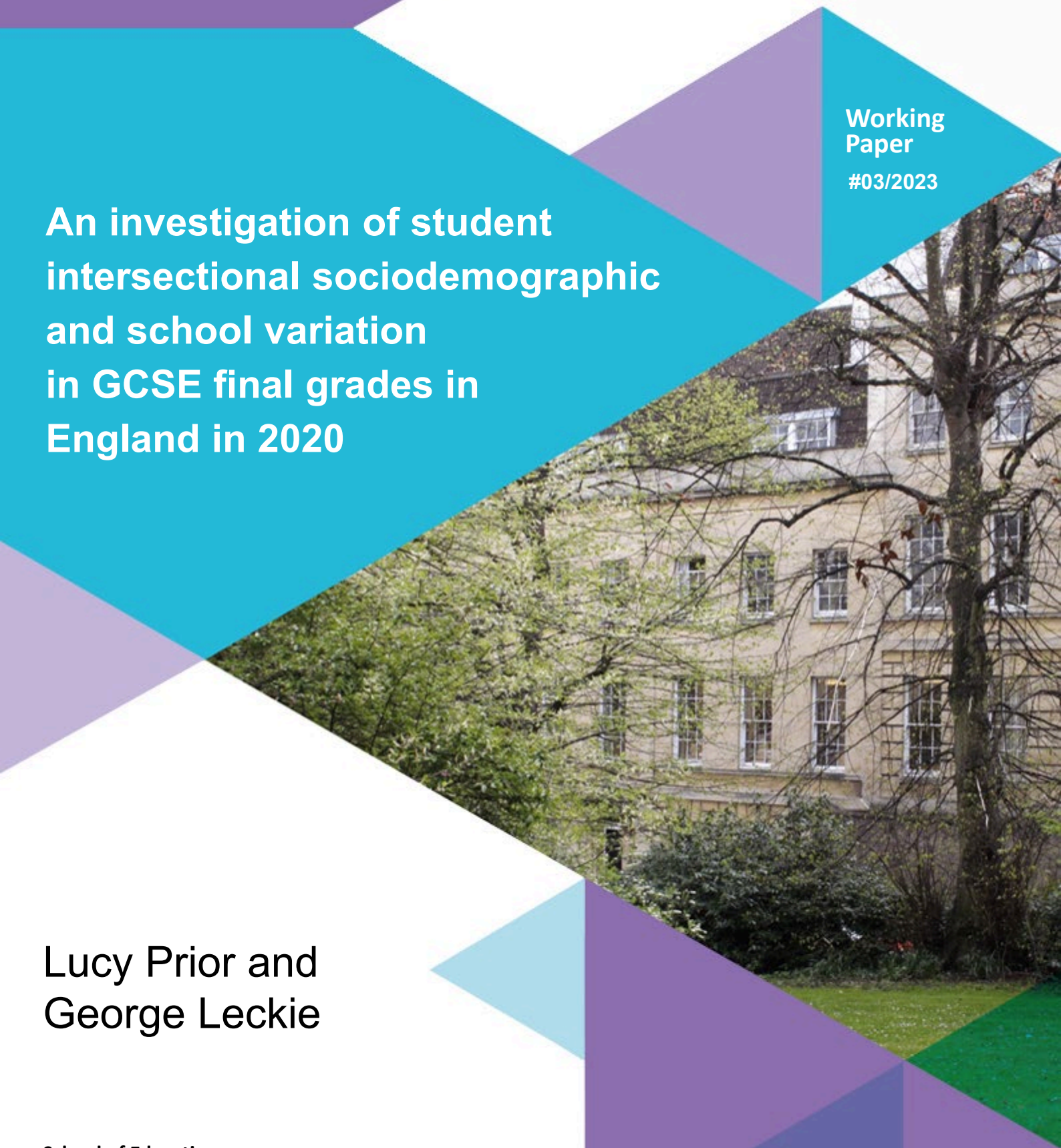
Bristol Working Papers
In Education Series

University of BRISTOL
School of Education

Working Paper
#03/2023

An investigation of student intersectional sociodemographic and school variation in GCSE final grades in England in 2020



Lucy Prior and
George Leckie

School of Education
University of Bristol
35 Berkeley Square
Bristol BS8 1JA

bristol.ac.uk/education/research/publications

Bristol Working Papers in Education #03/2023

An investigation of student intersectional sociodemographic and school variation in GCSE final grades in England in 2020

By Lucy Prior and George Leckie

Centre for Multilevel Modelling and School of Education, University of Bristol
September 2023

Corresponding author:
Dr Lucy Prior
lucy.prior@bristol.ac.uk
Telephone: +44 117 455 2823

University of Bristol,
35 Berkeley Square, Bristol,
BS8 1JA

Funding
This work was funded by Economic and Social Research Council (ESRC) grants ES/W000555/1 and ES/X011313/1.

Acknowledgements
This work was produced using statistical data from ONS which is Crown Copyright. The use of the ONS statistical data in this work does not imply the endorsement of the ONS in relation to the interpretation or analysis of the statistical data. This work uses research datasets which may not exactly reproduce National Statistics aggregates.

Disclosure statement
The authors declare no conflicts of interest.

Data availability
The dataset used in this analysis is available on application from the Office for National Statistics, doi: https://doi.org/10.57906/k68n-bt74.

University of BRISTOL

SCHOOL OF EDUCATION

An investigation of student intersectional sociodemographic and school variation in GCSE final grades in England in 2020

Abstract

In 2020, COVID-19 forced the cancellation of all student end-of-school examinations in England. Schools were asked to provide centre assessment grade*s* (CAGs), offering their best estimates for what students would have achieved had they sat their examinations. Although initially ignored in favour of grades calculated via an algorithm, students were eventually awarded their CAGs following widespread public outcry over the calculated grades. Whether CAGs were unfairly awarded across different student groups and schools in 2020 compared to previous years is a key question. However, existing analyses of bias in CAGs are limited by a lack of attention to potential interactions between student characteristics, and thus to hidden differential grade inflation across intersectional groups. We address this by examining student GCSE performance in 2018, 2019, and 2020 via a Multilevel Analysis of Individual Heterogeneity and Discriminatory Accuracy (MAIHDA) analysis of intersectional sociodemographic variation which we cross-classify with schools given their role in generating CAGs. Overall, a picture of stability emerges, where despite substantial overall grade inflation in 2020, the use of CAGs does not appear to have generated new or divergent intersectional relationships in comparison to previous years, suggesting CAGs showed a similar susceptibility to bias as normal examinations.

Keywords

Intersectionality, examinations, schools, bias, GCSE, MAIHDA

Introduction

Context

In 2020, COVID-19 caused widespread disruption to the educational system in England, and the decision was taken to cancel GCSE (age 15/16) and A-level (age 17/18) national examinations (GOV.UK, 2020). The Department for Education (DfE) and the exams regulator (Ofqual) were tasked with providing an alternative solution for student grades, enabling students to progress to the next stage of education, university, or other destinations. Schools were asked to provide their best estimate of the grade that students would have gained had they sat their examination (centre assessment grade, CAG) (Ofqual, 2020a). However, CAGs were overly optimistic and tasked with preventing grade inflation and maintaining grading consistency across years, Ofqual initially replaced CAGs with 'calculated' grades derived using their Direct Centre-level Performance (DCP) algorithm (Kelly, 2021; Ofqual, 2020b). This led to 40% of results being downgrading by one or more grades, and a public furore with media reports of bias and that students were 'robbed' of their deserved grades (BBC News, 2020; Guardian, 2020a; 2020b; Kelly, 2021). As a result of this outcry, the government reverted to using CAGs (or the calculated grade where this was higher) (Ofqual, 2020c).

Why investigate potential biases?

Like examinations in normal years, CAGs will directly impact the destinations of students, whether this be further study at A-level, apprenticeships, university, or entry into the labour market. It is important to understand potential biases as any distortions in grades may have affected the relative (dis)advantage of student groups in terms of their destinations. For instance, the grade inflation of CAGs presented universities with an oversubscription problem in 2020 with many more disadvantaged students meeting the criteria for entry than in usual years (Kelly, 2021). Given the longer-term impacts of student grades on employment opportunities, income, and associated social status (Boliver, 2011; Murphy and Wyness, 2020), understanding potential biases arising from the use of CAGs will provide insight into future social mobility patterns and wider social inequalities. Additionally, evaluating the 2020 grades could offer lessens for any future scenarios that may disrupt

examinations (e.g., future pandemics, teacher strikes, exam boycotts, centre malpractice etc.) and lead to the use of teacher judgements in the future.

Studying CAGs can also inform broader debates around teacher assessments and teacher biases. Teacher assessments are widely used in educational systems for summative assessment, in England and internationally (Harlen, 2005). For example, the predicted grades schools provide as part of university applications (UCAS, 2023) or the teacher assessments that form part of Key Stage tests in English schools (GOV.UK, 2023). In all these settings, it is important to understand any unfairness arising from the use of teacher judgements, particularly that which is related to student characteristics, to ensure that educational systems are not exacerbating social inequalities. Evaluation of the 2020 CAGs offers a beneficial situation in which to examine these issues, as we can explore them for an entire cohort of students at the crucial end of secondary schooling stage.

Conceptual background for teacher biases

A variety of different mechanisms could produce teacher biases. Explicit, conscious prejudice or discrimination against groups of students should hopefully be minimised through the protections arising from policies such as the Equality Act (2010) and we would not expect this to be a dominant source of bias. Stereotyping – the characterising of a group based on particular attributes that shapes how people interact with that group – is a commonly featured theoretical basis for biases (Magowan, 2023: Urhahne and Wijnia, 2021). Stereotyping is typically considered to be an unconscious process allowing efficient, if not necessarily accurate, judgements of student groups which may operate in positive or negative directions (Campbell, 2015; Ready and Wright, 2011). Stereotyping is also thought to feed through to teacher expectations and beliefs for pupils and thus possibly effect their assessments of students. For instance, statistical discrimination along gender lines or by socioeconomic status could emerge through beliefs around girls outperforming boys or the expected performance of low achievers (Gibbons and Chevalier, 2018; Lindahl, 2016). Teachers may also boost grades to attempt to compensate for perceived disadvantage, as a form of encouragement, or to give students a tactical advantage depending on the situation (Snell et al., 2008). Additionally, where bias is assessed in relationship to test scores, divergences could emerge because the test scores

themselves may not represent the 'true' ability of students (Marcenaro-Gutierrez and Vignoles, 2015). Likewise, it is possible that teacher assessments and tests are capturing different aspects of student performance. For example, examinations represent performance at one point in time and include aspects of examination technique, whilst teachers may be able to encompass a more holistic appraisal of a student. Often it is not possible to differentiate or confirm the action of these processes, rather researchers rely on the identification of systematic divergences along student characteristics as indicators of potential biases.

