol in ambitious high-school tracks, with underassessment also being a key contributor to the migrant-enrolment gap in high school [22].

One author compares blind and non-blind test scores among high school pupils in grades 6-11, exploiting quasi-random assignment of pupils to biased teachers [23]. She finds that teachers' gender biases in French and maths impacts pupils future progress and their likeliness to choose certain high school tracks, but in ways that are not straightforward. Bias against boys in maths does not impact boys progress in maths, but it does improve girls progress – while bias against boys in French reduces their progress. Bias in favor of girls in maths increases girls' probability of selecting a scientific track in high school.

One study finds that underassessment can impact students' future degree choices. The authors exploit a situation in Denmark where high school students are randomly allocated to an external exam in one subject (as well as being teacher assessed) [24]. They find that girls do relatively better in the maths exam than when there is teacher assessment, versus boys, and that assignment to the maths exam therefore reduces the gender gap in maths degree uptake.

Limitation and gaps (1,000 characters)

Investigating divergences between teacher assessment and external exams requires students to be assessed by both teachers and external exams at similar ages and educational stages, and for the researcher to observe test scores from both sources. But it also relies on a key assumption – that external exams provide an accurate measure of student attainment, and teacher assessments *should* be close to this measure. A failure of this assumption to hold could invalidate many of these findings, and it cannot be tested in most cases. Another issue is that different commonly used methods are often assumed to be directly comparable when in fact they are not [25]. The authors show how differences in the gender distributions of test scores can lead to biased estimates of bias.

Apparently, the only study that does not rely on this assumption is a field experiment in India in which the authors randomly assigned child characteristics (including gender and caste) to the cover page of the exam sheet, before teachers graded them [17]. This ensured there could be no systematic relationship between the observed characteristics and the exam quality, meaning any effect of the characteristics on test scores must be down to bias. However, the setting of this study in a developing country with quite different cultural and educational values makes it unclear if these results would hold in Western developed countries, such as the UK and the US. Therefore, investigating teacher bias in developed countries using alternative methods which rely on different assumptions is an important goal for future research.

Summary and policy advice (1,000 characters)

Studies from the US and Europe have revealed evidence of divergence between teacher assessment and externally marked exams. This implies that teachers may use criteria other than student ability when grading exams.

Much of the literature investigating gaps in teacher assessment by student characteristic has focused on gender, finding that girls are typically more likely to be favored by teachers. A smaller number of studies have focused on other characteristics, such as ethnicity, but there remains a lack of evidence on discrepancies in teacher judgement by socio-economic status.

Why do these discrepancies in teacher judgements versus exams exist? Evidence suggests factors other than pupil ability such as behavior impacts non-blind teacher assessments, or certain types of students may be favoured, such as those who have performed well in previous years, or who are among the minority in a field.

But does this matter? Evidence on the extent to which using teacher judgements versus external exams has consequences for student outcomes is limited. However, a small number of articles have shown that biased teacher assessments in certain subjects, such as maths, can impact pupil progress, their choice of academic track, and even their degree choices. More research in this area is needed.

Policymakers should be aware of the consequences of using teacher assessment versus externally marked exams. Moving away from external examinations – as some countries have discussed doing – may result in examination not reflecting the true ability of students, and may benefit certain student types over others, with long run consequences for inequality. Solutions include giving teachers better training on what assessment should include (and not include) or giving teachers training in unconscious bias, which has been show to successfully reduce biases [18]. However, if the aim is to award grades purely on the basis of ability, the evidence shows that a system of externally set and marked exams is likely to be the best way of achieving this.

Key References

Alesina, A. *et al.* (2024) 'Revealing Stereotypes: Evidence from Immigrants in Schools', *American Economic Review*, 114(7), pp. 1916–1948. Available at: https://doi.org/10.1257/aer.20191184. [18]

Breda, T. and Hillion, M. (2016) 'Teaching accreditation exams reveal grading biases favor women in male-dominated disciplines in France', *Science*, 353(6298), pp. 474–478. Available at: https://doi.org/10.1126/science.aaf4372. [21]

Brimi, H.M. (2011) 'Reliability of Grading High School Work in English', *Practical Assessment, Research, and Evaluation*, 16(1)(17). Available at: https://doi.org/10.7275/j531-fz38. [10]

Brookhart, S.M. (2013) 'The use of teacher judgement for summative assessment in the USA', *Assessment in Education: Principles, Policy & Practice*, 20(1), pp. 69–90. Available at: https://doi.org/10.1080/0969594X.2012.703170. [7]

Burgess, S. *et al.* (2022) 'The importance of external assessments: High school math and gender gaps in STEM degrees', *Economics of Education Review*, 88, p. 102267. [24]

Burgess, S. and Greaves, E. (2013) 'Test Scores, Subjective Assessment, and Stereotyping of Ethnic Minorities', *Journal of Labor Economics*, 31(3), pp. 535–576. Available at: https://doi.org/10.1086/669340. [15]

