CENTRE FOR EDUCATION POLICY
& EQUALISING OPPORTUNITIES

Grade Expectations: How well can we predict future grades based on past performance?

Centre for Education Policy and Equalising Opportunities (CEPEO)

Jake Anders, Catherine Dilnot, Lindsey Macmillan, and Gill Wyness

Working Paper No. 20-14

August 2020

Disclaimer

Any opinions expressed here are those of the author(s) and not those of the UCL Institute of Education. Research published in this series may include views on policy, but the institute itself takes no institutional policy positions.

CEPEO Working Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

Highlights

- Predicted grades are a common feature of the English education system, with teachers' predictions of pupils' A level performance forming the basis of university applications each year. Yet previous work has shown that these predictions are highly inaccurate.
- The recent Covid-19 pandemic has put these predictions under the spotlight, with the cancellation of exams meaning that all year 11 and year 13 pupils will instead receive 'calculated grades' based on teacher predictions.
- We ask whether an alternative approach to predicting A level grades, using statistical and machine learning methods based on pupil's prior achievement, can improve the accuracy of predictions.
- Using a wealth of administrative data, we can make only modest improvements on teacher predictions. Our models can correctly predict 1 in 4 pupils across their best three A levels, versus 1 in 5 for teacher predictions. The predictions generated by our models are incorrect for 74% of pupils.
- High achieving pupils in comprehensive schools are more likely to be under-predicted by our models, compared to their grammar and private school counterparts. This highlights the difficult task that teachers face each year, particularly for pupils with more variable trajectories from GCSE to A level.
- The fact that even with advanced statistical techniques, and rich achievement data, our models still generate low rates of prediction accuracy, with varying rates of accuracy across pupil achievement, school type and subjects, raises the question as to why predicted grades continue to form such a crucial part of our education system.

Why does this matter?

Predicted grades form the basis of university applications in England, determining the life chances of pupils in post-secondary education. Yet predicting grades accurately is very difficult, both for teachers and when using statistical and machine learning approaches. This raises the question as to why they play such a prominent role in our system.

Grade Expectations: How well can we predict future grades based on past performance?

Jake Anders*, Catherine Dilnot**, Lindsey Macmillan*, and Gill Wyness*

August 2020

*UCL Centre for Education Policy and Equalising Opportunities

**Oxford Brookes Business School

Abstract

The Covid-19 pandemic has led to unprecedented disruption of England's education system, including the cancellation of all formal examination. Instead of sitting exams, the class of 2020 will be assigned "calculated grades" based on predictions by their teachers. However, teacher predictions of pupil grades are a common feature of the English education system, with such predictions forming the basis of university applications in normal years. But previous research has shown these predictions are highly inaccurate, creating concern for teachers, pupils and parents. In this paper, we ask whether it is possible to improve on teachers' predictions, using detailed measures of pupils' past performance and non-linear and machine learning approaches. Despite lacking their informal knowledge, we can make modest improvements on the accuracy of teacher predictions with our models, with around 1 in 4 pupils being correctly predicted. We show that predictions are improved where we have information on 'related' GCSEs. We also find heterogeneity in the ability to predict successfully, according to student achievement, school type and subject of study. Notably, high achieving non-selective state school pupils are more likely to be under-predicted compared to their selective state and private school counterparts. Overall, the low rates of prediction, regardless of the approach taken, raises the question as to why predicted grades form such a crucial part of our education system.

Acknowledgements

The authors would like to thank Richard Murphy and John Jerrim for comments.

1. Introduction

The Covid-19 pandemic has led to unprecedented disruption in the education system in the UK, including the cancellation of all formal examination. Instead of sitting exams, the class of 2020 will be assigned “calculated grades” based on predictions by their teachers, with school-level moderation by the exam regulator. Importantly, this system of predicting grades is not unprecedented in the UK. In fact, teacher-predicted grades are a regular annual feature of the English education system for those applying to university, a legacy from the paper-based applications of our centralised system, meaning that pupils apply to university long before they sit their exams. Therefore, unlike anywhere else in the world, predicted grades are a fundamental part of determining access to university courses, and wider life chances of pupils in post-secondary education. Yet previous research (Delap, 1994; 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) has shown that only 16% of university applicants were correctly predicted across their best three A levels when comparing teacher’s predictions to university applicants’ actual grades achieved.