Previous research

Previous research has indicated there may be systematic differences between teacher assessments and student test scores according to student sociodemographic characteristics, though the evidence for the strength and direction of these relationships is varied (Lee and Walter, 2020; Lee and Newton, 2021; Urhahne and Wijnia, 2021). For example, teachers are commonly found to overestimate girls' scores in comparison to boys (Timmermans et al., 2015; Plewis et al., 1997; Ready and Wright, 2011; Marcenaro-Gutierrez and Vignoles, 2015). A more mixed picture emerges in relation to ethnicity: some indicate potential biases against minority ethnic students (Ready and Wright, 2011; Tenenbaum and Ruck, 2007; Plewis et al., 1997), others find gaps in favour of minority students (Burgess and Greaves, 2013; Gibbons and Chevalier, 2008), or no significant differences (Lindahl, 2016; Marcenaro-Gutierrez and Vignoles, 2015). Students with Special Educational Needs (SEN) are generally marked lower by teachers than on tests, leading to greater differences between the two assessment methods for these students (Campbell, 2015; Burgess and Greaves, 2013; Reeves et al., 2001; Thomas et al., 1998). However, the overall conclusion by Thomas et al. (1998) was that any differences across methods of assessment were modest at best.

Bias is identified against students of lower socioeconomic status across various aspects of teacher judgements (Timmermans et al., 2015; Boone and Van Houtte, 2013; Burgess and Greaves, 2013; Ready and Wright, 2011). However, in relation to the difference between predicted grades (as provided to universities for applications) and achieved grades in examinations, studies have identified that disadvantaged students are generally overpredicted (Wyness, 2016). Given the 2020 CAGs were

meant to be the teachers' best prediction of how well a student would have done had they sat their examinations, rather than a contemporaneous estimate of student achievement, we may be more likely to see a positive bias towards disadvantaged students in the following analysis. Interviews with teachers around their CAG grading decisions suggest teachers were likely to be optimistic, representing student potential on a 'good' day (Holmes et al., 2021).

Additionally, it is important for research to consider how potential biases relate to prior attainment. Gibbons and Chevalier (2008) find the strongest differences between teacher and test assessments are observed by prior attainment rather than sociodemographic characteristics, with low attaining students tending to receive more favourable ratings from teachers relative to their test scores. Regarding predicted examination grades, research highlights the potential ceiling and floor effects that may be present at the extreme ends of the achievement spectrum (Dhillon, 2005). Possible interaction effects whereby high achieving disadvantaged students receive lower predicted A-level grades than their more socially advantaged counterparts have also been noted (Wyness, 2016; Murphy and Wyness, 2020).

Interaction effects between student characteristics are less often studied in this research area (Urhahne and Wijnia, 2021). Ready and Wright (2011) provide some evidence that the association of teacher perceptions and a student's language status differ by a student's ethnic background, whilst others have explored interactions between student gender and behaviour in how this impacts teacher ratings (Urhahne and Wijnia, 2021). Previous research has recognised the importance of considering the wider institutional context of student experience, whether this be as moderators of teacher bias or as an important factor in itself. For instance, Martínez et al. (2009) explore how classroom assessment practices could moderate the relationship between standardised tests and teacher scores. Timmermans et al. (2015) did not find teacher expectations interacted with the characteristics of the classroom population, however, they did find expectations were generally higher for classrooms characterised by high performing and advantaged children. This latter finding is echoed by Ready and Wright (2011) who find that student socioeconomic status is more strongly related to teacher inaccuracy in classrooms with more disadvantaged students. The wider school environment could also play a role in

differences between teacher estimates/predictions and achieved test scores with studies indicating that grading deviations may vary by school type (Dhillon, 2005; Marcenaro-Gutierrez and Vignoles, 2015). Explicit differences between individual schools are less often given attention, however, given the role schools played in the generation of CAGs in 2020 it is important to investigate potential school differences in differential grade inflation in this study.

Previous analyses of CAGs

Ofqual has investigated equality concerns for the 2020 grades (Lee et al., 2020; Stratton et al., 2021). Lee et al. (2020) compare relationships between student sociodemographic characteristics and the 2020 grades (calculated, CAGs, and final grades) with these same relationships for examination results from 2018 and 2019. They conclude there is little evidence that the 2020 grades disadvantaged student groups based on their characteristics, with the strongest overall difference between years being the large increase in mean grades. Stratton et al. (2021) focus more specifically on the CAGs and include student, subject, and school features when considering differing relationships between years. They evidence a ceiling effect whereby the highest prior attainers saw smaller GCSE grade increases over previous years. Additionally, they suggest that CAGs may have closed the gap somewhat between candidates from more versus less deprived areas, and that grades for independent schools and small cohorts may have shown more grade inflation in 2020. Stratton et al. (2021) also explore a limited number of interaction effects between student characteristics; however, these are not the central focus of the analysis, are restricted to two-way interactions, and show few notable changes in 2020 compared to the previous years. Indeed, their overall conclusion was that whilst the CAGs were higher on average, most relationships with student, subject, and school characteristics had not substantially changed in 2020.

These Ofqual reports utilise a new linked administrative dataset called GRADE (GRading and Admissions Data for England) (Office for National Statistics, 2021), which is the only available resource to assess the 2020 grading situation in England. A fundamental issue with using GRADE to explore potential teacher bias is the identification of a suitable benchmark on which to base comparisons to isolate unusual change outside of normal year-on-year variation. Lee et al. (2020) and

Stratton et al (2021) take the approach of using the last two 'normal' years (the 2017/18 and 2018/19 academic years) as the schema of grading (using the reformed 9-1 system) is the same. However, two years is not enough to establish 'typical' variation. Therefore, researchers must rely on subjective judgements of what should be considered a substantively important change. Moreover, even where notable differences are identified, this does not constitute proof of bias. Real substantive changes in grading gaps year-on-year can be confounded with any systematic changes in bias. Additionally, it is important to recall that teacher judgements are not the only form of assessment susceptible to bias (Lee and Newton, 2021). This is not to say that we should not evaluate the situation in 2020, but rather that any analysis of these data must acknowledge upfront the limitations of the data resource and to appreciate the necessarily cautious nature of any claims of change in 2020. Furthermore, given the subjective nature of judgements, it is doubly important for researchers to provide independent verification of the claims of no differential grade inflation drawn by Ofqual.

Analyses of the 2020 grades arising outside of official Ofqual reports are still rare, one exception is a recent study by Magowan (2023) who examined the differences between the 2020 CAGs and predicted grades derived from fitting prediction models to the 2018 and 2019 data. They found that whilst relative bias across different student characteristics was small, when three-way interactions were considered, the differences became more substantial. For example, the largest intra-group range Magowan (2023) identified in terms of main effects was just over a grade in total grade score (1.097, for their deprivation measure), whereas the largest intra-group range for studied three-way interactions was almost 4 grades (3.771) for ethnicity × deprivation × prior attainment combinations. Magowan (2023) therefore highlights the importance of considering interaction effects, and we expand on these multidimensional perspectives using a novel quantitative approach to studying intersections of student characteristics.

Intersectionality

The evaluation of potential biases in teacher assessments according to interactions of student characteristics warrants further examination. Neglecting interactions can mean important differences across student groups are missed, whilst investigating interactions of student characteristics can

provide a more nuanced portrayal of inequality. The study of interactions links to the conceptual background of intersectionality, (Crenshaw, 1989), which focuses on how social systems of oppression (e.g., racism, classism, sexism) are inherently interrelated. The characteristics of individuals position them at the intersections of these mutually constituted social systems, giving rise to heterogeneous experiences of (dis)advantage which are beyond what may be discerned from a purely unidimensional understanding of identity. Thus, stereotypes and expectations for students may develop differently for different intersections of student characteristics. For example, teachers may perceive the behaviour of Black boys more negatively than other student groups, potentially influencing judgements of ability (Wint et al., 2022). Interlocking systems of (dis)advantage may also influence student performance, for instance, disadvantaged White students are considered a group 'forgotten' by the UK educational system (House of Commons Education Committee, 2021). Therefore, in evaluating potential bias in the 2020 grades, we can draw upon intersectional perspectives to provide a richer understanding of the dimensions of inequality and to reveal potentially hidden marginalisation in grading practices.

This study

In this study, we will explore intersectional and school-level variation in the 2020 final grades for GCSE students in England, comparing this to variation found for examination grades in 2018 and 2019. By evaluating intersectional interaction effects and in explicitly considering individual school effects for possible changes in 2020, we address key gaps in current studies of the 2020 CAGs and studies of teacher biases more widely. This study also serves as an important independent investigation to verify the claims made by Ofqual of there being little to no differential grade inflation resulting from the switch to CAGs in 2020.

Data

This study uses data from the new GRADE linked administrative dataset (Office for National Statistics, 2021). This dataset is a joint venture, combining information from Ofqual, the DfE National Pupil Database (NPD), and the Universities and Colleges Admissions Service (UCAS). We utilise

data on student GCSE grades from 2018 and 2019 (the last 'normal' years pre-pandemic) and final grades from 2020. We will refer to these 2020 final grades as CAGs as these were used in most cases (only 4.8% of individual grades in our sample represent calculated grades). In the interests of space, we do not also evaluate differential grade inflation at A-level, however, we encourage similar studies of these higher-level qualifications.

We use data on student sociodemographic characteristics from the NPD. Our sample is comprised of 'typical' GCSE students: those in state schools receiving their grades age 16, who took five or more GCSEs including English and mathematics, and for whom we had complete sociodemographic information. We exclude students from independent schools as these are largely not captured in the NPD and would lead to unacceptably high rates of missingness. In 2018 the sample consists of 398,181 students within 3,328 schools, for 2019 the sample is 435,599 students in 3,406 schools, and in 2020 it is 425,031 students in 3,437 schools.

For each year we calculate an average GCSE score to serve as our outcome, pooling across all subjects taken. This accounts for the differing number of GCSEs that students' take, however, we do not consider subject specific biases. We focus on the following student sociodemographic characteristics: sex (Male, Female), ethnicity (White, Black, Asian, Chinese, Mixed, Other), and Income Deprivation Affecting Children Index (IDACI) score split into tertiles (High, Medium, Low). Additionally, we use a combined English and mathematics score from Key Stage 2 (KS2, age 11) split into deciles to capture prior attainment differences. We combine information on student sex, ethnicity, deprivation, and prior attainment to create intersectional strata representing the combination of these social identities, giving 360 intersections in total ($10 \times 2 \times 6 \times 3$). We chose these characteristics to balance capturing the most salient proxies for prominent systems of marginalisation, and what has previously been identified as of potential importance for biases, with manageable interpretation. Preliminary descriptive analyses summarising student's average grades across these student characteristics are detailed in the supplementary material (Supplementary Figure S1). Additional student characteristics (Special Educational Needs (SEN), Free School Meal status (FSM), and speaking English as an Additional Language (EAL)) were also explored in these preliminary analyses

to help identify which characteristics might exhibit differential grade inflation and were ruled out on this basis (Supplementary Figure S2).

