Grade expectations: how well can past performance predict future grades?

Gill Wyness, Lindsey Macmillan, Jake Anders & Catherine Dilnot

Published online: 19 Sep 2022.

Submit your article to this journal ☑

Article views: 819

View related articles ☑

View Crossmark data ☑

Routledge
Taylor & Francis Group

Grade expectations: how well can past performance predict future grades?

Gill Wyness [a,b], Lindsey Macmillan [a], Jake Anders [a] and Catherine Dilnot [c]

ªUCL Centre for Education Policy and Equalising Opportunities, London, UK; ᵇCentre for Economic Performance, LSE, London, UK; ᶜOxford Brookes Business School, Oxford, UK

ABSTRACT

Students in the UK apply to university with teacher-predicted examination grades, rather than actual results. These predictions have been shown to be inaccurate, and to favour certain groups, leading to concerns about teacher bias. We ask whether it is possible to improve on the accuracy of teachers' predictions by predicting pupil achievement using prior attainment data and machine learning techniques. While our models do lead to a quantitative improvement on teacher predictions, substantial inaccuracies remain. Our models also underpredict high-achieving state school pupils and low socio-economic status pupils, suggesting they have more volatile education trajectories. This raises questions about the use of predictions in the UK system.

1. Introduction

The Covid-19 pandemic has led to unprecedented disruption in education systems across the globe, with many countries either postponing, or cancelling formal examinations (UNESCO 2020). In the UK, the setting of this study, all formal examinations for the classes of 2020 and 2021 were cancelled. Instead of the usual system of externally set and marked exams, teachers played a fundamental role in how grades were assigned. Instead of sitting exams, pupils were assigned grades based on predictions by their teachers. In 2021, teacher assessment (informed by in-class tests, coursework and other methods) formed the basis of all grades assigned. Similar models of replacing exams with teacher assessment were used in France, Italy, the Netherlands and Norway in 2020 (Oppos 2020).

While these are extraordinary measures brought on by the pandemic, teacher-predicted grades are used regularly in England, for those applying to university.[1] This is because students apply to university well in advance of sitting their school leaving exams (A levels). In brief, around a year before entering university, students must apply to up to five courses via the UK's centralised university application system (the University and College Applications Service, or UCAS) using their teacher predicted grades. Universities then make students offers, conditional on achieving certain grades. Students then have a short period to choose a first and backup course. The student is committed to their first choice course if they gain the required A-level grades, and the university is committed to accepting them. If they do not meet the grades of their first choice, they are committed to their backup choice, and again the backup university-course is committed to them. Only after this decision process is complete do students then sit their A-level exams, receive the results, and know which university-course they will attend. To our knowledge, England is the only country in which this system is used.[2]

Predicted grades have many similarities with teacher-assessed grades, and in this paper we use these concepts interchangeably. However, there are some distinctions. Teacher assessment tends to take the form of in-class tests, coursework and homework marked by teachers. Predicted grades of the sort used solely in the UK may be based on results of in-class tests and homework, and so are an informed decision. However, they may also feature an element of 'aspiration' whereby the teacher may set a target grade for a pupil to aspire to as a means of incentive.

These predicted grades are therefore crucial determinants of students' university applications. But previous research (Delap 1994; Gill and Benton 2015; Everett and Papageourgiou 2011; UCAS 2016; Murphy and Wyness 2020) has shown these predictions to be inaccurate. For example, research by Murphy and Wyness (2020) showed that only 16% of university applicants were correctly predicted, and that high achievers from poorer backgrounds received less generous predictions. As well as leading to widespread concerns about teacher bias, this also has important consequences for the students involved. For example, Murphy and Wyness (2020) show that 'underpredicted' students more likely to apply to universities that they are overqualified for, potentially leading to mismatch in higher education (Campbell et al. 2022). Thus, these predictions can impact the wider life chances of pupils in post-secondary education.

In this paper, we ask for the first time, whether we can improve on teacher predictions of pupil attainment, by predicting pupil performance at the end of formal schooling based purely on their past performance. This is an empirical question that we look to answer using non-linear regression modelling and a highly-flexible machine learning approach. On the one hand, the use of models based on large-scale data may avoid issues of teacher bias and/or manipulation of grades.[3] On the other hand, our models will fail to capture the full information set to which teachers have access, including recent test and mock exam results, the knowledge of the trajectories of individual pupils, and external influencing factors that teachers are able to account for in making their professional judgements. We also ask whether certain groups of pupils (such as higher achieving pupils, or those studying certain subjects) perform particularly worse or better than their past results would suggest, offering an explanation for why some pupils may be 'harder to predict'. Understanding whether pupils' trajectories can be accurately predicted using their prior academic attainment is important in and of itself. Examinations in many countries (including the UK) may continue to be disrupted over the coming years (e.g. England's contingency plan for 2022 involves awarding Teacher Assessed Grades, in the event that exams are not able to go ahead as planned[4]), and even if they do go ahead, may be considered less trustworthy than in typical years. This ongoing exam disruption as a result of the Covid-19 pandemic may lead further and higher education providers to put more weight on applicants' prior attainment as a source of information on their ability. Hence, understanding whether this may produce biases is important.

Aside from the disruption caused by the pandemic, several countries (e.g. Norway and the Netherlands) typically use a mixture of both exams, and teacher-graded pupil performance over the year to determine a proportion of their final school-leaving grade. Understanding whether there may be biases arising from the teacher-based element of these assessments is of interest for policy in these countries.

Moreover, understanding whether empirical approaches to predicting grades can improve on teachers' performance is important in the context of the existing UK system. If pupils' grades can be more accurately predicted using their prior test scores, then this may be a preferable alternative (or, at least, a supplement) to teacher prediction in some cases. If there are particular groups of pupils who are 'harder to predict', this would help to guide which predictions may need to be treated with particular caution, or supplemented with more evidence, or which pupils may appear on paper to be poor future bets but may, in fact, outperform expectations. On the other hand, if our results show that, even with a comprehensive set of prior attainment results and pupil characteristics, pupil grades cannot be predicted accurately, then this highlights the difficulty faced by teachers. It also provides further evidence that the UK's predicted grades system should be re-examined, and pupils should be admitted to universities with achieved grades rather than predicted grades.

Our empirical analysis finds that using information on previous achievement in exams at age 16 leads to a quantitative improvement on teacher predictions, though considerable inaccuracies still

remain, with the total proportion of pupils correctly predicted across their best three A level grades just 26-27%, versus 16% accuracy from teacher predictions. We can improve on this further for a restricted sample of higher achievers, for whom prior achievement includes 'related' GCSEs (for example, pupils studying chemistry at age 18 who have also studied chemistry at age 16), improving the accuracy of predictions to 1 in 3. These findings are consistent across both non-linear models and Random Forest machine learning approaches. Despite having access to detailed information from linked-administrative data on past performance and demographic characteristics including school attended, this is the best that we can do with our models.

We also observe concerning heterogeneity in our ability to predict pupil grades, with differences in how well we can correctly predict, in terms of pupil attainment, school type, by SES within school type, and subject type. We find that higher achievers are more accurately predicted compared to lower achievers. As with teacher predictions, ceiling effects play a role – mechanically the higher up the achievement distribution, the lower the probability of over-prediction.[5]

Looking across school type, we find that high achieving non-selective state school pupils are 12ppts more likely to be under-predicted by 2 grades or more, relative to high achieving grammar and private school pupils. This is driven by low SES pupils who are more likely to be under-predicted relative to their high SES state-educated counterparts, within high achievers. This suggests that high achieving non-selective state school pupils' trajectories between GCSE and A level are more 'noisy' than for their grammar and private school counterparts. This is particularly the case for low SES students relative to high SES students. Studies of teacher prediction accuracy (Murphy and Wyness 2020) found similar patterns, with teachers under-predicting high achieving low SES and state school pupils' compared to more advantaged pupils, and those from fee-paying schools. Our work offers one explanation for this – that the task that teachers are faced with in predicting grades is difficult, and may be more challenging for some groups of pupils.

We also observe concerning heterogeneity in our ability to predict in certain subjects. For example, maths is easier to predict among high achievers than other subjects such as history and chemistry, but for average and low achievers, the opposite is true. For those subjects without 'related' GCSEs, the task is even more challenging, with lower prediction rates across the board. For subjects such as economics and politics, there is more accuracy in predictions among high achievers, while for psychology and sociology, predictions are more accurate among low achievers.

Our paper makes a key contribution to the academic literature on the accuracy of teacher-predicted grades. There is limited research on the impact of predicted grades, though research typically reveals a high degree of inaccuracy, and a tendency toward overprediction. Delap (1994) and Everett and Papageourgiou (2011) analyse prediction accuracy by individual grade, both showing around half of all predictions were accurate, while 42–44% were over-predicted by at least one grade, and only 7–11% of all predicted grades were under-predicted.

Studies by Gill and Benton (2015), UCAS (2016) and Murphy and Wyness (2020) examine prediction accuracy according to the best 3 A-levels, both showing the majority of grades were overpredicted. All three studies found that higher grades were more accurately predicted than lower grades, and as a result high achieving pupils (i.e. those in independent schools and from high SES backgrounds) tend to be more accurately predicted than those from state schools and low SES backgrounds. Murphy and Wyness (2020) are the only study to examine predictions within achievement level, showing that among high attainers, those from low SES backgrounds receive slightly less generous grades than those from high SES backgrounds.

Our paper contributes to this small literature by comparing the efforts of teacher predictions with predictions based on empirical models. The fact that our models only make marginal improvements on the efforts of teachers highlights the challenges they face in making these predictions, but also shows that any attempts to improve teacher predictions, for example by training them to use past achievement data in their efforts, or indeed by replacing teacher predictions with those based on empirical models, would have limited success.