In this paper, we ask, given the centrality of predicted grades in our system, whether there is any way to improve the accuracy of these age 18 grade predictions, by instead predicting pupils’ grades based on their past performance using either non-linear regression modelling or a highly-flexible machine learning approach. This is an empirical question. On the one hand, the use of models based on large-scale data may avoid issues of teacher bias and/or manipulation of grades. [1] 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 empirical approaches to predicting grades can improve on teachers’ performance is important. If pupils’ grades can be more accurately predicting using their prior test scores, then this may be a preferable alternative (or, at least, a supplement) to teacher

[1] https://www.theguardian.com/education/2020/jun/24/top-public-school-asks-teachers-to-exaggerate-exam-predictions

estimation 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 rich set of detailed prior attainment results and pupil characteristics, pupil grades cannot be predicted accurately, then this highlights the difficulty faced by teachers, and provides further evidence that the UK’s predicted grades system should be re-examined.

A small number of studies have examined the accuracy of teacher predictions using data on students’ predicted and actual exam grades. These studies generally point to a high degree of inaccuracy in predicted grades, and typically find that teachers tend to err on the side of optimism in their predictions. Delap (1994) and Everett and Papageourgiou (2011) analyse prediction accuracy by individual subject, both showing around half of all predicted grades were accurate, while 42-44% were over-predicted by at least one grade, and only 7-11% of all predicted grades were under-predicted. More relevant to this paper, studies by UCAS (2016) and Wyness and Murphy (2020) examine prediction accuracy according a students’ best 3 A levels, with the latter finding that only 16% of students received accurate predictions, with 75% overpredicted and just 8% underpredicted.

All studies also found that higher grades tend to be more accurately predicted than lower grades (though this is likely mechanical: teachers tend to overpredict, and this is impossible for the top grades, so the default will be towards accuracy). This highlights the importance of examining prediction accuracy within pupil achievement level. While there was little robust evidence of any systematic bias in teacher predictions according to pupil characteristics, Wyness and Murphy (2020) do find that among high-achievers, low SES pupils were under-predicted across schools.

Our results are also relevant to the literature on teacher bias. Work by Burgess and Greaves (2013) examined 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 were more likely to be under-predicted, 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. Diamond and Persson (2016) show that pupils in Sweden are much less likely to be marked just below a grade threshold (resulting in positive signalling effect in the labour market), implying teacher manipulation may be present in some settings.

We find that using information on previous achievement in exams at age 16 leads to only a marginal improvement on teacher predictions, 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 a wealth of 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 differences in how well we can correctly predict pupil grades, in terms of pupil attainment, 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.[2] Importantly, when we look 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. While our data do not provide any insights as to why this might be the case, it indicates 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, highlighting the difficulty of the task that teachers are faced with in predicting grades.

We also observe 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

[2] We divide our sample asymmetrically by attainment, which increases the presence of ceiling effects relative to floor effects.

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.

In summary, our findings imply that even with detailed information on pupil prior attainment and demographics, predicting their future outcomes is a very challenging task. This raises the question of why we continue to define such a crucial stage of our education system using predictions.

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 disruption to the 2020 examination system

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 Wyness and Murphy, 2020 for a detailed outline of this process).

This year, these predicted grades have been thrust into the limelight due to the cancellation of formal examinations caused by the Covid-19 pandemic. Teachers must provide new predictions for the grades of every pupil due to take GCSEs and A levels in summer 2020, and also rank each pupil within year group and subject. While the exam regulator, the Office of Qualifications and Examinations Regulation (Ofqual) will moderate grades at the school level,[3]

[3] See https://www.gov.uk/government/news/ofqual-publishes-initial-decisions-on-gcse-and-a-level-grading-proposals-for-2020 for details.

these individual rankings will remain intact, making teacher predictions a fundamental element of pupil exam grades for 2020.

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). 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.

Descriptive statistics are presented in Table 1. 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 in three groups: low achievers (below CCC), middle achievers (CCC to ABB) and high achievers (AAB or above[4]). Of these pupils, 45% achieved below CCC grades, 36% achieved between CCC and AAB, and 19% achieved higher than AAB. Table 1 illustrates that we have a higher proportion of female pupils (54%) than males, with females typically outperforming males.

We also present our analysis across 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 present our analysis for two samples: a) all pupils and b) a restricted sample of pupils who take at least three A levels in subjects where they have taken a 'related' GCSE. Appendix Table A1 lists these 37 A levels and the 'related' GCSE subjects. Many are straightforward with the same subject being studied at GCSE and A level, such as maths, history, geography, modern

[4] 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).

languages, electronics, physical education (PE), and design technology (DT). For science A levels, we include both the separate award at GCSE (triple science) or the double award where taken, since this involves taking separate papers in each of the three sciences, but not the single award. This restricted group of pupils 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 teaching. We can then only compare within achievement groups across samples.