We also include school type and school size as these characteristics were highlighted in media and other reports of differential grade inflation (Guardian, 2020b; Stratton et al., 2021). School type is split into academies (schools funded directly by the Government with more control over how they run), comprehensives (schools run by the Local Authority), selective (grammar schools that actively select students based on high achievement), Sixth Form Colleges, and Other (covering all remaining school types, such as Further Education and Tertiary establishments). School size is the number of students attending the school grouped as: 50 or fewer students; between 50 and 100 students; between 100 and 200; and over 200 students. Descriptive summaries of student grades by school type and size are available in the supplementary material (Supplementary Figure S1).

Methods

We draw upon a novel technique utilising multilevel linear regression to quantitively assess intersectional variation in student average grades: multilevel analysis of individual heterogeneity and discriminatory accuracy (intersectional MAIHDA) (Evans et al., 2018; 2020; Merlo, 2018). This analytical approach involves treating intersectional social identities as contexts in which individuals are situated. This is based on the conceptual understanding that individuals occupying particular intersecting positionalities (i.e., Black, disadvantaged girls or White advantaged boys) may share social experiences and resources, including in educational settings. Thus, we treat our students as nested within intersectional strata defined by combinations of student characteristics. Under the MAIHDA approach we can parsimoniously evaluate many intersections of multiple social dimensions simultaneously and directly quantify the power of intersectional strata to classify individuals according to their outcomes (Evans et al., 2018; 2020). As we are also interested in school-level variation and the assessment of differences in average grades across individual school effects we extend the MAIHDA two-level model by cross-classifying students (level-1) as simultaneously but separately nested within their intersectional identities (level-2) and their schools (also conceptually at level-2) (Leckie, 2013a).

The analytical strategy involves fitting a set of three multilevel models, repeated separately for each of the three years (2018, 2019, and 2020). Model 1 is an unadjusted two-way cross-classified model with no covariates. From this first model we can assess to what degree intersectional strata and schools explain overall variation in student average grades. In the case of the intersectional strata, it is important to note that baseline variation at this level represents the action of both the main effects of the sociodemographic components defining the strata as well as their interactions. In Model 2 we control for the sociodemographic components of the intersectional strata through adding them as variables into the fixed portion of the model. In this way we account for the main effects of these characteristics, with any remaining variation at the intersectional stratum level representing the action of two-and higher-way interactions between components. At the school level, by controlling for sociodemographic main effects we account for differences in school mean grades which are predicted by school variation in these factors. Thus, the model moves closer to isolating the effect of school practices and policies on student average grades. In our final model (Model 3) we further adjust the models by entering the school characteristics as main effects covariates. We fit all models by maximum likelihood estimation using the mixed command in Stata (Leckie, 2013b).

Results

Intersectional stratum and school variation

Table 1 provides the intercepts and variance components from our separate models predicting average GCSE grades in 2018, 2019 and 2020. Firstly, from the intercepts of the unadjusted model (Model 1), the overall grade inflation arising from the use of CAGs in 2020 over the 'normal' examination years is evident: the average grade in 2018 is 5.13, 5.07 in 2019, and 5.59 in 2020, representing an increase of approximately half a grade on average. The increase in 2020 over 2019 is clearly substantively important. For example, a half grade difference on average would be the equivalent of a student achieving one grade higher in four out of eight subjects. This overall grade inflation is well known and was a major impetus for the initial use of calculated grades in place of CAGs.

The unadjusted model also shows that student-level variation is smaller in 2020 (1.14) than the previous years (1.22 in 2018, 1.23 in 2019). The CAGs also show smaller school-level variation (0.13 in 2020 versus 0.19 in 2018 and 0.18 in 2018). In contrast, stratum-level variation is highest in 2020 (1.65), however this is closer to the 2019 figure (1.60) than the variation in 2019 is to that in 2018 (1.39). Therefore, we cannot conclude with confidence that this is evidence of a greater reliance on student intersectional sociodemographic characteristics for the 2020 CAGs.

In Model 1, there is a similar amount of stratum-level variation (as a percentage of total variation) in all three years, with slightly less in 2018 (49.7%) and slightly more in 2019 and 2020 (53.1% and 56.4% respectively). Therefore, a strong degree of variability in average grades is associated with the intersections of sociodemographic characteristics to which a student belongs, likely primarily driven by student prior attainment. In the unadjusted model this variation represents the action of both main and interactional effects. However, it is important to recall that approximately 40% of overall variation remains at the student level (within stratum and within school) representing the action of other student characteristics not used in defining our stratum. In contrast, there is a much smaller degree of variation positioned between schools, the highest percentage being 6.8% in 2018, followed by 6.0% in 2019, and just 4.6% in 2020. This suggests that schools are of lesser importance to students' average grades in 2020 than in the previous years (as much as we can establish with only three years to evaluate).

In Model 2 as we control for the main effects of the constituent components of our intersectional strata, any remaining stratum-level variation corresponds to the action of any two- and higher-way interaction effects. Stratum-level variance drops dramatically in Model 2 for all three years, with the sociodemographic main effects explaining the same proportion of stratum-level variation (99.6%) between the unadjusted and adjusted models. This leaves a similar small proportion of remaining stratum-level variation in all three years (0.4% in 2018 and 2019 and 0.5% in 2020). Therefore, the strength of the intersectional strata lies in the main effects of the constituent characteristics with very little attributable to interaction effects above and beyond these main effects. It is notable this patterning of variation is remarkably stable across all three years studied, suggesting

that the intersectional characteristics analysed have similar power over grades despite the unusual situation in 2020.

Additionally, by controlling for student characteristics in Model 2 we are accounting for any element of school-level variation which is comprised by cohort differences between schools in these characteristics, with remaining variation now more likely to reflect the action of school practices and context on student learning and assigning of CAGs. The controls in Model 1 are much less powerful at explaining school-level variation than stratum-level (the percentage explained between Model 1 and Model 2 is 0.2% in 2018 and 2020 and 0.1% in 2019). That the student sociodemographic characteristics have similar explanatory power in 2020 as they do in previous years could be seen as a positive result regarding equalities considerations: were the explanatory power of these sociodemographic variables to be much higher at the school level in 2020 versus previous years this might have suggested that teachers relied heavily on these characteristics in constructing the student CAGs. As a result of the dramatic drop in the stratum-level variance, the percentage of remaining CAG variation seen at the school level is higher in Model 2 than in Model 1 (13.4% in 2018, 12.8% in 2019, and 10.4% in 2020. We also continue to see smaller school-level variation in 2020 than the previous normal examination years.

In Model 3 we are particularly interested in seeing how our school-level factors (type and size) impact on the remaining school-level variation. Our school characteristics explain more school-level variance in 2018 (12.8%) and 2019 (11.9%) than in 2020 (5.5%). That these school-level factors are not as powerful in 2020, and by a notable margin (the percentage change is half that seen in the previous years) suggests the process of producing CAGs may have worked to reduce mean school differences by these school characteristics.

Coefficient comparison

Figure 1 (top) plots the regression coefficients from Model 2 along with their 95% confidence intervals (full results given in Supplementary Table S1). We will briefly note any divergent main effects relationships in 2020, following Stratton et al. (2021) in deeming differences greater than 0.10

of a grade, and where the difference between 2020 and 2019 is greater than that between 2018 and 2019, as notable. One standard deviation in the overall mean grade scores across all three years is 1.7 grade points, therefore, our threshold criteria is highly cautious (less than $1/10^{th}$ of a standard deviation), allowing us to identify even marginal effects.

For the student characteristics, none of the 2020 coefficients meet both elements of our criteria for notable divergences. The 2020 coefficient for the highest KS2 decile is 0.10 grade points lower than in 2019, which could indicate the action of a ceiling effect (you cannot predict above the top grades, limiting possible grade inflation). However, this is not out of bounds of 'normal' variability given the difference to the coefficient in 2018 is greater (0.30). Notably, excepting decile 10, every 2020 regression coefficient is greater in absolute magnitude than in 2018 or 2019, so KS2 appears to be a stronger predictor in 2020 than in the previous years. In contrast, the coefficients for ethnicity and deprivation are smaller in magnitude in 2020.

Figure 1 (bottom) provides the regression coefficients related to the school characteristics from Model 3 (see Supplementary Table S2 for full results). Selective schools show smaller grade inflation in 2020 (coefficient is 0.37 in 2020 compared to 0.48 in 2019), likely reflecting the impact of the ceiling effect. There is a very large difference (1.01 grade points) between the 2020 and 2019 coefficients for Sixth form colleges. However, the coefficient is not significant, likely due to the very small number of sixth forms in our GCSE sample (these institutions typically cater for higher education levels), so we will not overemphasise this result. Additionally, the 2020 coefficient for other school types (-0.51) is substantially smaller than in 2019 (-0.62): a 0.11 difference compared with a 0.09 difference between 2018 and 2019. However, this seems to be a continuation of a trend rather that a particular divergence for 2020. Finally, the smallest schools (<50 students) show a less negative relationship with average grade in 2020 (-0.09) than in 2019 (-0.28) or 2019 (-0.26). This difference is just over a 10^{th} of a standard deviation in the overall grade distribution at 0.18 points.

Our coefficient comparisons reveal that the predictive power of student characteristics appears to have shifted towards prior attainment in 2020. However, the model results generally reveal a pattern of relative stability through time, particularly regarding intersectional variation. A closer

look at the stratum effects is warranted to evaluate whether the apparent consistency holds for the individual intersectional strata, particularly those in the adjusted models representing the action of two-way or higher interactions.

Stratum effects

Figure 2 provides scatterplots comparing the estimated stratum effects for the unadjusted models (top row) and the model adjusted for prior attainment and sociodemographic characteristics (bottom row) across all year combinations (Supplementary Figure S3 provides the plot additionally controlling for school characteristics). In Model 1 the correlations between years are all extremely high (r = 0.99) evidencing a strong degree of consistency from 2018 through to 2020, driven by the dominant main effects. Therefore, the rank ordering of the strata is very stable despite the overall grade inflation in 2020. When we consider Model 2, where the stratum effects represent the action of any two- or higher-way interactions between the student characteristics, the correlations are lessened (r = 0.66 comparing 2018 and 2019 and r = 0.68 comparing 2019 and 2020). There is generally less stability year-on-year in interactional effects. However, we see neither a substantially lower correlation between 2020 and 2019 than between 2019 and 2018 (suggesting that sociodemographic interactions play out differently in 2020 than in previous years) nor a dramatically higher correlation (suggesting that there could have been excessive use of the 2019 results in assigning CAGs).