We also contribute to the literature on teacher bias. Work by Burgess and Greaves (2013) examines teacher assessment versus exam performance of black and minority pupils versus white pupils at age 11 (Key Stage 2), finding evidence that black and minority pupils are more likely to be under-assessed, adding to concerns about bias. Lavy and Sand (2015) look for evidence of gender bias in teacher grading behaviour by comparing their average marking of boys' and girls' in a 'non-blind' classroom exam to those in a 'blind' national exam marked anonymously. Their results show that math teachers' assessment in primary school is on average gender neutral, though there is a considerable variation in gender-biased behaviour among teachers. In cases where teachers favour boys, this can positively impact boys' future achievement, and negatively impact girls. More recent work by Burgess et al. (2022) also examines gender differences arising from teacher assessment. They exploit the random assignment of students in Denmark to a semi-external assessment (SEAM) for a subset of their subjects (as opposed to receiving a teacher-assessed grade). The study shows that the effect of SEAM is positive for female students, reducing the gender gap in graduation from post-secondary STEM degrees when boys and girls are assigned to SEAM in maths. The authors conclude that external exams give students the opportunity to demonstrate what they can actually do, as opposed to purely what their teacher thinks they can do. It is often assumed that teacher bias is at the heart of the issues of inaccuracy in teacher-predicted grades in the UK[6] – our findings offer some context to this.

Our work speaks to the related literature on differences in the effect of information on past performance on individuals' future performance. Bandiera, Larcinese, and Rasul (2015) and Azmat and Iriberri (2010) show that increasing information about previous performance can positively impact student's future performance, with stronger impacts for those with less information and for more able students. De Paola and Scoppa (2017) and Paserman (2007) find evidence that women react differently to men, in terms of future performance, to poorer past performance, albeit in very specific settings (professional tennis matches). Murphy and Weinhardt (2020) show that students who are higher ranked in primary school achieve higher test scores during secondary school, with low SES students gaining more from being highly ranked. Our results suggest that past performance indicators may be more useful to certain groups of students, such as high achievers, in a similar vein.

Finally, we contribute to the literature on educational trajectories. It is well established that pupil performance in one time period is highly correlated with their performance in previous time periods, and that differences in educational achievement emerge early in life, at primary school or even pre-school (Chowdry et al. 2013; Cunha and Heckman, 2006; Demack, Drew, and Grimsley 2000). A related literature highlights how educational trajectories differ by SES, illustrating that high-achieving pupils from low SES backgrounds fall behind their average-achieving high SES counterparts during school (Crawford, Macmillan, and Vignoles 2016; Feinstein 2003; Jerrim and Vignoles 2013). Our results suggest one possible explanation for discrepancies between external assessments and teacher assessments could be that certain groups of pupils have more noisy trajectories than others, making the task more difficult.

In summary, our findings imply that even with a comprehensive set of information on pupil prior attainment and demographics, predicting pupils' future outcomes is a very challenging task. This raises the question of why the UK continues to define such a crucial stage of the education system (transition to higher education) using predictions. It also highlights the risks in using pupils' historical attainment to make judgements upon them in the absence of current examination. Ongoing exam disruption as a result of the Covid-19 pandemic may lead further and higher education providers to put more weight on applicants' prior attainment as a source of information on their ability. Our results suggest such information should be used with caution.

In the next section, we provide some background information on the UK's system of predicted grades and the unique situation brought about by the Covid-19 pandemic. In Section 3, we detail the administrative data records used for our analysis, before outlining our methods and approach in Section 4. Section 5 summarises our main findings across all pupils, by school type, and by

subject studied. Section 6 ends with some brief conclusions and discussion of the implications for this year's situation and the future of predicted grades more generally.

2. Background and context

Unlike any other country, predicted grades are a common feature of the UK education system in 'normal' times. The UK has a centralised system of university applications, and for historical reasons, applications to university are made almost a year in advance of university entry, and, crucially, before pupils sit their exams. Applicants must therefore make their applications based on their high-school teachers' predictions of their school-leaving examination grades (A levels) rather than their actual grades. Universities make pupils offers, usually conditional on achieving their predicted grades, and pupils must then commit to their first choice and reserve courses. Only then do pupils actually sit the exams which will determine entry (see Murphy and Wyness 2020 for a detailed outline of this process).

Since 2020, these predicted grades have been under far greater scrutiny due to the cancellation of formal examinations caused by the Covid-19 pandemic. In 2020, teacher predictions formed the basis of the GCSE and A level grades awarded, alongside their rankings of pupils. The original plan was for the exam regulator, the Office of Qualifications and Examinations Regulation (Ofqual) to moderate grades at the school level,[7] holding constant the individual rankings. However, the results of the moderation process were met with public outcry, with many reports of pupils' moderated grades being vastly below their original predictions. The government were eventually forced to capitulate and allow the new predictions to stand, resulting in significant grade inflation in 2020 (Nye and Thompson 2020). In 2021, teacher assessment formed the sole basis for grades awarded at both GCSE and A level. Teachers were allowed to inform pupil grades with assessment and coursework carried out throughout the year, though no moderation to standardise grades across settings was used. The lasting impact of this change in examination grading is unclear. Both 2020 and 2021 saw unprecedented grade inflation, with many pupils advantaged, since they were able to gain places at more selective universities on the basis of these grades. On the other hand, the classes of 2020 and 2021 may find their exam results are treated as less informative than those of previous cohorts.

The UK was not the only country to experience disruption during the exam period as a result of the pandemic, but countries responded in a variety of different ways. Italy chose to adopt the same model as the UK in replacing exams with teacher assessment. As documented by Oppos (2020), several countries followed a similar, but slightly different approach by replacing exams with average grades calculated from their scores on classroom tests and homework which had already been set throughout the year. This was the case in France, Norway (where 80% of students' final grades are determined by classwork anyway) and the Netherlands (where earlier school exams which normally contribute 50% of the final grade were used). Meanwhile, Spain, China, Azerbaijan, Ghana and Vietnam simply chose to move the exams to a later date, while Hong Kong went ahead with their exams using social distancing. Finally, the USA adapted their university entry test (the AP) by rewriting the exams, and allowing students to complete the exams remotely.

In the aftermath of the UK's exam cancellation, the ongoing system of the use of predicted grades came under heavy scrutiny, with many calling for their abolition and replacement with a post-qualification admissions system (PQA). The UK government consulted on this, but has ultimately chosen to retain the current system of predicted grades. This continues to be a controversial policy, however, and the issue is likely to be revisited in years to come (as it has been in the past).

3. Data

We study the group of pupils who took post-compulsory age 18 exams (A levels) in UK schools. We use administrative records from the National Pupil Database for a cohort of state and privately educated pupils who took their A levels in 2008 ($N = 238,898$).[8] From these records, we can observe information about pupils' final A level performance, including their grades and subjects studied, as well as

detailed information about their past performance in (compulsory) age 16 GCSEs, including the grade and subject of every prior qualification. The 238,898 pupils took at least one A level, and between them took a total of 639,298 A levels overall. This excludes community languages (Urdu, Turkish, Polish, etc.) and vocational A levels, which have since ceased to be offered. It also excludes General Studies and Critical Thinking which have also since been removed.

To predict A level performance in our models we use information on prior achievement including each grade (A*, A, B, C, below C, not entered) in each of 57 GCSE subjects, the total point score from GCSEs and equivalents, and a squared term for total point score. In Appendix Table A1, we present the proportion entering and achieving each grade in each subject. In our robustness checks, we test if our accuracy is improved by including additional individual-level predictors including gender, ethnicity, school type and a measure of socioeconomic status (SES), and in a separate model allowing predictions of cut-offs between grades to vary by school.[9] As well as GCSE scores, pupils are also tested during and at the end of primary school (Key Stages 1 and 2, at ages 7 and 11, respectively), as well as at age 13 (Key Stage 3). Key Stage 1 and 3 are both graded by teachers, and therefore we do not include them in our models. We have tested whether the inclusion of externally examined primary school achievement at age 11 (Key Stage 2 test scores), which is only available for state-educated pupils, improves our model further and find this adds no precision above our main model.[10]

SES is constructed following Chowdry et al. (2013) by combining information about pupil's free school meals (FSM) eligibility, with small local area (Lower Super Output Area[11]) level information about where the pupil lived from the Census (2011), including the proportion of individuals in the neighbourhood that worked in professional or managerial occupations; the proportion holding a qualification at level 3 or above; and the proportion who owned their home. This is combined with the Index of Multiple Deprivation using principal components analysis, with the resulting score then split into 5 quintiles. Private school pupils are missing SES information and so for the purposes of this analysis are included in the highest SES quintile (as in Crawford 2014).[12] Table 1 shows that higher SES pupils outperform lower SES pupils. The average GCSE scores of high A level achievers are higher than that of low achievers.

Given the issue of ceiling effects for high achievers (as noted, it is easier to predict grades where the distribution is truncated) we split all of our analysis by achievement, grouping pupils into three groups: low achievers (below CCC), middle achievers (CCC to ABB) and high achievers (AAB or above[13]). Of these pupils, Table 1 shows that 45% achieved below CCC grades, 36% achieved between CCC and AAB, and 19% achieved higher than AAB. Table 1 illustrates that a higher proportion of female pupils (54%) sit A levels than males, with females typically outperforming males.

We present our analysis across sub-groups, including type of school attended using linked data from the 2008 Spring Census. We split the school type into three categories: non-selective state (where the vast majority of pupils are educated), selective state (grammar schools), and private schools. Table 1 shows that private school pupils and selective (grammar) school pupils outperform pupils from non-selective state schools. We also split our analysis by high and low SES within state schools for a sub-sample of our total population.