Cornwell, C., Mustard, D.B. and Parys, J.V. (2013) 'Noncognitive Skills and the Gender Disparities in Test Scores and Teacher Assessments: Evidence from Primary School', *Journal of Human Resources*, 48(1), pp. 236–264. Available at: https://doi.org/10.3368/jhr.48.1.236. [11]

Falch, T. and Naper, L.R. (2013) 'Educational evaluation schemes and gender gaps in student achievement', *Economics of Education Review*, 36, pp. 12–25. Available at: https://doi.org/10.1016/j.econedurev.2013.05.002. [14]

Ferman, B. and Fontes, L.F. (2022) 'Assessing knowledge or classroom behavior? Evidence of teachers' grading bias', *Journal of Public Economics*, 216, p. 104773. Available at: https://doi.org/10.1016/j.jpubeco.2022.104773. [19]

Hanna, R.N. and Linden, L.L. (2012) 'Discrimination in Grading', *American Economic Journal: Economic Policy*, 4(4), pp. 146–168. Available at: https://doi.org/10.1257/pol.4.4.146. [17]

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), pp. 2083–2105. Available at: https://doi.org/10.1016/j.jpubeco.2008.02.009. [12]

Lindahl, E. (2007) *Comparing teachers' assessments and national test results - evidence from Sweden*. 2007:24. IFAU - Institute for Evaluation of Labour Market and Education Policy. Available at: https://www.ifau.se/en/Research/Publications/Working-papers/2007/Comparing-teachers-assessments-and-national-test-results---evidence-from-Sweden/ (Accessed: 13 June 2024). [13]

Starch, D. and Elliott, E.C. (1912) 'Reliability of the Grading of High-School Work in English', *The School Review*, 20(7), pp. 442–457. [9]

Terrier, C. (2020) 'Boys lag behind: How teachers' gender biases affect student achievement', *Economics of Education Review*, 77, p. 101981. [23]

Additional References

Botelho, F., Madeira, R.A. and Rangel, M.A. (2015) 'Racial Discrimination in Grading: Evidence from Brazil', *American Economic Journal: Applied Economics*, 7(4), pp. 37–52. Available at: https://doi.org/10.1257/app.20140352.

Brookhart, S.M. *et al.* (2016) 'A Century of Grading Research: Meaning and Value in the Most Common Educational Measure', *Review of Educational Research*, 86(4), pp. 803–848.

Chetty, R., Deming, D.J. and Friedman, J.N. (2023) 'Diversifying Society's Leaders? The Determinants and Causal Effects of Admission to Highly Selective Private Colleges'. National Bureau of Economic Research (Working Paper Series). Available at: https://doi.org/10.3386/w31492.

Delaney, J.M. and Devereux, P.J. (2023) 'Gender Differences in Teacher Judgement of Comparative Advantage'.

von der Embse, N. *et al.* (2018) 'Test anxiety effects, predictors, and correlates: A 30-year meta-analytic review', *Journal of Affective Disorders*, 227, pp. 483–493. Available at: https://doi.org/10.1016/j.jad.2017.11.048.

Jürges, H., Schneider, K. and Büchel, F. (2005) 'The Effect of Central Exit Examinations on Student Achievement: Quasi-Experimental Evidence from Timss Germany', *Journal of the European Economic Association*, 3(5), pp. 1134–1155. Available at: https://doi.org/10.1162/1542476054729400.

Leonhardt, D. (2024) 'The Misguided War on the SAT', *The New York Times*, 7 January. Available at: https://www.nytimes.com/2024/01/07/briefing/the-misguided-war-on-the-sat.html (Accessed: 14 February 2024).

Murphy, R. and Wyness, G. (2020) 'Minority report: the impact of predicted grades on university admissions of disadvantaged groups', *Education Economics*, 28(4), pp. 333–350. Available at: https://doi.org/10.1080/09645292.2020.1761945.

Rangvid, B.S. (2015) 'Systematic differences across evaluation schemes and educational choice', *Economics of Education Review*, 48, pp. 41–55. Available at: https://doi.org/10.1016/j.econedurev.2015.05.003.

Rimfeld, K. *et al.* (2019) 'Teacher assessments during compulsory education are as reliable, stable and heritable as standardized test scores', *Journal of Child Psychology and Psychiatry*, 60(12), pp. 1278–1288. Available at: https://doi.org/10.1111/jcpp.13070.

Woessmann, L. (2018) 'Central exit exams improve student outcomes', *IZA World of Labor* [Preprint]. Available at: https://doi.org/10.15185/izawol.419.

Zwick, R. (2002) 'Is the SAT a "Wealth Test"?', *Phi Delta Kappan*, 84(4), pp. 307–311. Available at: https://doi.org/10.1177/003172170208400411.