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 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.[5] We have also tested whether the inclusion of primary school achievement (Key Stage 2 test scores) improves our model further and find this adds no precision above our main model.

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[6]) 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

[5] 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.

[6] Around 700 households or 1,500 individuals

(as in Crawford, 2014).[7] Table 1 shows that higher SES pupils outperform lower SES pupils. The average GCSE scores of high A level achievers is higher than that of low achievers.

4. Methods

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 to 5 representing A level grades achieved.[8] 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_k to τ_{k-1} subject to the following relationship:

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

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

$$y_i^* = \sum_{j=2}^{6} \beta_1^j I(GCSEsubj1_i = j) + \sum_{j=2}^{6} \beta_2^j I(GCSEsubj2_i = j) \dots. + \sum_{j=2}^{6} \beta_s^j I(GCSEsubjs_i = j) + \beta_{s+1}\ GCSEpts_i + \beta_{s+2}\ GCSEpts_i^2 + \varepsilon_i \qquad (1)$$

Where $I(GCSEsubjk_i = j)$is the achieved grade j, by pupil i, in GCSE subject s, 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.

[7] 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.

[8] Note that our cohort predates the introduction of A* at A level, so we only predict between grades A-E and 'ungraded'.

We have tested various alternative models in order to try to improve on the overall predictive accuracy of the models, which we report in the Appendix. 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. [9] 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:

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

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 of 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

[9] Controlling for performance at age 11 for state school pupils (for whom such data are available – age 11 performance is not available for pupils at private schools) also made no substantive difference to the proportion of correct predictions. Results available from the author on request.

full cohort of pupils, and a second more restricted sample of those pupils who take ‘related’ GCSE subjects.

Random Forests

In our second approach, we estimate the predictability of A level grades by employing the supervised machine learning algorithm of Random Forests (Brieman, 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, 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 2,000) each on a bootstrapped sample of the dataset. At each step of the decision tree a random sub-set 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 & 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).

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 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).[10]

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.[11] Appendix Table A2 shows a very similar distribution to that found in Table 2 if we restrict the sample to only those taking 3 A levels, suggesting that including those with reduced risks of misclassification (those with fewer A levels) are not driving our findings.

Columns 2 to 4 of Table 2, and Figures 2-4, 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

[10] 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.

[11] 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.

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 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.

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. 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).[12]

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 for 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 are more likely to be over-predicted, with 76% over-predicted compared to 57% for the full sample.

[12] Results by achievement groups available on request.

Table 3 repeats the results in Table 2, but using 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 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.

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. 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-achieving less affluent (or non-selective state) pupils are harder to predict, regardless of the approach.

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.

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.

Across A level Subjects

Are certain A level subjects easier to predict than others? We explore this question for the for the top 5 most studied A level subjects with and without a 'related' GCSE. Table 5 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.

Figures 5-7 show 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 (Figure 7), 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 (Figure 7). 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, 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 (Figures 5 and 6), 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).

Figures 8-10 shows the distribution of predictions among the 5 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 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 (Figure 8), while economics and politics, with the highest average point scores are more accurately predicted among high achievers (Figure 10). Law is the least accurately predicted across the achievement distribution, and is also the lowest scoring subject in terms of average points.

6. Conclusion

Unlike any other country, predicted grades are a common feature of the UK education system in 'normal' times, being used to determine offers from university courses in advance of formal examinations. This year, teacher predictions have become even more high stakes due to the cancellation of examinations caused by the Covid-19 pandemic. Teacher predicted grades will now replace exam results for both GCSEs and A levels in summer 2020, and while Ofqual will moderate grades at the school level, these individual rankings will remain intact. This huge task for teachers has been shown in the past to be very difficult, with only 16% of pupils being accurately predicted in their final grade outcomes across 3 A levels (Murphy and Wyness, 2020). In this paper, we ask whether this 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.

Our models 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 wealth 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 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, and 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 to 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.

References

BIS (2011), “Students at the heart of the system”, Higher Education White Paper. Department for Business, Innovation and Skills, London

Burgess, S, & Greaves, E (2013) "Test scores, subjective assessment, and stereotyping of ethnic minorities." *Journal of Labor Economics* 31, no. 3, 535-576.