Finally, we consider how these predictions vary across the most popular subjects studied at A level for (a) the 5 most popular subjects with a 'related' GCSE (Maths, Biology, Chemistry, History, and English literature), and (b) the 5 most popular subjects where there is no 'related' GCSE (Psychology, Sociology, Economics, Government and Politics, and Law). For those with a 'related' GCSE, we are focusing on a restricted group of pupils who are more likely to be higher achieving than the full sample, with only 18% achieving below CCC, 47% achieving between CCC and ABB, and 35% achieving AAB or higher. The gender split is slightly more balanced than the full sample, although males are still more likely to be lower achieving and females higher achieving. There is a higher proportion of private and selective school pupils in the restricted sample, and on average they are typically from higher SES families. They also have higher prior achievement and final A level scores. That the restricted sample is more skewed towards higher attaining pupils may be because of the more traditional nature of the subjects with related GCSEs which such pupils/schools may have a tradition of

Table 1. Descriptive statistics for full sample (all those with at least one counting A level in the data) and for the restricted sample of all those with at least three A levels who had done the related GCSE.

	Pupils with at least one A level	<CCC	CCC-ABB	AAB+	Restricted sample	<CCC	CCC-ABB	AAB+
All	238,898	108,146	86,442	44,310	48,464	8,900	22,623	16,941
Gender								
Female	0.54	0.51	0.57	0.56	0.52	0.44	0.54	0.53
Male	0.46	0.49	0.43	0.44	0.48	0.56	0.46	0.47
School type								
Non selective state	0.78	0.91	0.75	0.53	0.69	0.85	0.74	0.55
Selective state	0.09	0.04	0.1	0.16	0.13	0.08	0.12	0.18
Private	0.13	0.05	0.15	0.30	0.17	0.07	0.14	0.27
Mean SES quintile[19]	3.38	2.91	3.53	4.05	3.61	3.1	3.52	3.97
Attainment								
GCSE point score[20]	490	450	503	564	533	476	520	581
Points from best three A levels[21]	8.9	4.9	11.0	14.6	11.5	6.4	11.2	14.7

Notes: Full sample of all students achieving at least one A level. Grade boundaries are chosen to replicate Murphy and Wyness (2020). Low achievers with grades below 3 Cs, middle achievers with grades from CCC to ABB, and high achievers from AAB to A*A*A*.

teaching. We can then only compare within achievement groups across samples. We show our main results for both our full and restricted samples for completeness.

4. Methods

4.1. Ordered probit

To model performance we use two different approaches. Our first approach is a latent variable formulation of a variable y^*, being the underlying performance in A level assessments, with observed outcomes 0–5 representing A level grades achieved.[14] In this case, y^* has real existence (marks awarded by examination boards before grade boundaries are determined), and the observed outcomes increase monotonically in the value of y^*.

Assume k categories of the grade (here $k = 6$), with $k-1$ cut points at the grade boundaries, where τ_k is the value of the latent variable at cut point k. The models are fit to estimate τ_1 to τ_{k-1} subject to the following relationship:

$$y = \begin{cases} y_1 & \text{if } y^* < \tau_1 \\ y_2 & \text{if } \tau_1 \leq y^* < \tau_2 \\ \dots \\ y_{k-1} & \text{if } \tau_{k-2} \leq y^* < \tau_{k-1} \\ y_k & \text{if } y^* \geq \tau_{k-1} \end{cases}$$

and y_i^* is modelled for pupil i across GCSE grades j, and subjects m, for our full sample as follows:

$$y_i^* = \sum_{j=2}^{6} \beta_1^j I(GCSEsubj1_i = j) + \sum_{j=2}^{6} \beta_2^j I(GCSEsubj2_i = j) \dots .$$

$$+ \sum_{j=2}^{6} \beta_s^j I(GCSEsubjm_i = j) + \beta_{s+1} \, GCSEpts_i + \beta_{s+2} \, GCSEpts_i^2 + \varepsilon_i$$

(1)

where $I(GCSEsubjm_i = j)$ is the achieved grade j, by pupil i, in GCSE subject m, and *GCSEpts* is the total point score (and squared) for both GCSEs and equivalent qualifications, to account for equivalent qualifications in addition to GCSE achievement.

We have tested various alternative models in order to try to improve the overall predictive accuracy of the models, which we report in Appendix Table A3. We show that including individual demographic characteristics, such as gender, ethnicity, and socio-economic status quintile, and including school-level indicators, does very little to improve the accuracy of the models. As previously described, our dataset also contains pupil attainment at age 11 for state school pupils, but this age 11 attainment data are not available for pupils at private schools. Including this attainment data also made no substantive difference to the proportion of correct predictions (with results available on request). We therefore focus on prior achievement at age 16 in our main specifications for clarity of what is being used to predict grades.

It can be shown that the probability of an individual i falling into category k or below can be given by the link function:

$$g(\gamma_{ik}) = \Phi^{-1}(y_{ik}) = \tau_k + \boldsymbol{x_i}\boldsymbol{\beta}$$

where $\gamma_{ik} = \Pr(y_i \leq k)$ and $\Phi^{-1}(y_{ik}) =$ cdf of the error term ε_i, which is assumed to be standard normal. We make the parallel regression assumption that the vector of coefficients $\boldsymbol{\beta}$ have the same relationship with the latent variable across all grade boundaries, allowing us to use this ordered probit formulation. Our second, machine learning, approach relaxes this assumption.

For each A level subject (60 in total) we run a separate ordered probit model to estimate the predicted probability of achieving each grade for each pupil based on their prior achievement, using the cut points and coefficients from our predictor variables. The probability for each individual falling into each grade category is predicted, and the grade with the highest probability assigned. This is then compared with the actual grade in that A level subject received for each individual. We subtract the actual grade from the predicted grade, so positive numbers imply over-prediction and negative numbers imply under-prediction. Across all pupils, this gives us the distribution of over- and under-predictions for all A level subjects.

To enable us to compare our findings with those from previous work looking at the accuracy of teacher's predicted grades (Murphy and Wyness 2020), we calculate 'best three' distributions by aggregating results for individual pupils. For each pupil, the A levels with the three highest marks are identified and the over- or under-prediction for these three aggregated to give a net over/under prediction for 'best three'. We present our results for two samples, the full cohort of pupils, and a second more restricted sample of those pupils who take 'related' GCSE subjects. The more restricted sample are the focus of our analysis when we consider differences by subject type.

4.2. Random forests

In our second approach, we estimate the predictability of A level grades by employing the supervised machine learning algorithm of Random Forests (Breiman 2001) to carry out this prediction task. This has twin advantages: (1) it is extremely flexible in its approach to how A levels are predicted from GCSE grades, removing the need for assumptions such as the parallel regressions assumption; (2) it is robust to concerns about overfitting of the model which would artificially boost within-sample prediction rates but reduce out-of-sample prediction rates.

Random Forests work by 'growing' a large number of decision trees (in our case 2000) each on a bootstrapped sample of the dataset. At each step of the decision tree a random subset of the predictor variables (in our case 8 out of 58 predictors, in line with the suggested default of the square root of the number of predictors) are tested to determine the best split in one of the selected variables in order to classify the outcomes. While each individual decision tree is likely over-fit on its bootstrapped sample, this issue for out-of-sample prediction is overcome by aggregating the

trees and using 'votes' from each of the trees to determine the predicted classification from the forest as a whole. The method is highly flexible to potential interaction between predictors as there is no assumption that there would be the same split on a different variable conditional on the first. We apply the Random Forests algorithm using the R package developed by Liaw and Wiener (2002).

We grow a separate Random Forest for each A level subject among pupils who have an observed grade for this subject; in each case potential predictors available to the algorithm include the total GCSE points score (it is not necessary to include the squared term given that a decision tree is based on non-continuous splitting of predictors), and a full set of GCSE grades across subjects as used in the regression modelling – missingness due to lack of entry is imputed as −1 with the Random Forest algorithm effectively able to recognise this as a categorical difference and use this information for prediction, as appropriate.

Predictions compared to observed grades in each subject are then aggregated across individuals using the same approach as following the regression analysis in order to provide analogous estimates of precision. Note that these would not necessarily be exactly the same as the out-of-bag accuracy estimates computed internally by the Random Forests algorithm, but are used for maximum comparability with the regression model predictions and remain robust to overfitting concerns by virtue of the overall prediction classification approach of the algorithm (and, in any case, have been compared and are extremely similar).[15]

5. Results

Table 2 shows the distribution of over- and under-prediction of 'best three' A level grades, for the full sample of pupils across the distribution of achievement for our ordered probit model. Our model correctly predicts the best 3 A level grades for 27% of pupils. This is 11 percentage points higher than the 16% of correct predictions found in Murphy and Wyness (2020) based on teacher's assessment of 'predicted grades' used in the university applications system in the UK. A further 34% are over- or under-predicted by 1 grade, while 25% are over-predicted by 2 grades or more, and 14% are under-predicted by the same amount. In total, 44% of pupils are over-predicted (which is lower than the proportion found by Murphy and Wyness) and 29% are underpredicted, hence both the findings of Murphy and Wyness and ours point to a tendency towards overestimation of future grades either by teachers (as in Murphy and Wyness) or purely according to past results (as in this paper).[16]

Figure 1 shows the corresponding distribution for individual A level grades (as opposed to pupils' best 3), illustrating just below 50% are correctly predicted, with over 20% being over-predicted by one grade, and a further 10% being over-predicted by two grades or more. Around 20% are under-predicted, while fewer than 5% are under-predicted by two grades or more.[17]

In Appendix Table A1 we show which GCSE subjects are significant predictors of the four most popular A level subjects (maths, English, biology and psychology). Interestingly, performance at English language GCSE is not a significant predictor of maths A level grade, but performance at maths GCSE is a significant predictor of English A level grade.