The correlations between schools in Model 1 (Supplementary Figure S4) are still high but lower than for the intersectional strata (r = 0.80 comparing 2018 and 2019, r = 0.82 comparing 2019 and 2020). There is generally less stability year-on-year in school effects on average grades, which aligns with the literature on the instability of school performance over time (Prior et al., 2021). In Model 3 (the model more pertinent to the examination of school effects), the correlation remains the same for the 2019 to 2020 comparison (r = 0.82). This corroborates our earlier findings suggesting our school characteristics have lower explanatory power in 2020 than in the previous years and suggests there could have been excessive reliance on the 2019 results in generating CAGs.

Intersectional interactions

To further evaluate the intersectional interactional effects, we plot the predicted stratum effects from Model 2 by the sociodemographic components of the strata (Figure 3). These stratum effects represent to what degree the average grades of students composing an intersectional group differ on average from that predicted by their combination of main effects. Positive effects show stratum groups who tend to score more highly on average than their main effects predict, negative effects those who score lower. For the most part we see relative consistency in how these interactions play out across the three years, or where there is variation, this is present between all three years rather than the 2020 year appearing particularly divergent. Additionally, the general scale of these interactional stratum effects is small.

If we apply the same criteria as before regarding notable effects (a shift of 0.1 of a grade or more from the previous year, where the 2020 to 2019 change is greater in scale than the 2019 to 2018 change) we find only 11 (3.1%) notable effects from the 360 intersectional strata (see Supplementary Table S3). These strata represent a mix of characteristics, though all the identified groups feature ethnic minority students with only the Chinese and White groups not appearing. The most extreme differences seen between 2020 and 2019 are still not very substantial, being over one tenth of the standard deviation in overall grades across all three years (0.17) at 0.18 (Other ethnicity, female, high deprivation, KS2 decile 6) and -0.20 (Black, female, mid deprivation, KS2 decile 4). However, we will not place too much weight on these 'notable' stratum changes as similarly sized differences are also observed between 2018 and 2019, where 3.9% (14 strata) meet the 0.1 grade difference threshold (Supplementary Table S4).

In contrast to the stratum effects, if we consider the differences between years in the predicted school effects from our fully adjusted model (Model 3) there are considerably larger differences present (Supplementary Figure S5). Whilst most school differences are close to zero (representing relative stability between years), some schools shift by half a grade or more, with a small number reaching a one plus grade shift in average GCSE score. The overall variation is smaller for changes between 2020 and 2019 (standard deviation 0.22 for 2020 to 2019, compared with 0.26 for 2019 to 2018), with a greater frequency of schools appearing to make little change between the two years.

This aligns with our previous findings regarding the smaller variability between schools in 2020 and that there was likely considerable reliance on the school 2019 results in producing the CAGs.

Discussion

In this study we contribute to the growing body of work analysing the student CAGs awarded in England in lieu of examination grades when GCSE examinations were cancelled in response to COVID-19 in 2020. Drawing upon the notion of intersectionality, we focus on addressing the need for further examination of interaction effects in assessments of whether the switch to CAGs brought about any new or divergent relationships with student background characteristics. Overall, the picture that emerges is one of stability in intersectional relationships over time. Our intersectional MAIHDA analysis reveals that the combination of social and demographic identities to which a student belongs has the same explanatory power in all three years considered. Additionally, correlations between intersectional effects are similar through time, suggesting the use of CAGs did not dramatically alter the rank ordering of students. Moreover, only 3.1% of interactional stratum effects in 2020 show notable (>0.10 grade points) differences from 2019, with a similar percentage of such differences also found between 2018 and 2019. It appears that the move to CAGs in 2020 largely did not introduce any substantial divergences from previous years in relationships with student characteristics, even when we consider intersectional interaction effects. Therefore, this independent investigation draws similar conclusions to the Ofqual equalities analyses (Lee et al., 2020; Stratton et al., 2021).

As with the previous analyses (Stratton et al., 2021), our results are indicative of a ceiling effect resultant from the overall grade inflation and we highlight a shift in predictive power towards KS2 score in 2020. This could suggest heavy reliance on prior attainment by teachers and schools when assigning CAGs, rather than on individual achievements or performance in the intervening years. This may have reduced the accuracy of CAGs as a reflection of a student's current ability. The higher predictive power of prior attainment may also reflect the fact that CAGs factored out some of the typical unpredictable sources of variability in student outcomes, such as last minute 'cramming' for exams, shock events, or exam question variation. Teachers were giving students the 'benefit of the doubt' (Holmes et al., 2021). There may also have been a conscious or unconscious effort by teachers

not to base judgements on student demographics to avoid bias, lessening their predictive power in favour of prior attainment.

The most notable differences arising in 2020 were between schools. Selective schools show less grade inflation. As selective school students are expected to be multiply advantaged, in ways beyond that captured by our student characteristics, this result likely reflects the impact of the ceiling effect on highest grades. Additionally, smaller schools appear to benefit more from the move to CAGs than schools with larger cohorts. Smaller schools, and therefore smaller classes, could represent a different mix of subjects than at larger institutions, and these small-class subjects may be marked particularly optimistically under the 2020 CAGs. Elsewise, teachers in smaller schools may feel indirect pressure towards greater optimism due to potentially closer relationships with their students, and subsequently feeling more invested in student outcomes. There may also be less resources for CAG moderation processes at smaller schools, and Holmes et al. (2021) indicate smaller centres tended to take less data-driven approaches to CAG generation, relying more on subjective evidence.

Again, these results are alike to those identified by Stratton et al. (2021). Though these shifts were generally small in size, they highlight potential sources of unfairness in the awarding of the 2020 CAGs and avenues for further research. Whilst we controlled for important student background characteristics, and thus many cohort differences between schools, there are likely to be other student factors (and thus potential sources of bias) associated with attendance at selective or small size schools that we do not capture. Additionally, the distribution of different school types is not evenly distributed around England, meaning the exploration of geographic dimensions to potential inequalities arising from the use of CAGs in 2020 will be an important opportunity for future work.

However, we demonstrate that school-level variation is lower in 2020 than the previous 'normal' examination years and that our studied school characteristics had substantially less explanatory power in 2020. Where between-school differences are smaller this could suggest schools had similar approaches to producing CAGs although it could also reflect the naturally smaller variation arising from the overall grade inflation and the associated ceiling effect. Moreover, our results highlighted that there may have been heavy reliance on the 2019 results when determining

2020 grades at the school level. Teachers, particularly those at larger centres using more data driven approaches (Holmes et al., 2021), may have relied on knowledge built up from past cohorts with the most recent (2019) being the most prominent.

As highlighted in the introduction, this study is necessarily limited by the data available from GRADE (Office for National Statistics, 2021). Most notably, our ability to determine what constituted 'normal' year-on-year variation was limited to just two years where we could maintain comparability in grading scales. We chose a relatively modest criteria for identifying 'notable' shifts and changes at 0.1 of a grade (less than one tenth of a standard deviation in overall average grades). That very few variables or intersectional stratum effects met this criterion means our overall conclusion remains that the switch to CAGs in 2020 largely did not noticeably widen or narrow pre-existing inequalities according to student, school or intersectional characteristics.

References

BBC News (2020. A-levels: Anger over 'unfair' results this year. Published 13/08/2020. URL: https://www.bbc.co.uk/news/education-53759832

Boliver, V. (2011). Expansion, differentiation, and the persistence of social class inequalities in British higher education. *Higher Education*, 61: 229-242.

Boone, S. and Van Houtte, M. (2013). Why are teacher recommendations at the transition from primary to secondary education socially biased? A mixed-method research. *British Journal of Sociology of Education*, 34(1): 20-38.

Burgess, S. and Greaves, E. (2013). Test scores, subjective assessment, and stereotyping of ethnic minorities. *Journal of Labor Economics*, 31(3): 535-576.

Campbell, T. (2015). Stereotyped at seven? Biases in teacher judgements of pupils' ability and attainment. *Journal of Social Policy*, 44(3): 517-543.

Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: a Black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum*, 1(8): 139-167.

Dhillon, D. (2005). Teachers' estimates of candidates' grades: Curriculum 2000 Advanced Level Qualifications. *British Educational Research Journal*, 31(1): 69-88.

Evans, C.R., Leckie, G., and Merlo, J. (2020). Multilevel versus single-level regression for the analysis of multilevel information: the case of quantitative intersectional analysis. *Social Science and Medicine*¸ 245: 112499.

Evans, C.R., Williams, D.R., Onnela, J-P., and Subramanian S.V. (2018). A multilevel approach to modelling health inequalities at the intersection of multiple social identities. *Social Science & Medicine*¸ 203: 64-73.

Gibbons, S. and Chevalier, A. (2008). Assessment and age 16+ education participation. *Research Papers in Education.* 23(2): 113-123.

Green, M.A., Evans, C.R., and Subramanian, S.V. (2017). Can intersectionality theory enrich population health research? *Social Science and Medicine*, 178: 214-216.

GOV.UK (2020). Further details on exams and grades announced. URL: https://www.gov.uk/government/news/further-details-on-exams-and-grades-announced

GOV.UK (2023). The national curriculum. URL: https://www.gov.uk/national-curriculum

Guardian, The (2020a). A-level results: almost 40% of teacher assessments in England downgraded. Published 13/08/2020. URL: https://www.theguardian.com/education/2020/aug/13/almost-40-of-english-students-have-a-level-results-downgraded

Guardian, The (2020b). England A-level downgrades hit pupils from disadvantaged areas hardest. Published 13/08/2020. URL: https://www.theguardian.com/education/2020/aug/13/england-a-level-downgrades-hit-pupils-from-disadvantaged-areas-hardest

Harlen, W. (2005). Trusting teachers' judgement: research evidence of the reliability and validity of teachers' assessment used for summative purposes. *Research Papers in Education*, 20(3): 245-270.

Holmes, S., Churchward, D., Howard, E., Keys, E., Leahy, F. and Tonin, D. (2021). *Centre Assessment Grades: Teaching Staff Interviews, Summer 2020*. Ofqual: London.

House of Commons Education Committee (2021). The forgotten: how White working-class pupils have been let down, and how to change it. First Report of Session 2021-22. London: House of Commons. Available at: https://committees.parliament.uk/publications/6364/documents/70802/default/ [Accessed 17/07/2023]

Kelly, A. (2021). A tale of two algorithms: the appeal and repeal of calculated grades systems in England and Ireland in 2020. *British Educational Research Journal*,

Leckie, G. (2013a) Cross-classified multilevel models – concepts. *LEMMA VLE Module 12*, 1-60. Available at: http://www.bristol.ac.uk/cmm/learning/online-course/

Leckie, G. (2013b). Cross-classified multilevel models – Stata practical. *Lemma VLE Module 12*, 1-52. Available at: http://www/bristol.ac.uk/cmm/learning/online-course/

Lee, M.W. and Newton, P. (2021). Systematic divergence between teacher and test-based assessment: literature review. Ofqual: London. URL: https://www.gov.uk/government/publications/systematic-divergence-between-teacher-and-test-based-assessment/systematic-divergence-between-teacher-and-test-based-assessment-literature-review [Accessed 22/05/2023]

Lee, M.W. and Walter, M. (2020). *Equality impact assessment: literature review.* Ofqual: London. URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/879605/Equality_impact_assessment_literature_review_15_April_2020.pdf [Accessed 27/06/2023]

Lee, M.W., Stringer, N. and Zanini, N. (2020). *Student-level equalities analyses for GCSE and A level.* Ofqual: London.