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

Chowdry, H., Crawford, C.,Dearden, L., Goodman, A., & Vignoles, A. (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

Delap, M. R. (1994). An investigation into the accuracy of A-level predicted grades. *Educational Research*, *36*(2), 135-148.

Diamond, R., & Persson, P. (2016). “The long-term consequences of teacher discretion in grading of high-stakes tests” NBER Working Paper No. w22207. National Bureau of Economic Research.

Everett & Papageorgiou (2011), “Investigating the Accuracy of Predicted A Level Grades as part of 2009 UCAS Admission Process”

Lavy, V., & Sand, E. (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 Wiener, M. (2002). Classification and Regression by randomForest. R News 2(3), 18--22.

Murphy, R., & Wyness, G. (2020). “Minority Report: the impact of predicted grades on university admissions of disadvantaged groups”. *Education Economics*, 1-18.

UCAS (2015). “Factors associated with predicted and achieved A level attainment”, University and College Admissions Service, Gloucestershire

Figure 1 Distribution of over predictions by individual A level – full sample

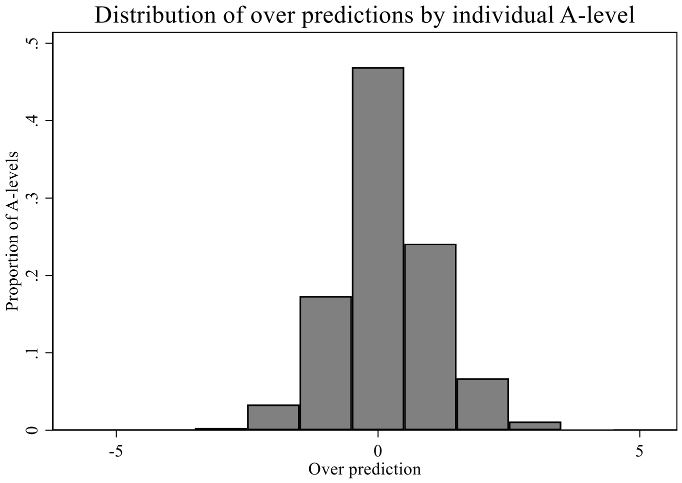


Figure 2 Distribution of total over-prediction by pupil: low achievers (<CCC)

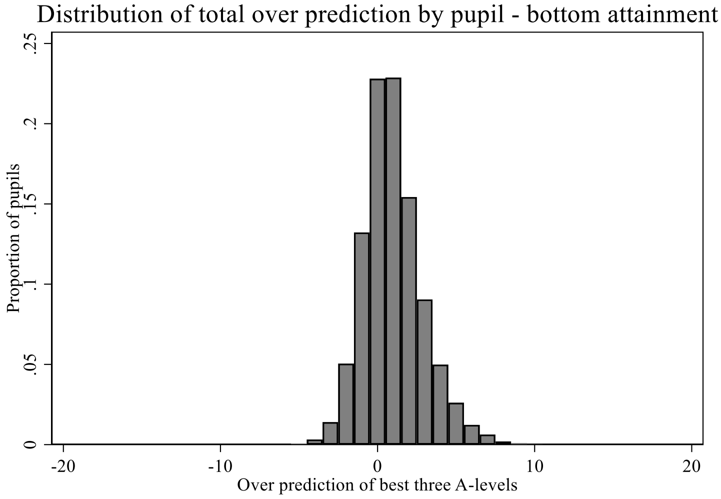


Figure 3 Distribution of total over-prediction by pupil: average achievers (CCC-ABB)

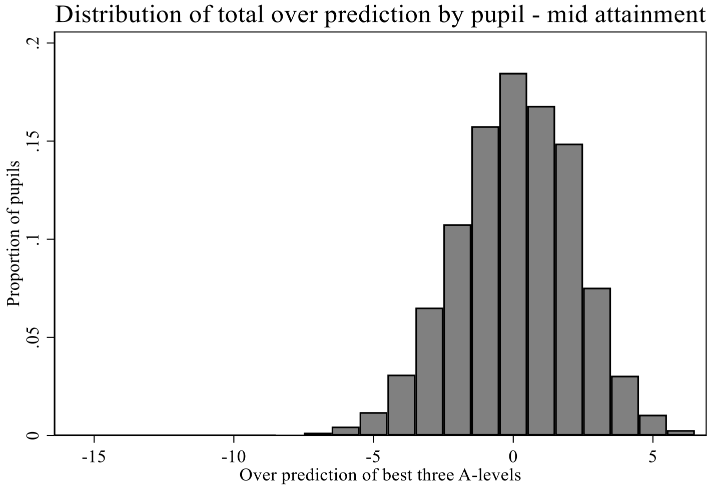


Figure 4 Distribution of total over-prediction by pupil: high achievers (AAB+)