Appendix Table A2 restricts the sample to only those taking 3 A levels, showing a very similar distribution to that found in Table 2. This suggests that including those with reduced risks of misclassification (those with fewer A levels) are not driving our findings.

Columns 2–4 of Table 2, and Figure 2 illustrate that a far higher proportion of 'high achievers' (AAB or above) are correctly predicted, with 55% of this group assigned the same predicted grades from our model as the grades they went on to achieve. This highlights the important point (as also found in the literature on teacher predictions) that ceiling effects mean that it is easier to predict achievement of those at the top of the distribution. There is far more variability in the middle and lower parts of the achievement distribution, with only 19–23% of these pupils correctly predicted across (up to) their best three A levels, and far more over-prediction than seen for high achievers. 34% of low achievers and 27% of middle achievers are over-predicted by the model

Table 2. Distribution of over-prediction by pupil of best three A level grades using the ordered probit model, by A level attainment group.

Total over prediction	Full sample				Restricted sample			
	Total	<CCC	CCC-ABB	AAB+	Total	<CCC	CCC-ABB	AAB+
	%	%	%	%	%	%	%	%
−8	0	0	0	0.1	0	0	0.1	0.1
−7	0.1	0	0.1	0.2	0.1	0	0.1	0.1
−6	0.3	0	0.5	0.6	0.3	0	0.5	0.3
−5	0.7	0.1	1.2	1.3	0.7	0.2	1.0	0.7
−4	1.7	0.3	3.1	2.5	1.7	0.3	2.2	1.8
−3	3.8	1.4	6.5	4.6	3.6	1.4	4.9	2.9
−2	7.6	5.0	10.8	7.9	6.4	3.2	8.2	5.6
−1	13.8	13.2	15.8	11.3	10.5	6.9	12.7	9.5
0	27.2	22.8	18.5	54.7	31.5	12.4	16.6	61.6
1	19.5	22.9	16.8	16.6	17.4	16.3	17.8	17.3
2	12.4	15.4	14.9	0	12.3	18	19.2	0
3	6.8	9.0	7.5	0	7.5	15.1	10.2	0
4	3.4	5.0	3.0	0	4.1	11.1	4.4	0
5	1.6	2.6	1.1	0	2.0	6.9	1.6	0
6	0.7	1.2	0.3	0	1.0	4.0	0.5	0
7	0.3	0.6	0	0	0.5	2.5	0	0
8	0.1	0.2	0	0	0.2	0.9	0	0
9	0	0.1	0	0	0.1	0.5	0	0
10	0	0	0	0	0	0.2	0	0
11	0	0	0	0	0	0.1	0	0
Number of pupils	238,898	108,146	86,442	44,310	48,464	8900	22,623	16,941
% overpredicted	44.8	57.0	43.6	16.6	45.1	75.6	53.7	17.3
% underpredicted	28.0	20.0	38.0	28.5	23.3	12.0	29.7	21.0

Notes: Full sample of all students achieving at least one A level. Grade boundaries are chosen to replicate Murphy and Wyness (2020). Low achievers with grades below 3 Cs, middle achievers with grades from CCC to ABB, and high achievers from AAB to A*A*A*.

by two grades or more. Interestingly, average achievers are more likely to be underpredicted compared to high or low achievers, with 38% of average achievers are under-predicted, compared to 29% of high achievers, and just 20% of low achievers.
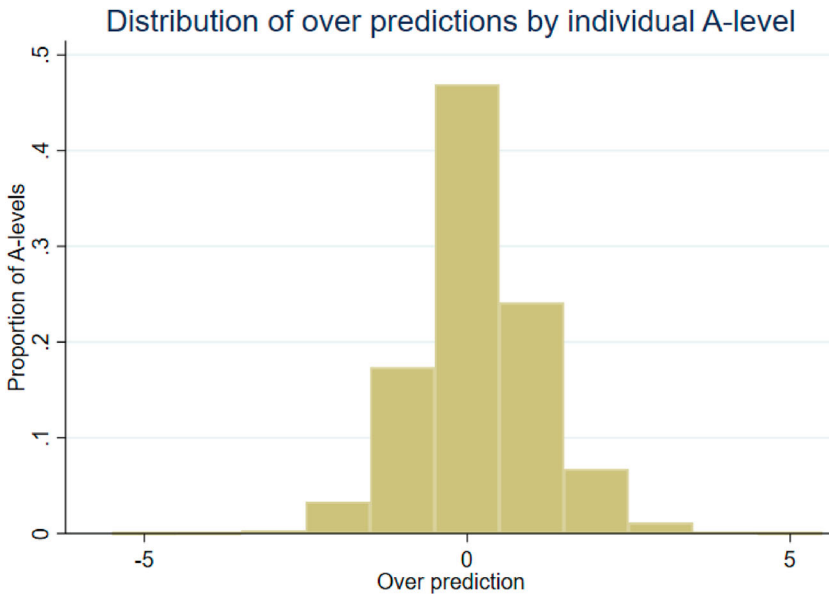

Figure 1. Distribution of over predictions by individual A level – full sample.
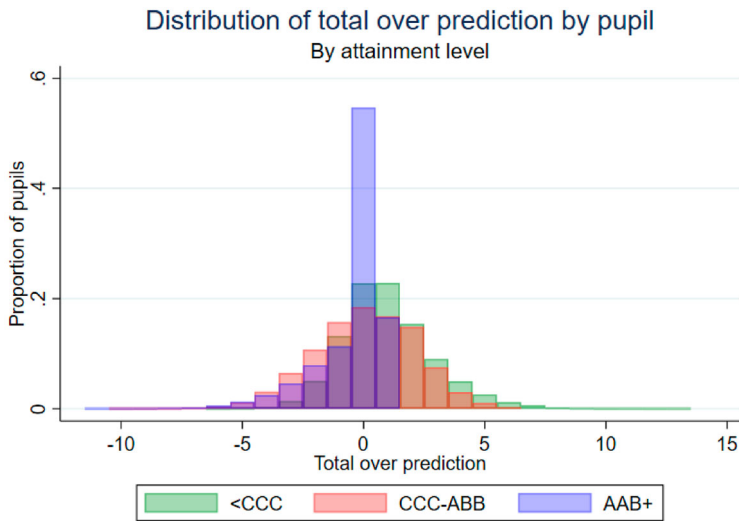
Figure 2. Distribution of total over-prediction by pupil: low achievers (<CCC), average achievers (CCC-ABB) and high achievers (AAB+).

Appendix Table A3 repeats our results with alternative specifications to attempt to improve the predictive power of our model. Demographic variables (gender, ethnicity, SES quintile and school type) were added to model 2 to see if their inclusion increased correctly predicted proportions, with the same rates of total prediction overall (27%). To include individual school indicators (i.e. school random effects), we create a simplified model for computational reasons, controlling for GCSE subject dummies, and total point score overall. Model 3 presents this simplified specification for our main model. Model 4 then compares these predictions to a model including school indicators. In A3 we also include results for a restricted sample which excludes those at private schools or with a missing SES quintile (run for main specification, and M2, which includes demographics), which does not affect our overall conclusions. Overall, there is very little difference between a model based only on school achievement and demographic indicators (M3), compared to a model using school achievement and school indicators (M4).[18]

The second panel of Table 2 focuses on our restricted sample of pupils who take 3 A levels having previously studied 'related' GCSE subjects. Given the differences in the composition of the restricted sample, it is important to compare those with similar levels of achievement. We can see that those pupils with potentially more useful information about their prior achievement in the subject of study improves the prediction among high achievers, with 62% of pupils correctly predicted across their three best A level grades. Average achievers are similarly predicted across the sample models, although with slightly higher rates of over-prediction (54% compared to 44% in full sample) and lower rates of under-prediction (30% compared to 38% in full sample). Low achievers with 'related' GCSEs are less likely to be correctly predicted, with only 12% correctly predicted compared to 23% for full sample. This group is more likely to be over-predicted, with 76% over-predicted compared to 57% for the full sample. Table 3 repeats the results in Table 2, but uses the predictions from the flexible Random Forest machine learning approach rather than from the ordinal probit regression models. The rates of accurate prediction across the two alternative approaches are almost identical, suggesting that even with a fully flexible approach to modelling the prior achievement data, we struggle to improve on correctly predicting more than a quarter of pupils. For this machine learning approach, 26% of pupils are correctly predicted across their three best A levels, compared to 27% using ordered probit. The prediction rate across levels of achievement, and in terms of over- and under-predicting is also almost identical, with a higher proportion of high

Table 3. Distribution of over-prediction by pupil of best three A level grades using Random Forest, by A level attainment group.