Lindhal, E. (2016). Are teacher assessments biased? Evidence from Sweden. *Education Economics*, 24(2): 224-238.

Magowan, L. (2023). Centre assessment grades in 2020: a natural experiment for investigating bias in teacher judgements. *Journal of Computational Social Science*, https://doi.org/10.1007/s42001-023-00206-x

Marcenaro-Gutierrez, O. and Vignoles, A. (2015). A comparison on teacher and test-based assessment for Spanish primary and secondary students. *Educational Research*, 57(1): 1-21.

Martínez, J.F., Stecher, B. and Borko, H. (2009). Classroom assessment practices, teacher judgements, and student achievement in mathematics: evidence from the ECLS. *Educational Assessment*, 14(2): 78-102.

Merlo, J. (2018). Multilevel analysis of individual heterogeneity and discriminatory accuracy (MAIHDA) within an intersectional framework. *Social Science & Medicine*, 203: 74-80.

Murphy, R. and Wyness, G. (2020). Minority report: the impact of predicted grades on university admissions of disadvantaged groups. *Education Economics*, 28(4): 333-350.

Office for National Statistics, (2021) ONS SRS Metadata Catalogue, GRading and Admissions Data England-Ofqual-DfE-UCAS, dataset, released 14 December 2021, https://doi.org/10.57906/k68n-bt74

Ofqual (2020a). How GCSEs, AS & A levels will be awarded in summer 2020. Published 03/04/2020. URL: https://www.gov.uk/government/news/how-gcses-as-a-levels-will-be-awarded-in-summer-2020 [Accessed 20/03/2023]

Ofqual (2020b). Awarding GCSE, AS, A level, advanced extension awards and extended project qualifications in summer 2020: interim report. Coventry: Ofqual. Available at: https://www.gov.uk/government/publications/awarding-gcse-as-a-levels-in-summer-2020-interim-report

Ofqual (2020c). GCSE and A level students to receive centre assessment grades. Published 17/08/2020. URL: https://www.gov.uk/government/news/gcse-and-a-level-students-to-receive-centre-assessment-grades

Plewis, I. (1997). Inferences about teacher expectations from national assessment at Key Stage One. *British Journal of Educational Psychology*, 67: 235-247.

Prior, L., Jerrim, J., Thomson, D. and Leckie, G. (2021). A review and evaluation of secondary school accountability in England: Statistical strengths, weaknesses and challenges for 'Progress 8' raised by COVID-19. *Review of Education*, 9(3): e3299.

Ready, D.D. and Wright, D.L. (2011). Accuracy and inaccuracy in teachers' perceptions of young children's cognitive abilities: the role of child background and classroom context. *American Educational Research Journal*, 48(2): 335-360.

Reeves, D.J., Boyle, W.F. and Christie, T. (2001). The relationship between teacher assessments and pupil attainments in standard test tasks at Key Stage 2, 1996-98. *British Educational Research Journal*, 27(2): 141-160.

Snell, M., Thorpe, A., Hoskins, S. and Chevalier, A. (2008). Teachers' perceptions and A-level performance: is there any evidence of systematic bias? *Oxford Review of Education*, 34(4): 403-423.

Stratton, T., Zanini, N. and Noden, P. (2021). *An evaluation of centre assessment grades from summer 2020.* Ofqual: London.

Tenenbaum, H.R. and Ruck, M.D. (2007). Are teachers' expectations different for racial minority than for European American students? A meta-analysis. *Journal of Educational Psychology*, 99(2): 253-273.

Timmermans, A.C., Kuyer, H. and van der Werf, G. (2015). Accurate, inaccurate, or biased teacher expectations: Do Dutch teachers differ in their expectations at the end of primary education? *British Journal of Educational Psychology*, 85, 459-478.

Thomas, S., Madaus, G.F., Raczek, A.E. and Smees, R. (1998). Comparing teacher assessment and standard task results in England: the relationship between pupil characteristics and attainment. *Assessment in Education: Principles, Policy & Practice*, 5(2): 213-246.

UCAS (2023). Predicted grades – what you need to know for entry this year. URL: https://www.ucas.com/advisers/managing-applications/predicted-grades-what-you-need-know-entry-year

Urhahne, D. and Wijnia, L. (2021). A review on the accuracy of teacher judgements. *Educational Research Review*, 32: 100374.

Wint, K.M., Opara, I., Gordon, R. and Brooms, D.R. (2022). Countering educational disparities among Black boys and Black adolescent boys from pre-K to high school: a life course-intersectional perspective. *The Urban Review*, 54: 183-206.

Wyness, G. (2016). Predicted grades: accuracy and impact: A report for University and College Union. University and College Union: London. Available at: https://www.ucu.org.uk/media/8409/Predicted-grades-accuracy-and-impact-Dec-16/pdf/Predicted_grades_report_Dec2016.pdf

Tables

Table 1. Model 1, 2 and 3 results for 2018, 2019, and 2020, including intercepts plus strata, school, and student variances, variance partitioning coefficients (VPCs), and proportional changes in variance (PCVs).

			2018		2019		2020	
			Est.	S.E	Est.	S.E	Est.	S.E
Model 1: Unadjusted	Intercept		5.13	0.063	5.07	0.068	5.59	0.068
	Strata variance		1.39	0.105	1.60	0.121	1.65	0.124
	School variance		0.19	0.005	0.18	0.005	0.13	0.004
	Student variance		1.22	0.003	1.23	0.003	1.14	0.002
	VPC	Strata	49.7%		53.1%		56.4%	
		School	6.8%		6.0%		4.6%	
	PCV	Strata	-		-		-	
		School	-		-		-	
Model 2: Student characteristics	Intercept		4.37	0.023	4.28	0.022	4.82	0.022
	Strata variance		0.01	0.001	0.01	0.001	0.01	0.001
	School variance		0.19	0.005	0.18	0.005	0.13	0.004
	Student variance		1.22	0.003	1.23	0.003	1.14	0.002
	VPC	Strata	0.4%		0.4%		0.5%	
		School	13.4%		12.8%		10.4%	
	PCV	Strata	99.6%		99.6%		99.6%	
		School	0.2%		0.1%		0.2%	
Model 3: School characteristics	Intercept		4.44	0.024	4.35	0.024	4.85	0.024
	Strata variance		0.01	0.001	0.01	0.001	0.01	0.001
	School variance		0.17	0.005	0.16	0.004	0.13	0.004
	Student variance		1.22	0.003	1.23	0.003	1.14	0.002
	VPC	Strata	0.4%		0.4%		0.5%	
		School	11.8%		11.4%		9.9%	
	PCV	Strata	0.1%		0.0%		0.0%	
		School	12.8%		11.9%		5.5%	

Figures

Figure 1. Regression coefficients and their 95% confidence intervals for the student sociodemographic characteristics from Model 2 (top row) and for the school characteristics from Model 3 (bottom row) for 2018, 2019, and 2020. * indicates reference category.

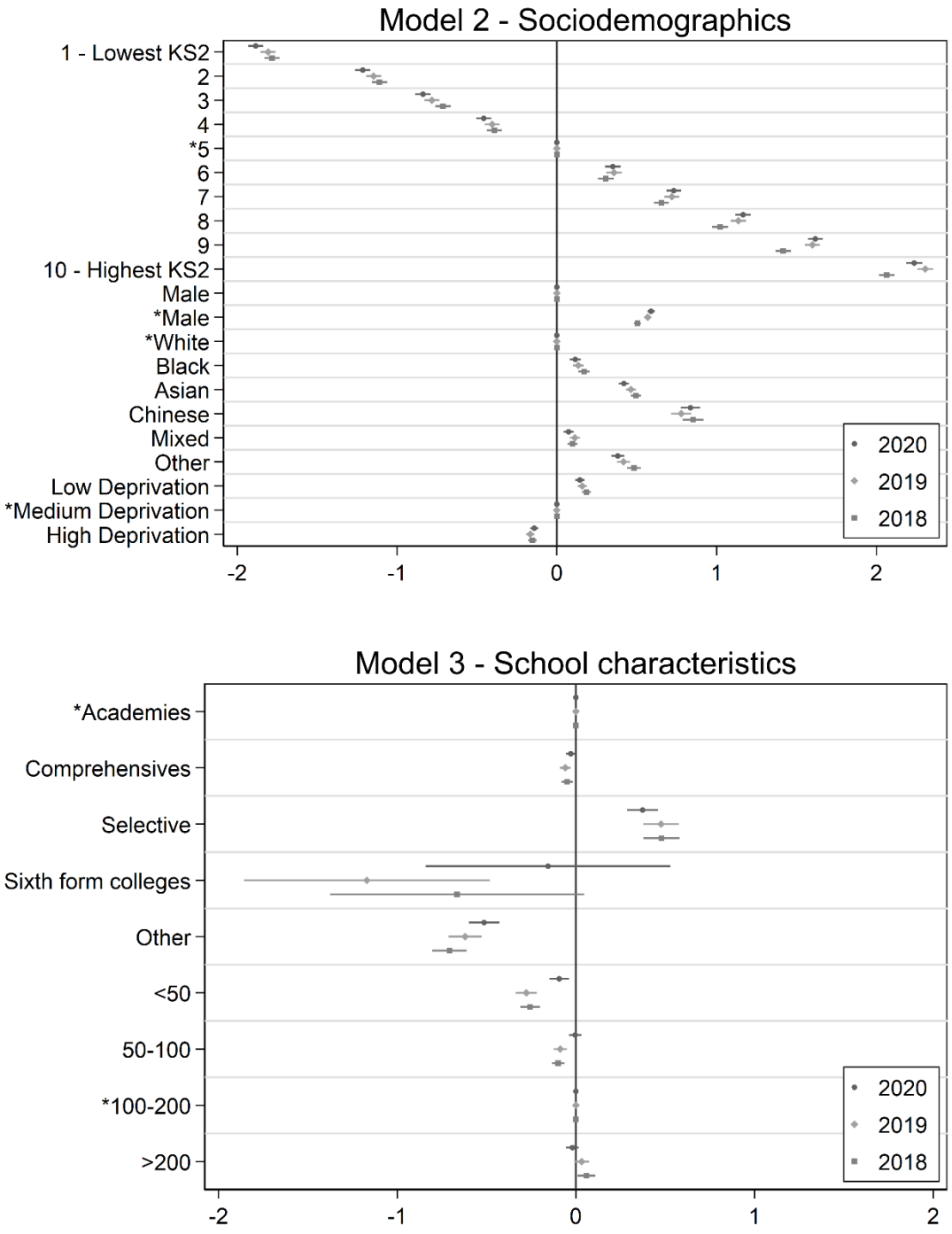


Figure 2. Scatterplots of stratum effects for the unadjusted model (Model 1, top row) and model adjusted for student sociodemographic characteristics (Model 2, bottom row) comparing 2018 and 2019 (left) and 2019 and 2020 (right).