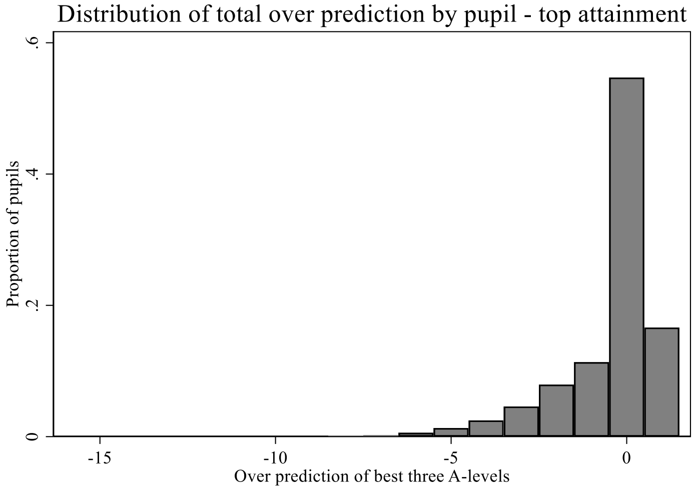
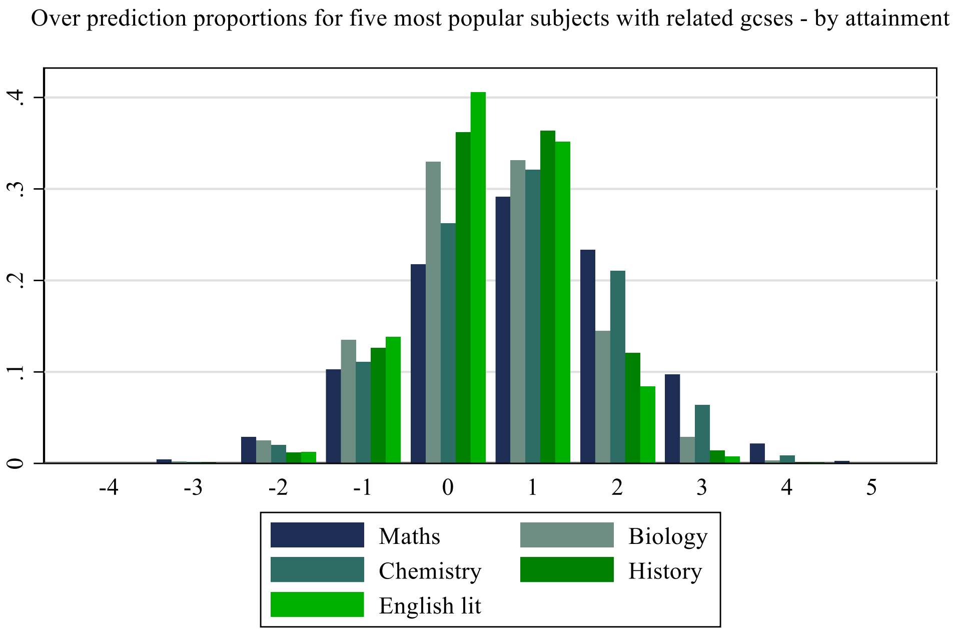
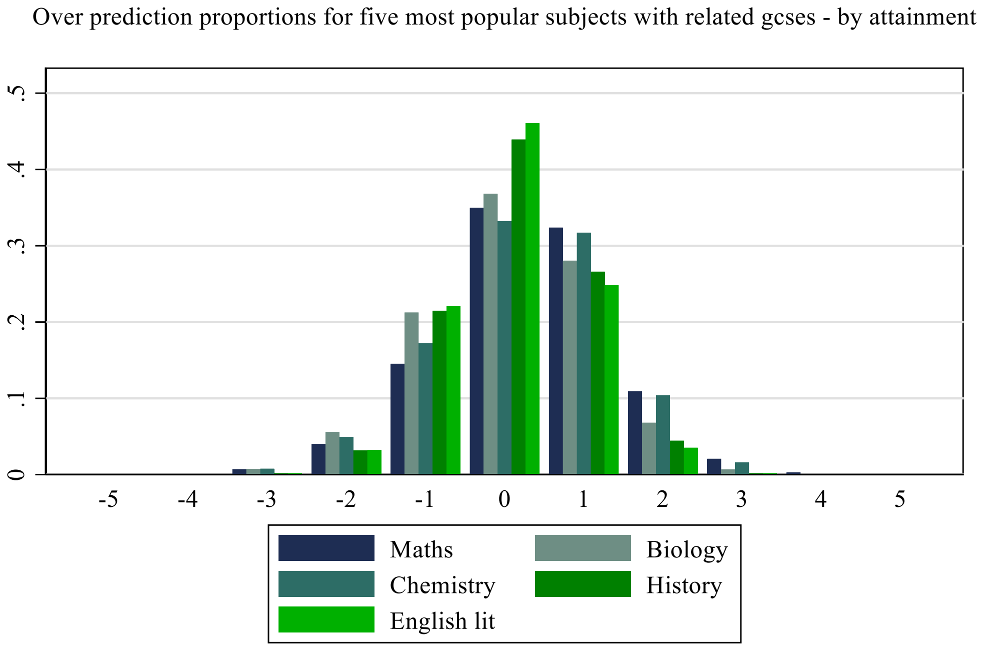