Total over prediction	full sample				restricted sample			
	Total %	<CCC %	CCC-ABB %	AAB+ %	Total %	<CCC %	CCC-ABB %	AAB+ %
−8	0.1	0	0.1	0.2	0.1	0	0	0.1
−7	0.2	0	0.2	0.4	0.2	0	0.2	0.2
−6	0.4	0	0.6	0.8	0.4	0	0.5	0.4
−5	0.9	0.1	1.5	1.5	0.8	0.1	1.1	0.9
−4	2	0.4	3.6	2.8	1.9	0.4	2.4	2.1
−3	4.2	1.7	6.9	5	3.8	1	5.1	3.5
−2	7.8	5.4	10.8	8.1	6.7	2.6	8.6	6.4
−1	13.7	12.8	15.1	13.4	11.1	6.1	13	11.1
0	25.8	21.4	17.6	52.7	30.3	10.8	15.9	59.7
1	18.6	21.6	16.7	15.1	16.7	16.1	17.7	15.6
2	12.4	15.6	14.7	0	12.1	18	18.8	0
3	7.1	9.8	7.5	0	7.6	16.2	10	0
4	3.7	5.5	3.3	0	4.2	11.4	4.4	0
5	1.8	3	1.2	0	2.2	7.8	1.7	0
6	0.8	1.5	0.3	0	1.1	4.6	0.5	0
7	0.3	0.7	0	0	0.5	2.8	0	0
8	0.1	0.3	0	0	0.2	1.3	0	0
9	0.1	0.1	0	0	0.1	0.6	0	0
10	0	0	0	0	0	0.2	0	0
11	0	0	0	0	0	0.1	0	0
Number of pupils	238,898	108,146	86,442	44,310	48,464	8900	22,623	16,941
% overpredicted	44.9	58.1	43.7	15.1	44.7	79.1	53.1	15.6
% underpredicted	29.3	20.4	38.8	32.2	25.0	10.2	30.9	24.7

Notes: Full sample of all students achieving at least one A level. Grade boundaries are chosen to replicate Murphy and Wyness (2020). Low achievers with grades below 3 Cs, middle achievers with grades from CCC to ABB, and high achievers from AAB to A*A*A*.

achievers correctly predicted, more under-prediction among average achievers, and more over-prediction among low achievers. We note, however, that this headline similarity likely disguises offsetting differences associated with the reasons for use of this technique (a) more flexibility in approach to prediction and (b) more robust to concerns about possible over-fitting. Thus, we would expect the Random Forest's prediction rates to hold up better if applied to new data (e.g. a subsequent year), whereas the ordered probit models would be more likely to struggle with such out-of-sample prediction.

6. Across school type

Table 4 uses the predictions from the ordered probit specification to consider whether there is any difference in predictions across the type of school attended by the pupils at age 18. Given the similarities in predictions from the two approaches noted above we focus only on this specification for brevity. Here we compare those in any non-selective state-funded institution, to those in selective (grammar) state schools, and those attending fee-paying private schools for our full sample of respondents.

Among high achievers, where under-prediction is most common, predicting A level grades based on GCSE performance leads to 23% of non-selective state school pupils being under-predicted (by 2 or more grades) – said another way, these pupils end up doing better than expected, given their GCSE performance – compared to just 11% of grammar and private school pupils. Over-prediction is similar across school type among high achievers. This suggests that there are larger differences, or greater amounts of mismatch, between the GCSE and A level grades of high achieving non-selective state school pupils compared to grammar and private school pupils. This finding is similar to that of Murphy and Wyness (2020) that high achieving low SES students are more likely to be under-predicted by teachers. While teacher bias is one explanation of this, another is that, as we find, high

Table 4. Distribution of over-prediction by pupil of best three A level grades, full sample, by school type within A level attainment group.

| | Full sample | | | | | | | | |
| | <CCC | | | CCC-ABB | | | AAB+ | | |
Total over prediction	Non sel state %	Grammar %	Private %	Non sel state %	Grammar %	Private %	Non sel state %	Grammar %	Private %
−8	0	0	0	0	0	0	0.2	0	0.1
−7	0	0	0	0.2	0	0	0.4	0.1	0.1
−6	0	0	0	0.6	0.1	0.2	0.9	0.2	0.2
−5	0.1	0	0.1	1.4	0.4	0.4	2.0	0.6	0.5
−4	0.3	0.2	0.1	3.6	1.5	1.8	3.6	1.2	1.2
−3	1.4	0.8	1.2	7.4	3.4	4.3	6.0	2.9	3.1
−2	5.2	2.5	4.2	11.7	7.1	8.4	10.0	5.6	5.6
−1	13.6	6.8	10.6	16.5	12.4	14.3	13.8	8.4	8.7
0	23.4	13.4	18.8	18.6	17.1	18.8	47.2	62.8	63.5
1	23.1	19.5	20.9	16.2	18.7	18.5	15.9	18.3	17.1
2	15.2	19.2	16.8	13.2	21.0	19.3	0	0	0
3	8.7	14.9	11.9	6.7	11.4	8.9	0	0	0
4	4.7	9.7	6.9	2.8	4.6	3.3	0	0	0
5	2.4	6.8	4.1	0.9	1.7	1.3	0	0	0
6	1.1	3.6	2.3	0.2	0.5	0.3	0	0	0
7	0.6	1.5	1.4	0	0	0	0	0	0
8	0.2	0.8	0.4	0	0	0	0	0	0
9	0.1	0.2	0.2	0	0	0	0	0	0
10	0	0.2	0	0	0	0	0	0	0
11	0	0.1	0	0	0	0	0	0	0
Number of pupils	98,875	4113	5158	64,904	8959	12,579	23,561	7237	13,512
% overpredicted	56.1	76.5	64.9	40.0	57.9	51.6	15.9	18.3	17.1
% underpredicted	20.6	10.3	16.2	41.4	24.9	29.4	36.9	19.0	19.5

Notes: Full sample of all students achieving at least one A level. Grade boundaries are chosen to replicate Murphy and Wyness (2020). Low achievers with grades below 3 Cs, middle achievers with grades from CCC to ABB, and high achievers from AAB to A*A*A*.

achieving less affluent (or non-selective state) pupils are harder to predict, regardless of the approach, due to more noisy trajectories. This therefore highlights the risks associated with the routine use of teacher predictions, which may put certain pupils at a disadvantage.

For middle and low achievers, non-selective state school pupils are more likely to be correctly predicted (or predicted within one grade), compared to grammar or private school pupils. 77% of low achieving grammar and 65% of low achieving private school pupils are over-predicted based on their GCSE performance – they achieve grades that are lower than expected given their GCSE results – compared to 56% of non-selective state school pupils.

To explore further the findings within state schools compared to private schools, Table 5 considers predictions from our ordered probit model, splitting the sample by low SES state-educated pupils, high SES state-educated pupils, and privately educated pupils, by their final A level achievement. Within low achievers, the extent of over- and under-prediction within state school pupils between those from low compared to high SES backgrounds is well balanced. For middle and high achievers we see that low SES pupils are more likely to be underpredicted compared to high SES pupils by 6 and 12 ppts, respectively.

Across the range of achievement, non-selective state school pupils are therefore more likely to be under-predicted, and less likely to be over-predicted, while selective state and private school pupils are more likely to be over-predicted and less likely to be under-predicted. This is driven by low SES pupils within non-selective state schools among middle and high achievers. Under-prediction is more balanced among low achievers across SES groups.

One important caveat to consider is that teacher predictions may themselves influence student outcomes. This issue is noted in Murphy and Wyness (2020) who suggest that a potential reason

Table 5. Distribution of over-prediction by pupil of best three A level grades, full sample, by socio-economic status within A level attainment group.

	Full sample								
	<CCC			CCC-ABB			AAB+		
Total over prediction	Low SES %	High SES %	Private %	Low SES %	High SES %	Private %	Low SES %	High SES %	Private %
−9	0	0	0	0	0	0	0.1	0	0
−8	0	0	0	0	0	0	0.2	0.1	0.1
−7	0	0	0	0.2	0.1	0	0.3	0.1	0.1
−6	0	0	0	0.6	0.3	0.2	1.0	0.3	0.2
−5	0.1	0.1	0.1	1.3	1	0.4	2.2	0.8	0.5
−4	0.4	0.3	0.1	4.2	2.4	1.8	4.2	1.9	1.2
−3	1.4	1.3	1.2	7.2	5.8	4.3	6.3	3.8	3.1
−2	5.0	5.3	4.2	11.3	9.9	8.4	9.7	7.2	5.6
−1	13.2	13.5	10.6	16.3	15.2	14.3	12.9	10.6	8.7
0	24.2	21.6	18.8	18.4	18.5	18.8	46.9	58.2	63.5
1	23.8	22.4	20.9	15.3	17.8	18.5	16.2	16.9	17.1
2	15.3	15.5	16.8	13.4	15.7	19.3	0	0	0
3	8.0	9.9	11.9	7.3	8.6	8.9	0	0	0
4	4.5	5.1	6.9	3.3	3.2	3.3	0	0	0
5	2.2	2.8	4.1	1.0	1.2	1.3	0	0	0
6	1.1	1.3	2.3	0.3	0.3	0.3	0	0	0
7	0.5	0.7	1.4	0	0	0	0	0	0
8	0.1	0.2	0.4	0	0	0	0	0	0
9	0.1	0.1	0.2	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0
Number of pupils	16,373	9491	5158	7813	10,735	12,579	2429	6603	13,512
% overpredicted	55.6	58.0	64.9	40.6	46.8	51.6	16.2	16.9	17.1
% underpredicted	20.1	20.5	16.2	41.1	34.7	29.4	36.9	24.8	19.5

Notes: Full sample of all students achieving at least one A level. Grade boundaries are chosen to replicate Murphy and Wyness (2020). Low achievers with grades below 3 Cs, middle achievers with grades from CCC to ABB, and high achievers from AAB to A*A*.

they observe more generous predictions for high SES versus low SES students (among high attainers) may be down to heterogenous incentive effects (e.g. Hvidman and Sievertsen 2021). High ability low SES students may be more likely to respond positively to an incentive compared to high ability high SES children. This would result in increased likelihood of high ability low SES students appearing under-predicted (since their efforts would narrow the gap between the predictions and reality relative to high SES students). The outcome – that low SES students do better than predicted, relative to high SES students, would also be reflected in our models (since the students in our sample would also have received A-level predictions) so could be responsible for the similar trend we observe.