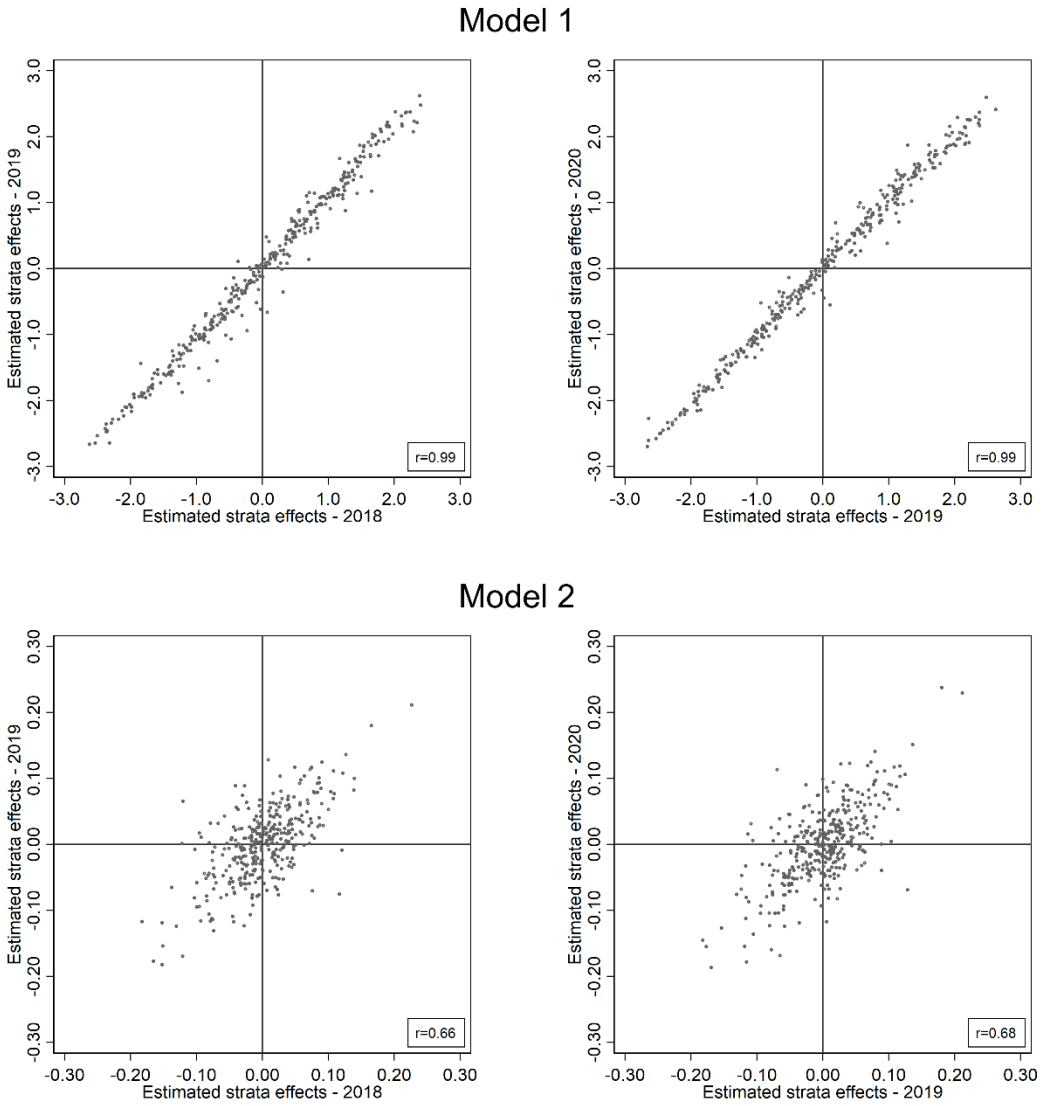
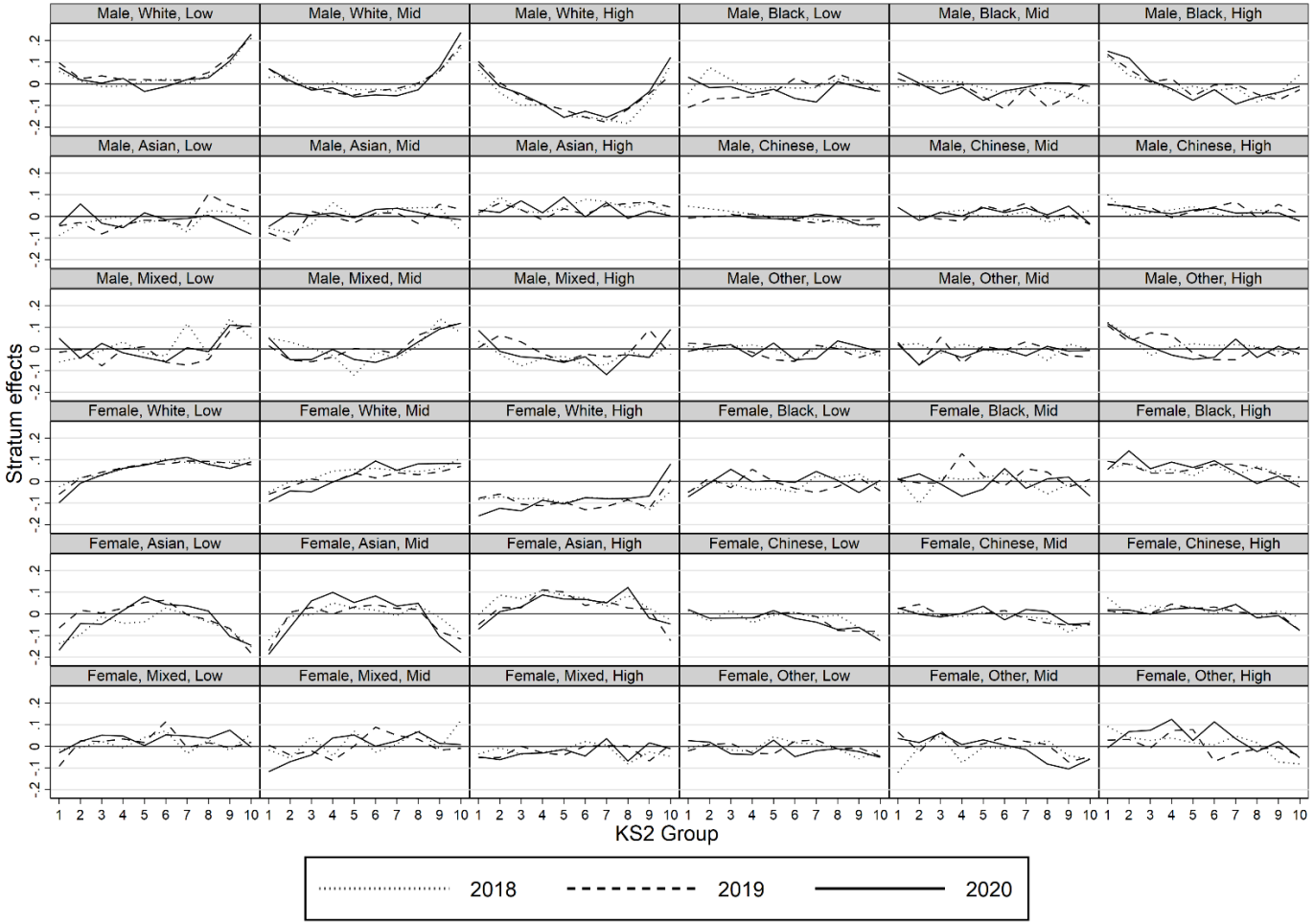


Figure 3. Model 2 predicted stratum effects for the model adjusted for student sociodemographic characteristics for 2018, 2019, and 2020, by KS2 group, gender, ethnicity, and IDACI tertile.



Supplementary material

Preliminary descriptive results

We summarise the average student grades across the sociodemographic components of the intersectional strata (KS2 group, gender, ethnicity, and IDACI tertile) and school characteristics (school type, school size) (Supplementary Figure S1). Equivalent plots for other sociodemographic characteristics not used in constructing the (FSM, EAL, SEN, IDACI quintile) are available in Supplementary Figure S2. Outside the overall grade inflation in 2020 versus previous years, the overall picture is one of stability in relationships between the characteristics and average grades.

However, some small shifts are identifiable. The highest (0.41 grade difference) and lowest (0.44 grade difference) prior achievement levels show smaller grade inflation in 2020 over 2019, compared with mid-range KS2 scores (e.g. 0.53 grade difference for KS2 group 5). This aligns with the notion of a ceiling effect (you cannot predict above the top grade, limited possible inflation), and with the idea that the most extreme grades are easier to predict accurately. Chinese students also see smaller grade inflation than other ethnic groups which could also relate to the ceiling effect for these generally high achieving students. There is some indication of differential grade inflation in favour of disadvantaged students (0.40 grade difference between 2020 and 2019 for the least deprived and 0.52 for the most deprived students), and therefore an apparent narrowing of the disadvantage gap, a finding also shown by Stratton et al. (2021). This result is not replicated for FSM and non-FSM students, where the grade difference is 0.47 for both student groups. IDACI score is tied to the home neighbourhood of the student, and therefore potentially with perceptions of place, whereas FSM status is largely determined by an income threshold. It may have been that the IDACI more closely aligned with teacher's perceptions of student disadvantage.

In terms of school characteristics, students taking their GCSEs in sixth form colleges see a much higher degree of grade inflation in 2020 over 2019 (1.23 grade difference) than other school types (e.g. 0.45 grade difference for academies, the most populous school type). However, we will not overstate this result as there are very few sixth form colleges in our sample as it is unusual for

students to sit these examinations at these types of centres. Additionally, the smallest schools show higher grade inflation in 2020 over 2019 (0.65) than the largest schools (0.40). This is similar to Ofqual findings regarding small cohort subjects showing higher inflation than subjects taken by many students (Stratton et al., 2021). Altogether, these descriptive summaries showcase that the overall picture is likely to be one of stability in relationships over time. However, there may still be some identifiable substantive changes and it is important to see if these relationships are maintained when we control for the other student and school characteristics simultaneously.