Figure 5 Distribution of over predictions for the five most popular A level subjects with related GCSEs– low achievers (<CCC)



Over prediction proportions for five most popular subjects with related gcses - by attainment

Figure 6 Distribution of over predictions for the five most popular A level subjects with related GCSEs– average achievers (CCC-ABB)



Over prediction proportions for five most popular subjects with related gcses - by attainment

Figure 7 Distribution of over predictions for the five most popular A level subjects with related GCSEs–high achievers (AAB+)

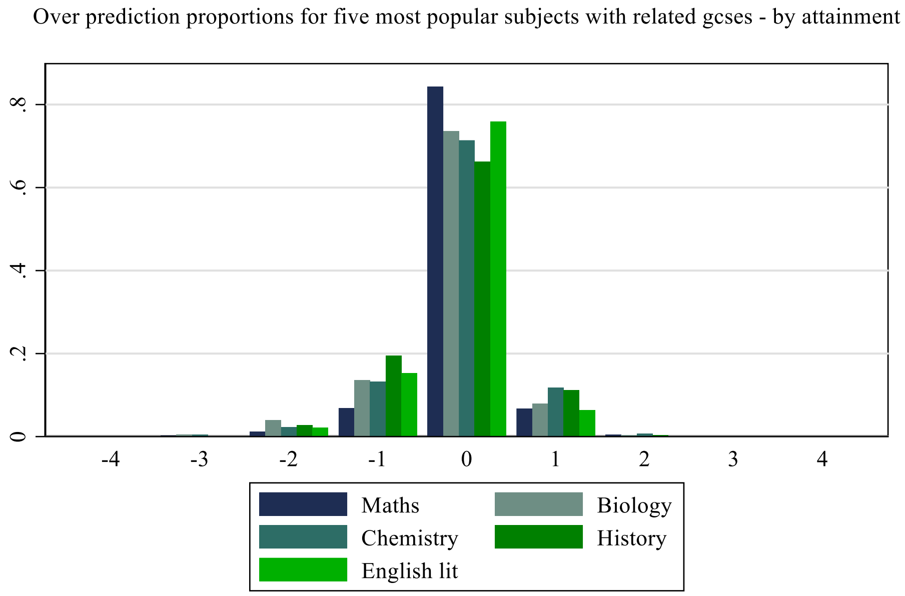


Figure 8 Distribution of over predictions for the five most popular A level subjects without related GCSEs– low achievers (<CCC)