7. Across A level subjects

Are certain A level subjects easier to predict than others? We explore this question for the top 5 most studied A level subjects with and without a 'related' GCSE. Table 6 shows the mean A level points for subjects in these two groups, showing that while points are slightly higher among those subjects with 'related' GCSEs, there is a similar range between the two groups. Average points for maths (with a 'related' GCSE) and economics are similarly high, while psychology, law (both no 'related' GCSE) and biology all have lower average points.

Figure 3 shows the distribution of under- and over-prediction for the top 5 most studied A level subjects with a 'related' GCSE. The story varies across the distribution of achievement.

For high achievers, maths has the highest proportion of accurate predictions, with over 80% of maths grades predicted exactly the same as their actual grades using information on GCSE performance including their GCSE maths grade. English Literature is also well-predicted for high achievers. Table 5 shows that these subjects, along with chemistry, have high average A level performance, meaning that they are more likely to benefit from ceiling effects than the other subjects. Indeed,

Table 6. Mean A level scores by subject for the 5 most popular A levels with a related GCSE, and for those without a GCSE.

	Full sample	
	Mean points	Number
Subjects with related GCSE		
Maths	3.8	50,674
Biology	3.3	43,272
Chemistry	3.6	33,412
History	3.5	40,559
English literature	3.6	43,993
Subjects with no related GCSE		
Psychology	3.2	47,213
Sociology	3.4	24,156
Economics	3.8	13,912
Gov't and politics	3.7	10,347
Law	3.2	13,343

among average and high achievers, there are only very small proportions who are under- or over-predicted across all five of these (notably facilitating) subjects by more than 1 grade.

For average and low achievers, English literature and history are the most accurately predicted subjects, while maths and chemistry are the least accurately predicted (with low achievers particularly likely to 'miss' their maths predicted score).

Figure 4 shows the distribution of predictions among the five most popular subjects without a 'related' GCSE, across the distribution of achievement. Note that in all cases, prediction rates are less accurate for these subjects, relative to those subjects with a 'related' GCSE, again indicating

Figure 3. Distribution of over predictions for the five most popular A level subjects with related GCSEs – low achievers (<CCC), average achievers (CCC-ABB) and high achievers (AAB+).
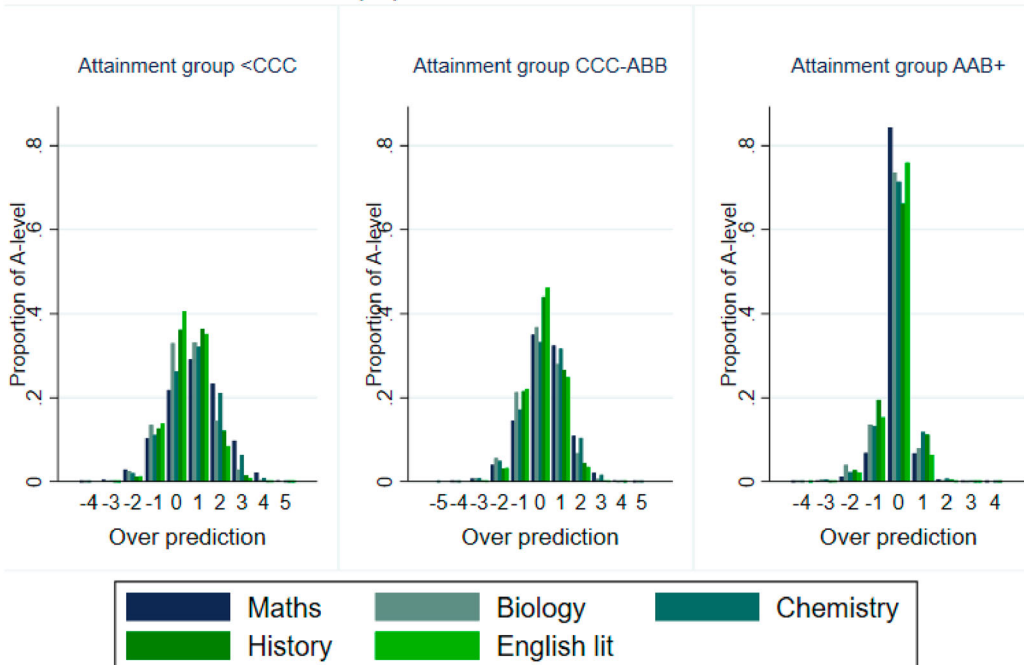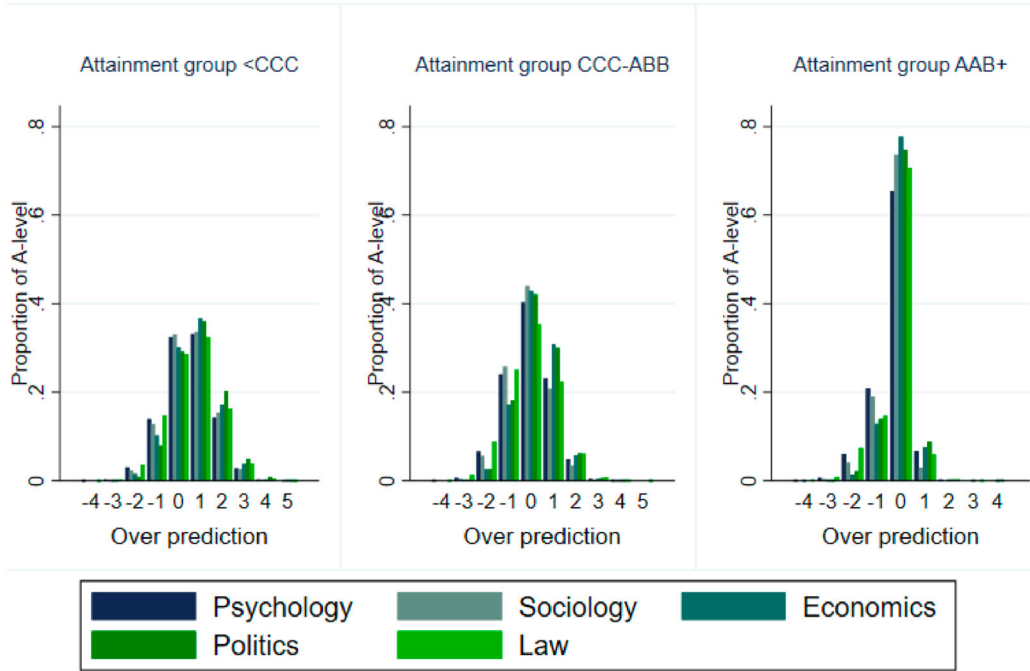
Figure 4. Distribution of over predictions for the five most popular A level subjects without related GCSEs – low achievers (<CCC), average achievers (CCC-ABB) and high achievers (AAB+).

that those A levels with direct prior achievement information are more accurately predicted. Here, there is a similar pattern to that seen for those subjects with a 'related' GCSE. Psychology and sociology, those with the lowest average points scores in Table 5, are more accurately predicted among low achievers, while economics and politics, with the highest average point scores are more accurately predicted among high achievers. Law is the least accurately predicted across the achievement distribution, and is also the lowest scoring subject in terms of average points.

8. Conclusion

While pupils' prior achievement is commonly used as an element of their school leaving grades, and to supplement admission decisions to university courses, unlike any other country, the UK uses predicted grades as a central feature of this process in advance of formal examinations. Research (Murphy and Wyness 2020; Campbell et al. 2022) suggests inaccuracies in predictions could result in students from poorer backgrounds applying to less selective courses than they might have done, had they known their true exam results upon application. In this paper, we ask whether the UK system can be improved upon, by using two different approaches to model predicted grades based on information from detailed administrative data including prior achievement, demographic information, and school-level data.

This question is important in and of itself, due to the decreasing value of examination grades caused by the Covid-19 pandemic. Many countries experienced widespread disruption to their education systems. This is likely to lead to universities (and employers) relying on previous (i.e. pre pandemic) formal examination results when deciding between applicants. If pupils' trajectories vary by key characteristics, then we might expect some groups' prior achievement to be less representative of where they are now.

Moreover, understanding whether empirical approaches to predicting grades can improve on teachers' performance is important in the context of the existing UK system. As described, students from low SES backgrounds have been shown to receive less generous predictions than their high SES counterparts. This has been shown to result in them applying to less prestigious institutions (Murphy and Wyness 2020). Campbell et al. (2022) show that low SES students are more likely to undermatch, enrolling in less academically selective courses than their richer counterparts.

Our models do improve the accuracy of teacher predictions, with just over 1 in 4 correctly predicted compared to their actual performance across their best three A levels using both our ordered probit and Random Forest approaches. This is an 11ppt improvement on teachers' predictions. Yet despite the comprehensive set of information available to us, 3 out of 4 pupils are still under- or over-predicted when using these approaches. There are also important differences across settings, showing that prediction accuracy varies depending on the group of interest. In particular, high achievers are more often correctly predicted due to ceiling effects, yet high achievers in non-selective state schools are 12ppts more likely to be under-predicted by 2 grades or more, relative to their high achieving counterparts in grammar or private schools. This is driven by low SES pupils within non-selective state schools who are disproportionately under-predicted among middle and high achievers, relative to their high SES counterparts. This highlights both the difficulty in predicting such crucial examination results, and important inequalities in these predictions.

There are also differences across subjects studied, with facilitating subjects easier to predict than other subjects, partly due to these subjects having 'related' GCSEs. While English literature is well-predicted across the range of achievement, maths and chemistry are harder to predict among low achievers, and more accurately predicted among high achievers. Among those A level subjects without a 'related' GCSEs, subjects studied more often at private schools, such as economics and politics, are more accurately predicted among high achievers, while subjects such as sociology and psychology, are more accurately predicted among low achievers. Law is hard to predict accurately across the distribution of achievement.