Supplementary tables

Table S1. Regression coefficients and standard errors from separate models predicting average student GCSE grades in 2018, 2019 and 2020, adjusted (Model 2) for student sociodemographic characteristics.

		2018		2019		2020	
		Beta	S.E	Beta	S.E	Beta	S.E
Fixed							
Intercept		4.37	0.023	4.28	0.022	4.82	0.022
KS2 Group	*ref. = 5*						
	1 - Lowest	-1.78	0.024	-1.81	0.024	-1.89	0.024
	2	-1.11	0.024	-1.15	0.024	-1.22	0.024
	3	-0.71	0.025	-0.78	0.024	-0.84	0.024
	4	-0.39	0.024	-0.40	0.024	-0.46	0.024
	6	0.31	0.025	0.36	0.025	0.35	0.025
	7	0.65	0.024	0.72	0.024	0.73	0.024
	8	1.02	0.026	1.14	0.025	1.17	0.025
	9	1.42	0.024	1.60	0.024	1.62	0.024
	10 - Highest	2.07	0.025	2.31	0.025	2.24	0.026
Gender	*ref. = Male*						
	Female	0.51	0.011	0.57	0.011	0.59	0.011
Ethnic group	*ref. = White*						
	Black	0.17	0.018	0.13	0.018	0.12	0.018
	Asian	0.50	0.016	0.46	0.016	0.42	0.016
	Chinese	0.85	0.034	0.78	0.032	0.84	0.031
	Mixed	0.10	0.016	0.11	0.016	0.07	0.016
	Other	0.48	0.022	0.42	0.021	0.38	0.021
IDACI tertile	*ref. = 2 - Mid*						
	1 - Low	0.19	0.015	0.16	0.014	0.14	0.014
	3 - High	-0.15	0.013	-0.17	0.013	-0.14	0.013
Random							
Strata variance		0.01	0.001	0.01	0.001	0.01	0.001
School variance		0.19	0.005	0.18	0.005	0.13	0.004
Student variance		1.22	0.003	1.23	0.003	1.14	0.002
VPC	Strata	0.4%		0.4%		0.5%	
	School	13.4%		12.8%		10.4%	
PCV	Strata	99.6%		99.6%		99.6%	
	School	0.2%		0.1%		0.2%	

Table S2. Regression coefficients and standard errors from separate models predicting average student GCSE grades in 2018, 2019 and 2020, adjusted (Model 3) for student sociodemographic characteristics and school characteristics.

		2018		2019		2020	
		Beta	S.E	Beta	S.E	Beta	S.E
Fixed							
Intercept		4.44	0.024	4.35	0.024	4.85	0.024
KS2 Group	*ref = 5*						
	1 - Lowest	-1.78	0.024	-1.81	0.024	-1.88	0.024
	2	-1.11	0.024	-1.15	0.024	-1.22	0.024
	3	-0.71	0.025	-0.78	0.024	-0.84	0.024
	4	-0.39	0.024	-0.40	0.024	-0.46	0.024
	6	0.31	0.025	0.36	0.025	0.35	0.025
	7	0.65	0.024	0.72	0.024	0.73	0.024
	8	1.02	0.026	1.14	0.025	1.16	0.025
	9	1.42	0.024	1.60	0.024	1.62	0.024
	10 - Highest	2.06	0.025	2.30	0.025	2.23	0.026
Gender	*ref = Male*						
	Female	0.50	0.011	0.57	0.011	0.59	0.011
Ethnic group	*ref = White*						
	Black	0.17	0.018	0.14	0.018	0.12	0.018
	Asian	0.49	0.016	0.46	0.016	0.42	0.016
	Chinese	0.85	0.034	0.78	0.032	0.84	0.031
	Mixed	0.10	0.016	0.11	0.016	0.07	0.016
	Other	0.49	0.022	0.42	0.021	0.38	0.021
IDACI tertile	*ref = 2 - Mid*						
	1 - Low	0.18	0.015	0.16	0.014	0.14	0.014
	3 - High	-0.15	0.013	-0.17	0.013	-0.14	0.013
School type	*ref = Academies*						
	Comprehensives	-0.05	0.016	-0.06	0.016	-0.03	0.014
	Selective	0.48	0.051	0.48	0.051	0.37	0.044
	Sixth form	-0.67	0.363	-1.17	0.351	-0.16	0.349
	Other types	-0.71	0.049	-0.62	0.047	-0.51	0.044
School size	*ref = 100-200*						
	<50	-0.26	0.028	-0.28	0.03	-0.09	0.028
	50-100	-0.10	0.018	-0.09	0.019	0.00	0.017
	>200	0.06	0.025	0.03	0.021	-0.02	0.018
Random							
Strata variance		0.01	0.001	0.01	0.001	0.01	0.001
School variance		0.17	0.005	0.16	0.004	0.13	0.004
Student variance		1.22	0.003	1.23	0.003	1.14	0.002
VPC	Strata	0.4%		0.4%		0.5%	
	School	11.8%		11.4%		9.9%	

PCV	Strata	0.1%	0.0%	0.0%
	School	12.8%	11.%	5.5%

Table S3. Notable intersectional stratum effect changes between 2019 and 2020 for Model 2 (adjusted for student sociodemographic characteristics).

Difference		Stratum effects			Sociodemographic characteristics				Number of students		
2020 - 2019	2019 - 2018	2018	2019	2020	KS2 group	Gender	Ethnicity	IDACI	2018	2019	2020
-0.20	0.12	0.01	0.13	-0.07	4	Female	Black	Mid	259	346	425
-0.12	0.02	-0.02	0.01	-0.12	1 - Lowest	Female	Mixed	Mid	374	331	389
-0.10	0.06	-0.04	0.02	-0.08	10 - Highest	Male	Asian	Low	432	309	343
-0.10	0.07	-0.14	-0.07	-0.17	1 - Lowest	Female	Asian	Low	279	237	214
0.10	-0.07	-0.01	-0.08	0.03	3	Male	Mixed	Low	177	170	233
0.10	-0.02	-0.04	-0.07	0.04	4	Female	Mixed	Mid	293	483	480
0.11	-0.09	-0.02	-0.11	0.01	8	Male	Black	Mid	103	154	141
0.12	-0.05	0.02	-0.03	0.09	10 - Highest	Male	Mixed	High	241	150	183
0.13	-0.04	-0.08	-0.11	0.02	2	Male	Asian	Mid	716	842	796
0.14	-0.06	-0.05	-0.11	0.03	1 - Lowest	Male	Black	Low	83	95	84
0.18	-0.08	0.01	-0.07	0.11	6	Female	Other	High	114	123	124

Note: A notable change is identified as larger than 0.1 in size and where the 2020-2019 change is larger than the 2019-2018 change.

Table S4. Notable intersectional strata effect changes between 2018 and 2019 for the model adjusted for student sociodemographic characteristics.

Difference		Stratum effects			Sociodemographic characteristics				Number of students		
2019 - 2018	2020 - 2019	2018	2019	2020	KS2 group	Gender	Ethnicity	IDACI	2018	2019	2020
-0.19	0.08	0.12	-0.08	0.01	7	Male	Mixed	Low	332	395	428
-0.15	0.05	0.08	-0.07	-0.02	2	Male	Black	Low	80	88	67
-0.13	0.02	0.12	-0.01	0.01	10 - Highest	Female	Mixed	Mid	271	232	165
-0.10	0.00	0.02	-0.08	-0.07	2	Male	Other	Mid	77	108	98
0.10	-0.02	-0.09	0.01	-0.01	10 - Highest	Male	Black	Mid	98	79	69
0.10	-0.03	-0.06	0.04	0.01	8	Female	Black	Mid	138	151	137
0.10	-0.07	-0.03	0.07	0.01	3	Male	Other	High	163	190	242
0.11	-0.06	-0.10	0.02	-0.04	2	Female	Asian	Low	223	212	210
0.11	-0.07	-0.08	0.03	-0.04	3	Male	Mixed	High	394	505	511
0.12	-0.05	-0.12	0.00	-0.05	5	Male	Mixed	Mid	320	435	464
0.12	-0.20	0.01	0.13	-0.07	4	Female	Black	Mid	259	346	425
0.12	-0.09	-0.03	0.09	0.00	6	Female	Mixed	Mid	200	242	304
0.13	-0.13	-0.04	0.09	-0.04	9	Male	Mixed	High	378	467	483
0.19	-0.03	-0.12	0.07	0.04	1 - Lowest	Female	Other	Mid	136	132	147

Note: A notable change is identified as larger than 0.1 in size.

Supplementary figures

Figure S1. Average GCSE grade summarised by student sociodemographic characteristics (KS2 score decile, gender, ethnicity, IDACI tertile) and school characteristics (type and size) for 2018, 2019, and 2020.