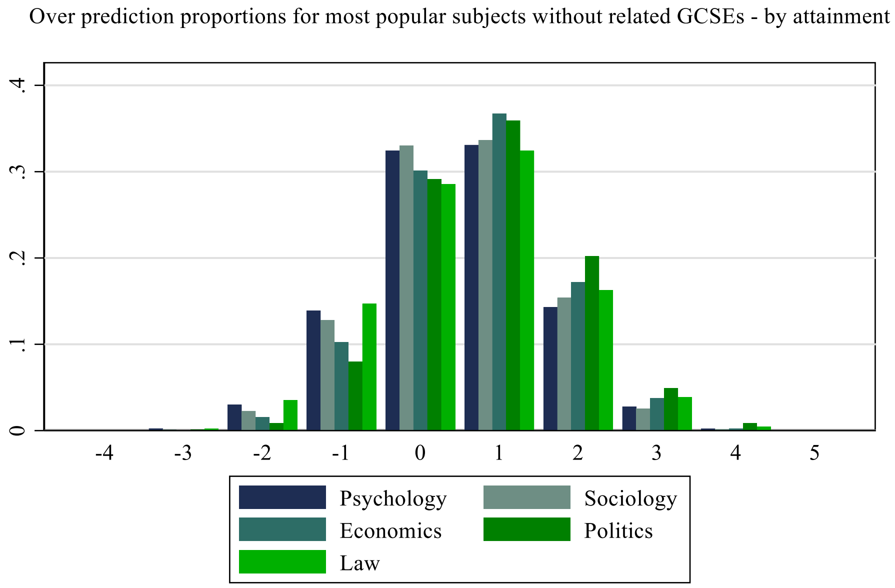


Figure 9 Distribution of over predictions for the five most popular A level subjects without related GCSEs– average achievers (CCC-ABB)

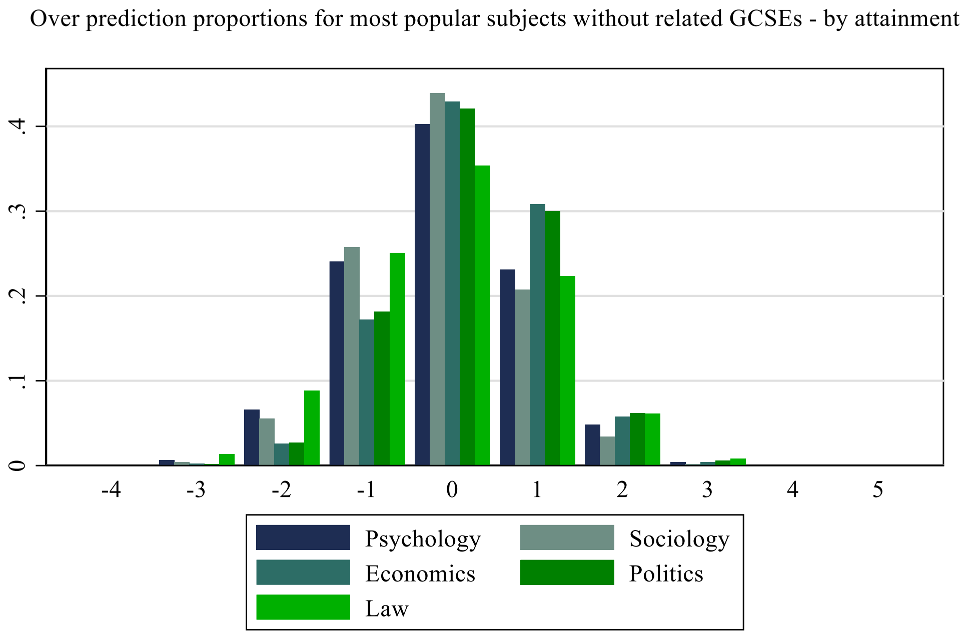


Figure 10 Distribution of over predictions for the five most popular A level subjects without related GCSEs– high achievers (AAB+)