Taken together, this analysis has shown the difficulties in accurately predicting A level grades, regardless of the method used. Accuracy of predictions varies across levels of achievement, school type, by SES, and by subject studied. This raises some significant questions about why such predictions play such a prominent role in the UK's education system given the amount of inaccuracy found in measuring them, and the risk of exacerbating inequalities in life chances for young people in different settings. Our results also highlight concerning instances where pupils are 'hard to predict' and go on to over-perform at A level, given their GCSE results – most notably high achieving non-selective state school pupils. There is scope for future research to understand why such pupils outperform expectations. However, if ongoing exam disruption as a result of the Covid-19 pandemic leads further and higher education providers to put more weight on applicants' prior attainment (or even on teacher predictions) as a source of information on students ability, this could put certain groups at a disadvantage.

Notes

1. The UK's centralised applications system was established 60 years ago, and a legacy from the original paper-based system is that pupils apply to university long before they sit their exams – using predicted, rather than actual exam grades. Therefore, unlike anywhere else in the world, predicted grades are a fundamental part of determining access to university courses.
2. More details of this can be found in Murphy and Wyness (2020).
3. https://www.theguardian.com/education/2020/jun/24/top-public-school-asks-teachers-to-exaggerate-exam-predictions
4. Department for Education (2021)
5. We divide our sample asymmetrically by attainment, which increases the presence of ceiling effects relative to floor effects.

6. https://www.itv.com/news/2020-06-17/concerns-report-negative-impact-disadvantages-predicted-grades-students-gcse-alevel-coronavirus-uk-education
7. See https://www.gov.uk/government/news/ofqual-publishes-initial-decisions-on-gcse-and-a-level-grading-proposals-for-2020 for details.
8. This was the most recent wave of data available to us for this analysis. While participation increased for all groups over this period, the participation gap between high and low SES students remained relatively stable (Murphy, Scott-Clayton, and Wyness 2019). It is therefore not surprising that our results across school type and SES (see page 18) are similar to those of Murphy and Wyness (2020) who use later 2013–2015 cohorts. Note that Murphy and Wyness use data from the University and College Admissions Service (UCAS). The key differences are that their data are aggregate whereas ours is individual level, and their data are focused on university applicants, whereas ours is on enrolled students.
9. Further Education Colleges did not return the Spring Census in 2008 and so information on SES is missing for them for 73,666 individuals. On average, these pupils are likely to be from families with lower SES.
10. Key Stage 2 tests are not available for private school students and so our chosen model focuses on measures that are available for all students at Key Stage 4 and Key Stage 5. The inclusion of these scores makes very little difference to our models, improving the predictive value for our sub-sample of state-educated pupils by 0.2%. Results available from authors on request.
11. Around 700 households or 1,500 individuals
12. Jerrim (2020) compares this SES index to average family income across childhood using the Millennium Cohort Study (MCS) and has found this to be a promising proxy of childhood circumstance.
13. We choose these groupings in part to align with work by Murphy and Wyness (2020) but also because pupils with AAB or higher are considered to be particularly high achieving pupils (BIS 2011).
14. Note that our cohort predates the introduction of A* at A level, so we only predict between grades A-E and ungraded.
15. Bearing in mind the main purpose of our paper is to investigate how well we can predict actual outcomes, using detailed data on prior achievement from administrative data, the two models we use have both strengths and drawbacks. The ordered probit has the benefit of being simple to interpret and easy to estimate due to being less computationally intensive, but makes some assumptions such as the parallel regression assumption. The random forest approach is more computationally intensive, but on the other hand more flexible than the ordered probit.
16. Note there are two main differences between our approach and that of Murphy and Wyness which will have opposing effects on the accuracy of predictions. On the one hand, our cohort preceded the introduction of A* grades, meaning that our predictions are over 5 grades rather than 6 grades in Murphy and Wyness. On the other hand, our predictions are based on the more heterogeneous sample of all A level students while Murphy and Wyness are restricted to only those attending university.
17. This distribution is less skewed towards over-prediction than our 'best 3' headline findings, illustrating that part of the asymmetry is driven by this aggregation, given the correlation in A level grades within pupils. In addition, the other driver of this asymmetry is the greater proportion of low achievers who are mechanically more likely to be over-predicted.
18. Results by achievement groups available on request.
19. SES quintiles are scored 1 (lowest) to 5 (highest). 31,255 private school pupils in the full sample and 8,416 private school pupils in the restricted sample were added to the top quintile in the absence of SES data in KS5. This is why the overall mean is greater than 3 for both samples.
20. Mean score of GCSEs and equivalents, with an A* 58 QCA points, and each grade then 6 points lower.
21. Scored as for calculation of over prediction – 5 points for A, 0 points for ungraded.

Acknowledgements

Disclosure statement

ORCID

Gill Wyness 🔵 http://orcid.org/0000-0002-2920-6649
Lindsey Macmillan 🔵 http://orcid.org/0000-0003-1262-303X
Jake Anders 🔵 http://orcid.org/0000-0003-0930-2884
Catherine Dilnot 🔵 http://orcid.org/0000-0002-3952-347X

References

Azmat, G., and N. Iriberri. 2010. "The Importance of Relative Performance Feedback Information: Evidence from a Natural Experiment Using High School Students." *Journal of Public Economics* 94 (7-8): 435–452.

Bandiera, O., V. Larcinese, and I. Rasul. 2015. "Blissful Ignorance? A Natural Experiment on the Effect of Feedback on Students' Performance." *Labour Economics* 34: 13–25.

BIS. 2011. "Students at the Heart of the System", Higher Education White Paper. Department for Business, Innovation and Skills, London.

Breiman, L. 2001. "Random Forests." *Machine Learning* 45: 5–32. doi:10.1023/A:1010933404324.

Burgess, S., and E. Greaves. 2013. "Test Scores, Subjective Assessment, and Stereotyping of Ethnic Minorities." *Journal of Labor Economics* 31 (3): 535–576.

Burgess, S., D. S. Hauberg, B. S. Rangvid, and H. H. Sievertsen. 2022. "The Importance of External Assessments: High School Math and Gender Gaps in STEM Degrees." *Economics of Education Review* 88: 102267.

Campbell, S., L. Macmillan, R. Murphy, and G. Wyness. 2022 (forthcoming). "Matching in the Dark? Inequalities in Student to Degree Match." *Journal of Labor Economics* 40 (4).

Chowdry, H., C. Crawford, L. Dearden, A. Goodman, and A. Vignoles. 2013. "Widening Participation in Higher Education: Analysis Using Linked Administrative Data." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 176 (2): 431–457.

Crawford, C. 2014. "Socio-economic Differences in University Outcomes in the UK: Drop-out, Degree Completion and Degree Class" IFS Working Paper W14/31. IFS.

Crawford, C., L. Macmillan, and A. Vignoles. 2016. "When and Why Do Initially High Attaining Poor Children Fall Behind?" *Oxford Review of Education* 43 (1): 88–108.

Cunha, F., J. Heckman, L. Lochner, and D. Masterov. 2006. "Interpreting the Evidence on Life Cycle Skill Formation." In *Handbook of the Economics of Education*, Vol. 1, edited by E. Hanushek and F. Welch, 697–812. Amsterdam: North-Holland.

Delap, M. R. 1994. "An Investigation Into the Accuracy of A-Level Predicted Grades." *Educational Research* 36 (2): 135–148.

Demack, S., D. Drew, and M. Grimsley. 2000. "Minding the Gap: Ethnic, Gender and Social Class Differences in Achievement at 16, 1988–95." *Race Ethnicity Education* 3: 112–141.

Everett & Papageorgiou. 2011. "Investigating the Accuracy of Predicted A Level Grades as part of 2009 UCAS Admission Process".

Feinstein, L. 2003. "Inequality in the Early Cognitive Development of British Children in the 1970 Cohort." *Economica* 70: 73–97.

Gill and Benton. 2015. "The Accuracy of Forecast Grades for OCR A Levels in June 2014", Cambridge Assessment Statistics Report Series No.90, Cambridge Assessment, Cambridge, UK.

Hvidman, U., and H. H. Sievertsen. 2021. "High-stakes Grades and Student Behavior." *Journal of Human Resources* 56 (3): 821–849.

Jerrim, J. 2020. "Measuring Socio-economic Background Using Administrative Data. What is the Best Proxy Available?" Social Research Institute working paper. http://repec.ioe.ac.uk/REPEc/pdf/qsswp2009.pdf.

Jerrim, J., and A. Vignoles. 2013. "Social Mobility, Regression to the Mean and the Cognitive Development of High Ability Children from Disadvantaged Homes." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 176 (4): 887–906.

Lavy, V., and E. Sand. 2015. "On the Origins of Gender Human Capital Gaps: Short and Long term Consequences of Teachers' stereotypical biases." No. w20909. National Bureau of Economic Research.

Liaw, A., and M. Wiener. 2002. "Classification and Regression by RandomForest." *R News* 2 (3): 18–22.

Murphy, R., J. Scott-Clayton, and G. Wyness. 2019. "The End of Free College in England: Implications for Enrolments, Equity, and Quality." *Economics of Education Review* 71 (2019): 7–22.

Murphy, R., and F. Weinhardt. 2020. "Top of the Class: The Importance of Ordinal Rank." *The Review of Economic Studies* 87 (6): 2777–2826.

Murphy, R., and G. Wyness. 2020. "Minority Report: The Impact of Predicted Grades on University Admissions of Disadvantaged Groups." *Education Economics* 28 (4): 1–18.

Nye, P., and D. Thompson. 2020. "GCSE and A-Level Results 2020: How Grades Have Changed in Every Subject". FFT Education Datalab blog. Fisher Family Trust. Last accessed 30/06/2022. https://ffteducationdatalab.org.uk/2020/08/gcse-and-a-level-results-2020-how-grades-have-changed-in-every-subject/.