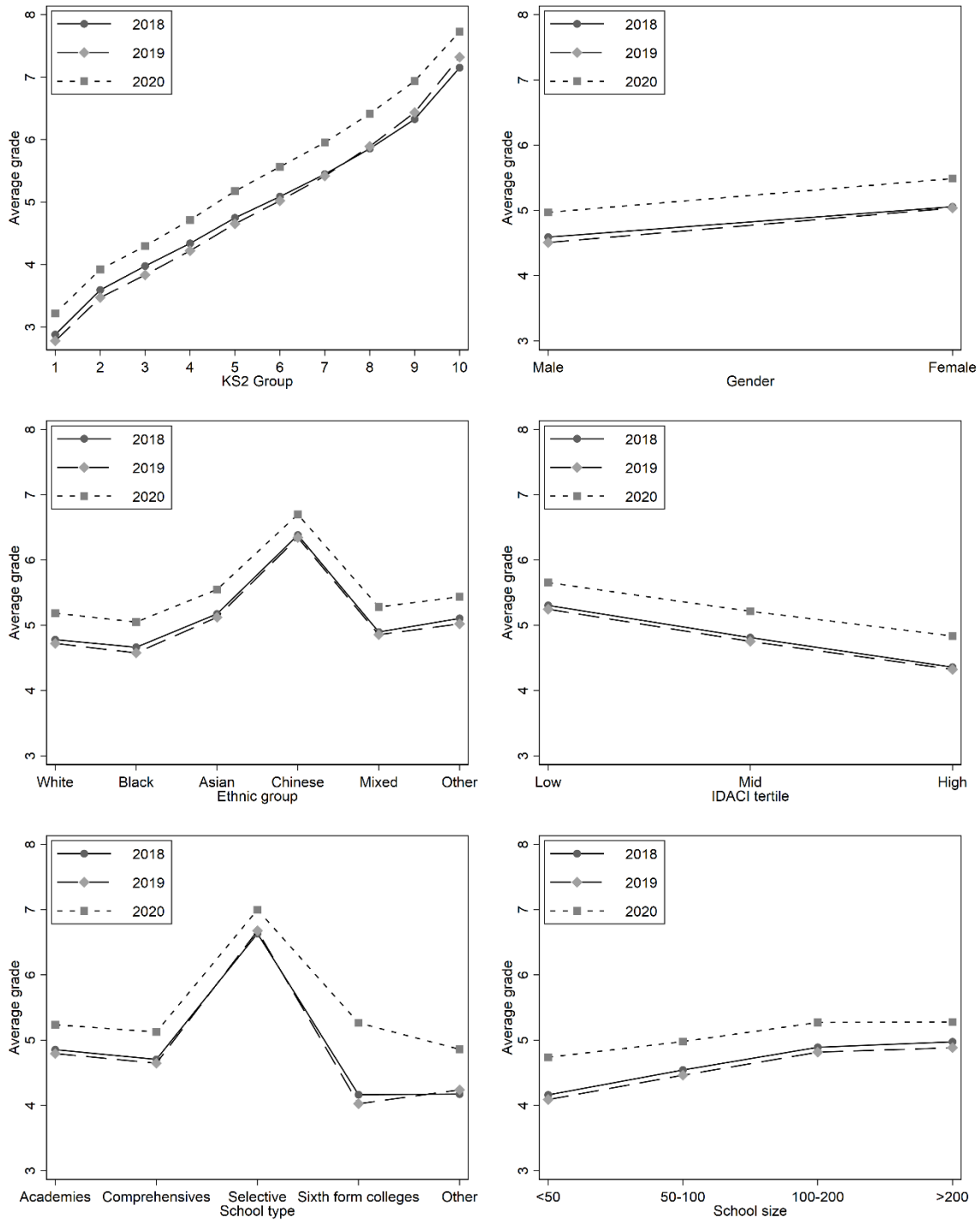


Figure S2. Average GCSE score summarised by student FSM status, EAL status, SEN status, and IDACI quintile for 2018, 2019, and 2020.

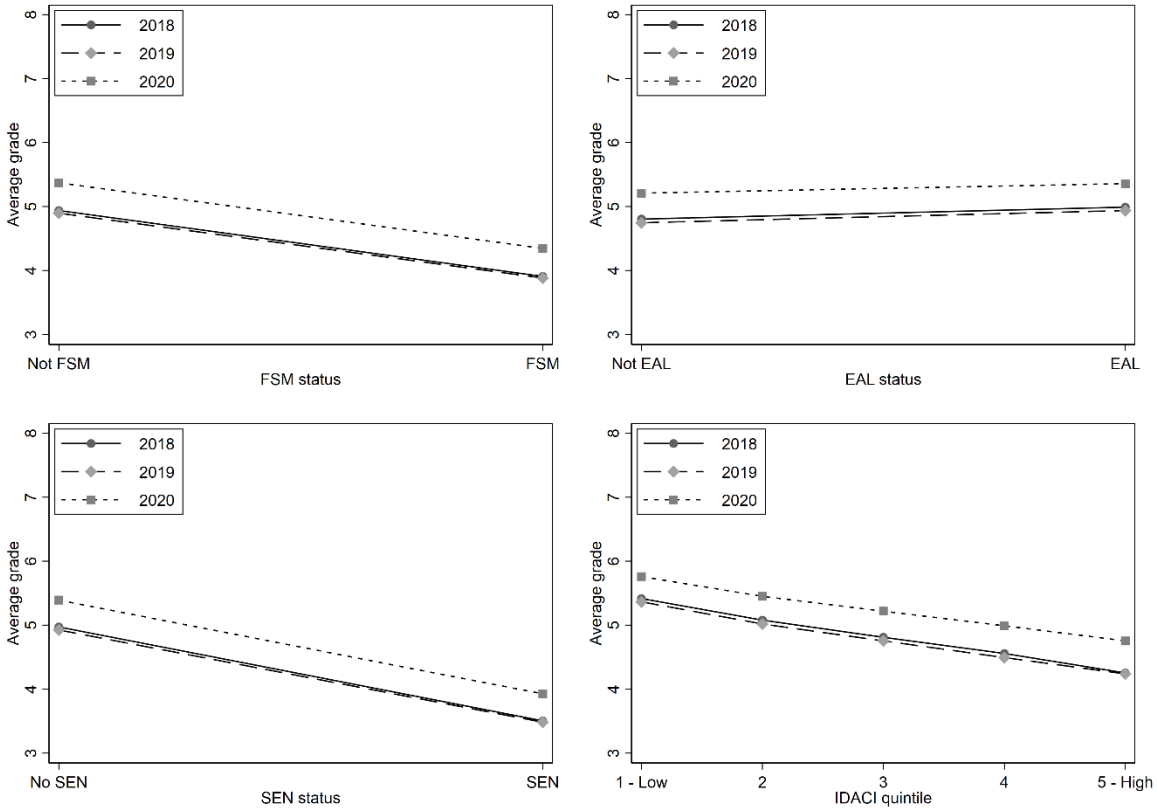


Figure S3. Scatterplots of stratum effects for the model adjusted for student sociodemographic characteristics and school characteristics (Model 3) comparing 2018 and 2019 (left) and 2019 and 2020 (right).

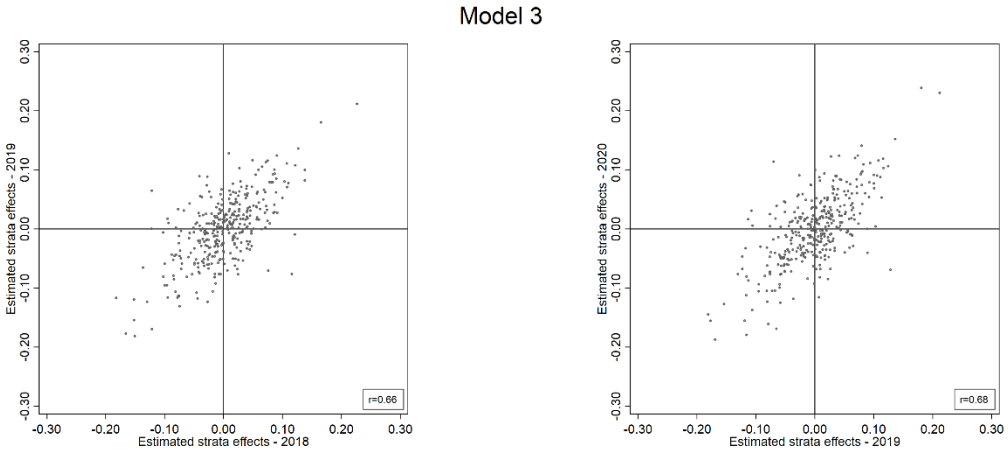


Figure S4. Scatterplots of school effects for the unadjusted model (Model 1, top row) and model adjusted for student sociodemographic and school characteristics (Model 3, bottom row) comparing 2018 and 2019 (left) and 2019 and 2020 (right).

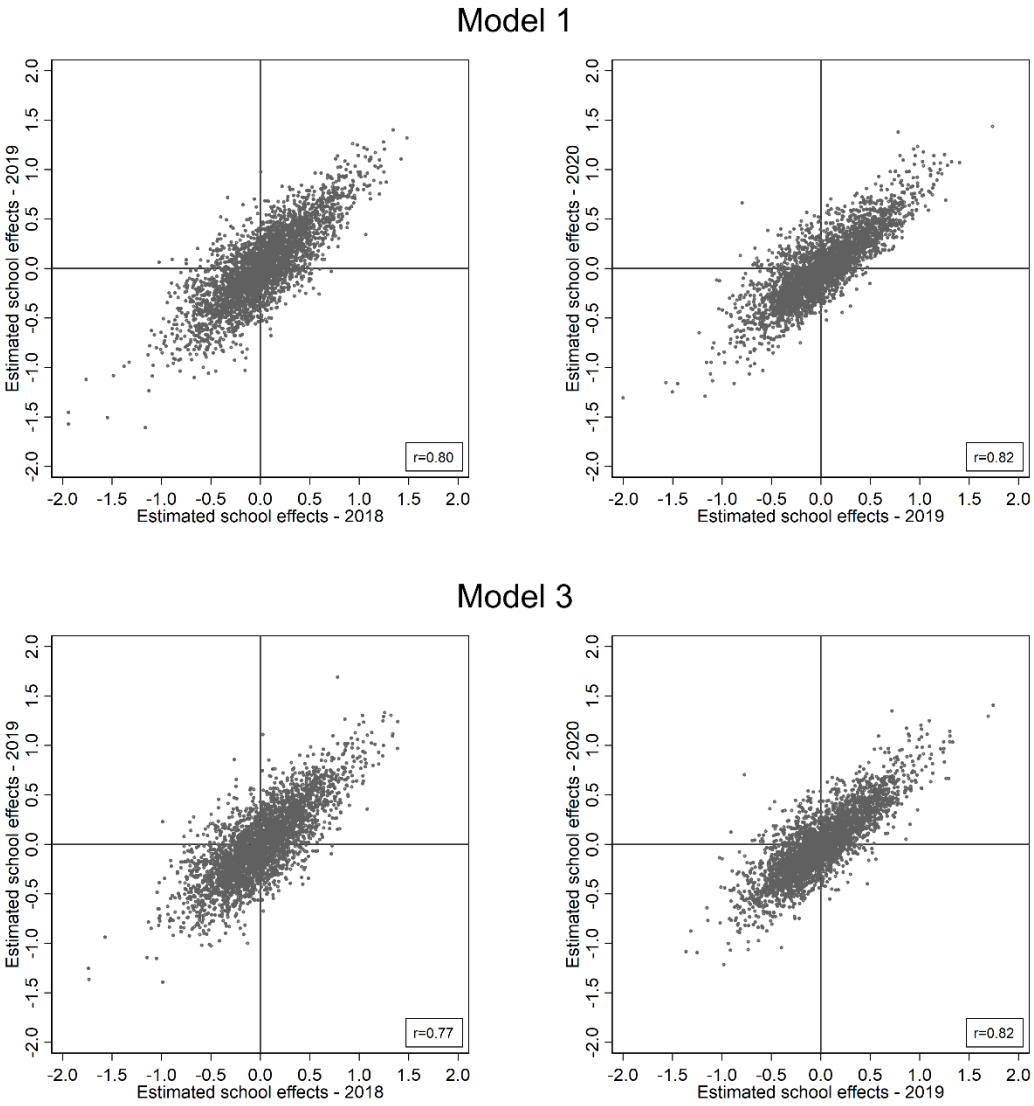


Figure S5. Histograms of the average difference between school effects comparing 2019 and 2018 (top row) and 2020 and 2019 (bottom row) for the model adjusted for student and school characteristics (Model 3).