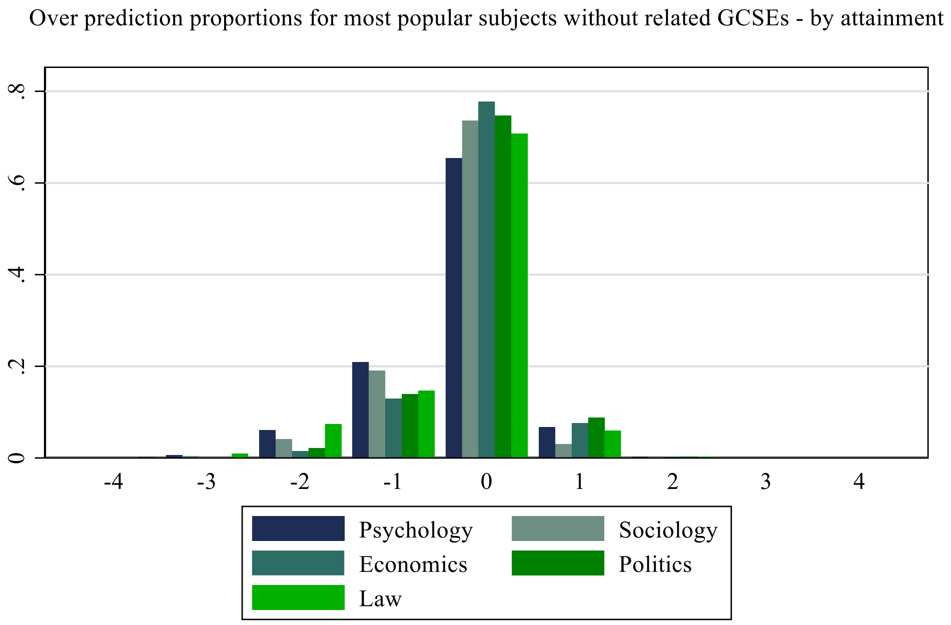


Table 1 Descriptive statistics for full sample (all those with at least one counting A level in the data) and for the restricted sample of all those with at least three A levels who had done the related GCSE

	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[13]	3.38	2.91	3.53	4.05	3.61	3.1	3.52	3.97
Attainment								
GCSE point score[14]	490	450	503	564	533	476	520	581
Points from best three A levels[15]	8.9	4.9	11.0	14.6	11.5	6.4	11.2	14.7

[13] 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. pupils

[14] Mean score of GCSEs and equivalents, with an A* 58 QCA points, and each grade then 6 points lower.

[15] Scored as for calculation of over prediction – 5 points for A, 0 points for ungraded.

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

	full sample				restricted sample			
Total over prediction	**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	8,900	22,623	16,941

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

	full sample				restricted sample			
Total over prediction	**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	8,900	22,623	16,941

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	4,113	5,158	64,904	8,959	12,579	23,561	7,237	13,512

Table 5: 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

A1: A levels with ‘related’ GCSEs

A level	GCSE1	GCSE2	GCSE3
A & D textiles	DT: Textiles tech		
A&D – all endorsements	DT: Graphic prods	Art & design	Fine art
Ancient Greek[16]	Classical Greek		
Biblical Hebrew	Biblical Hebrew		
Biology	Biology	Science double award	
Business	Bbusiness studies	Voc Business	
Chemistry	Chemistry	Science double	
Drama	Drama		
DT (food)	DT: Food tech		
DT control systems	DT: Systems & controls		
DT resistant mats	DT: Resistant mats		
DT Textiles	DT: Textiles tech		
Electronics	DT: Electronic prods		
English (all)	English lit	English	English lang
French	French		
Further maths	Maths		
Geography	Geography		
German	German		
History	History		
IT	Info tech	Info tech (short)	
Latin[17]	Latin		
Maths	Maths		
Media	Media, film, tv		
Music	Music		
PE	Physical education		
Physics	Physics	Science double award	
RS	Religious studies (full)	Religious studies (short)	
Science	Science double		
Spanish	Spanish		
Statistics	Statistics		

[16] There were no entries in dataset for classical Greek GCSE – so included in no related GCSE sample

[17] Some latin GCSE scores were problematic in the original data (wrongly recorded as double award scores). Therefore not used and Latin was included in no related GCSE sample

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
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

A3: Distribution of over-prediction by pupil of best three A level grades, with random effects and demographic controls for full and restricted samples

Total over prediction	**Main specification**	**M2**	**M3**	**M4**
	%	%	%	%
-8	0	0	0.1	0.1
-7	0.1	0.1	0.2	0.3
-6	0.3	0.3	0.5	0.7
-5	0.7	0.7	1.0	1.5
-4	1.7	1.7	2.1	2.9
-3	3.8	3.8	4.1	5.0
-2	7.6	7.7	7.1	8.2
-1	13.8	13.9	11.1	12
0	27.2	27.5	21.8	21.7
1	19.5	19.5	16.9	16.3
2	12.4	12.3	12.9	12.0
3	6.8	6.7	8.8	7.8
4	3.4	3.3	5.6	5.0
5	1.6	1.5	3.4	2.9
6	0.7	0.6	2.0	1.7
7	0.3	0.3	1.1	1.0
8	0.1	0.1	0.6	0.5
9	0	0	0.3	0.2
10	0	0	0.2	0.1
11	0	0	0.1	0.1
Number of pupils	238,898	238,898	238,898	238,898
Total GCSE points and points squared	x	x	x	x
GCSE grades in all subjects	x	x		
GCSE entry flags in all subjects			x	x
Gender, ethnicity, SES quintile, school type		x	x	x

@cepeo_ucl

ucl.ac.uk/ioe/cepeo