Oppos, D. 2020. "The Impact of the Coronavirus Outbreak on Exams Around the World", The Ofqual blog. Office of Qualifications and Examinations Regulation, London. https://ofqual.blog.gov.uk/2020/05/22/the-impact-of-the-coronavirus-outbreak-on-exams-around-the-world/.

De Paola, M., and V. Scoppa. 2017. "Gender Differences in Reaction to Psychological Pressure: Evidence from Tennis Players." *European Journal of Work and Organizational Psychology* 26 (3): 444–456.

Paserman, M. D. 2007. "Gender Differences in Performance in Competitive Environments: Evidence from Professional Tennis Players." *IZA Discussion Paper series*, June 2007 (No 2834).

UCAS. 2016. "Factors Associated with Predicted and Achieved A Level Attainment", University and College Admissions Service, Gloucestershire.

UNESCO. 2020. "COVID-19. A Glance of National Coping Strategies on High Stakes Examinations and Assessments". Last accessed 30/06/2022. https://en.unesco.org/sites/default/files/unesco_review_of_high-stakes_exams_and_assessments_during_covid-19_en.pdf. United Nations Educational, Scientific and Cultural Organization, Paris France.

Appendix

Table A1. GCSEs entered and predictors of key A level grades.

GCSE predictor variables	Proportion entered	A*	A	B	C	Fail	Not entered	Significant predictor of A-level grade – all gcse grades significantly different from A*			
								Maths A-level	English Lit A-level	Biology A-level	Psychology A-level
GCSE total points	100.0%							***	***		***
GCSE total points squared	100.0%							***	***	*	***
English literature	0.0%	10%	28%	34%	20%	5%	3%	***	***	***	***
English	0.0%	10%	27%	34%	19%	2%	8%		***	***	***
Maths (see note below)	0.0%	11%	22%	31%	21%	6%	9%	***	***	***	***
Double award science	0.0%	10%	15%	20%	23%	7%	25%	***	***	***	***
French	0.0%	8%	11%	13%	13%	7%	48%	***	***	***	***
History	0.0%	9%	15%	13%	8%	4%	51%	***	***	***	***
Geography	0.0%	7%	11%	11%	8%	3%	60%	***	***	***	***
Religious studies (short)	0.0%	5%	9%	10%	8%	5%	63%		***	***	***
Art and design	0.0%	5%	9%	9%	7%	2%	69%		***	***	***
Religious studies (full)	0.0%	6%	10%	9%	5%	2%	69%		***	***	***
Physical education	0.0%	3%	6%	6%	4%	3%	77%			***	***
German	0.0%	3%	5%	6%	6%	3%	78%	***	***	***	***
Information technology (short)	0.0%	1%	3%	5%	5%	5%	82%		***	***	***
Biology	0.0%	4%	6%	5%	3%	1%	82%	***	***	***	***
Chemistry	0.0%	4%	5%	5%	3%	1%	83%	***		***	***
Drama	0.0%	2%	5%	6%	3%	1%	83%		***	***	***
Physics	0.0%	4%	5%	4%	3%	1%	83%	***	***	***	***
Information technology (full)	0.0%	2%	4%	5%	4%	2%	83%	***		***	
Business	0.0%	1%	3%	4%	4%	2%	85%		***	***	***
Design & Tech (graphics)	0.0%	1%	3%	4%	3%	2%	86%				
Spanish	0.0%	3%	3%	3%	3%	2%	86%	***	***	***	
Statistics	0.0%	1%	3%	3%	4%	1%	87%	***		***	
Music	0.0%	2%	4%	4%	2%	1%	87%	***		***	
Design & tech (food tech)	0.0%	1%	4%	3%	3%	1%	88%			***	***
Design & tech (resistant materials)	0.0%	1%	3%	3%	3%	1%	89%			***	
Fine art	0.0%	1%	3%	2%	2%	0%	91%		***		
Design & tech (textiles tech)	0.0%	1%	3%	2%	1%	1%	92%				

Table A1. Continued.

GCSE predictor variables	Proportion entered	A*	A	B	C	Fail	Not entered	Significant predictor of A-level grade – all gcse grades significantly different from A* English			
								Maths A-level	Lit A-level	Biology A-level	Psychology A-level
Social science citizenship	0.0%	0%	2%	2%	2%	1%	93%				
Media, film & television	0.0%	1%	2%	2%	2%	0%	93%		***		***
Applied ICT	0.0%	0%	1%	2%	2%	2%	94%				
Office technology	0.0%	1%	1%	1%	1%	0%	95%				
Single award science	0.0%	0%	0%	1%	1%	1%	97%	***	***		
Design & tech (electronic products)	0.0%	0%	1%	1%	1%	0%	97%			***	***
Applied business	0.0%	0%	1%	1%	1%	0%	97%				
Home economics: child development	0.0%	0%	1%	1%	1%	0%	98%				
Design & tech (systems and control)	0.0%	0%	1%	1%	0%	0%	98%	***		**	
Health and social care	0.0%	0%	0%	1%	0%	0%	98%				***
Applied science	0.0%	0%	0%	0%	1%	0%	99%				
Italian	0.0%	0%	0%	0%	0%	0%	99%				
Latin	0.0%	0%	0%	0%	0%	0%	99%				
Chinese	0.0%	1%	0%	0%	0%	0%	99%				
Urdu	0.0%	0%	0%	0%	0%	0%	99%				
Russian	0.0%	0%	0%	0%	0%	0%	100%				
Gujurati	0.0%	0%	0%	0%	0%	0%	100%				
Arabic	0.0%	0%	0%	0%	0%	0%	100%				
Japanese	0.0%	0%	0%	0%	0%	0%	100%				
Panjabi	0.0%	0%	0%	0%	0%	0%	100%				
Bengali	0.0%	0%	0%	0%	0%	0%	100%				
Modern Greek	0.0%	0%	0%	0%	0%	0%	100%	**			
Turkish	0.0%	0%	0%	0%	SUP	SUP	100%				
Modern Hebrew	0.0%	0%	0%	0%	SUP	SUP	100%				
Portuguese	0.0%	0%	0%	0%	SUP	SUP	100%				
Polish	0.0%	0%	0%	0%	SUP	SUP	100%				
Persian	0.0%	0%	0%	0%	SUP	SUP	100%				
Dutch	0.0%	0%	0%	0%	SUP	SUP	100%				

Table A2. Distribution of over prediction by pupil of best three A level grades excluding those with fewer than three A levels, by A level attainment group.

Total over prediction	Total	<CCC	CCC-ABB	AAB+
	%	%	%	%
−8	0.1	0	0	0.1
−7	0.1	0	0.1	0.2
−6	0.4	0	0.4	0.6
−5	0.9	0.1	1.2	1.3
−4	2.2	0.3	3.0	2.5
−3	4.6	1.2	6.3	4.6
−2	8.1	3.5	10.4	7.9
−1	12.5	7.8	15.5	11.3
0	26.6	13.3	18.1	54.7
1	17.3	18.5	17.1	16.6
2	12.2	19.0	15.4	0

Table A2. Continued.

Total over prediction	Total	<CCC	CCC-ABB	AAB+
3	7.5	14.9	7.8	0
4	4.0	9.9	3.2	0
5	2.0	6.0	1.1	0
6	0.9	3.0	0.3	0
7	0.4	1.6	0	0
8	0.1	0.5	0	0
9	0.1	0.2	0	0
10	0	0.1	0	0
11	0	0	0	0
Total number pupils	167,937	40,333	83,288	44,310

Notes: Full sample of all students achieving at least three A levels. Grade boundaries are chosen to replicate Murphy and Wyness (2020). Low achievers with grades below 3 Cs, middle achievers with grades from CCC to ABB, and high achievers from AAB to A*A*A.

Table A3. Distribution of over-prediction by pupil of best three A level grades, with random effects and demographic controls for full and restricted samples (sample excluding those at private schools or with a missing SES quintile run for main specification, and M2, which includes demographics).

Total over prediction	Main specification	Main specification restricted	M2	M2 restricted	M3	M4
	%	%	%	%	%	%
−8	0	0	0	0	0.1	0.1
−7	0.1	0.1	0.1	0.1	0.2	0.3
−6	0.3	0.2	0.3	0.2	0.5	0.7
−5	0.7	0.7	0.7	0.6	1.0	1.5
−4	1.7	1.7	1.7	1.7	2.1	2.9
−3	3.8	3.9	3.8	3.9	4.1	5.0
−2	7.6	8.0	7.7	8.0	7.1	8.2
−1	13.8	14.5	13.9	14.7	11.1	12
0	27.2	27.1	27.5	27.2	21.8	21.7
1	19.5	19.6	19.5	19.7	16.9	16.3
2	12.4	12.3	12.3	12.2	12.9	12.0
3	6.8	6.5	6.7	6.4	8.8	7.8
34	3.4	3.1	3.3	3.1	5.6	5.0
5	1.6	1.3	1.5	1.3	3.4	2.9
6	0.7	0.6	0.6	0.5	2.0	1.7
7	0.3	0.2	0.3	0.2	1.1	1.0
8	0.1	0.1	0.1	0.1	0.6	0.5
9	0	0	0	0	0.3	0.2
10	0	0	0	0	0.2	0.1
11	0	0	0	0	0.1	0.1
Number of pupils	238,898	133,986	238,898	133,986	238,898	238,898
Total GCSE points and points squared	x	x	x	x	x	x
GCSE grades in all subjects	x	x	x	x		
GCSE entry flags in all subjects					x	x
Gender, ethnicity, SES quintile, school type			x	x	x	x
School random effects						x