A case study of using AI for General Certificate of Secondary Education (GCSE) grade prediction in a selective independent school in England

Gyorgy Denes[a,*]

[a] *The Perse School, Cambridge, United Kingdom*

Abstract

The COVID-19 pandemic has created significant challenges for UK schools, but a time of cancelled exams and uncertainty around future examinations can provide opportunities to explore novel assessment methods. Hence, the 2020 proposal of the Ofqual algorithm which combines teachers' estimated grades and schools' historical performance seemed timely. However, the algorithmically calculated grades resulted in a public backlash and withdrawal of the proposal. While the failed Ofqual algorithm could be considered an example of AI, we do not yet have a thorough understanding of its numerical accuracy and how it performs in comparison to other AI models. This paper investigates this novel application: the potential use of a range of AI models as assessment tools in a selective, independent, secondary school in England. The following questions were examined: (1) how accurate are modern AI models in predicting GCSE exam grades? (2) what are the differences in model accuracy across subjects and can these be explained by qualitative differences in teachers' grading practices? Results indicate that while models yield acceptable mean absolute errors, individual mispredictions can be larger than desired. Subject differences highlighted that grading subjectivity is less significant in science, technology, engineering, and maths (STEM) subjects, which could explain why objective models fail to predict non-STEM grades more frequently. In summary, numerical results indicate that grade prediction could be an interesting novel application of AI, but more research is needed to reduce outliers.

Keywords: grading, grade prediction, machine learning, examinations

1. Introduction

The COVID-19 pandemic has created significant challenges for secondary schools in the UK with prolonged remote learning, lack of socialisation, and the cancellation of the 2020 and 2021 GCSE and A-level exams. All of these are expected to have a long-lasting impact on students.

GCSE exams are a crucial stepping stone for young people in England. Years of learning are assessed in a single exam period, the result of which often determines future career options. The grades are also of considerable importance to schools and teachers, as they form the basis of school league tables that can then influence future school applicants and funding, especially in independent schools.

With the cancellation of the 2020 and 2021 GCSEs, the Office of Qualifications and Examinations Regulation (Ofqual) was forced to consider alternative approaches to award grades. GCSE results are critical for university applications, hence there was great pressure to make the new approach comparably accurate to formal exams to ensure a fair allocation of university places. Ofqual's initial proposal (2020a; 2020b) consisted of centre-assessed grades (CAGs) that were based on teacher predictions, and a standardisation algorithm that mapped CAGs to a predicted distribution derived from each centre's historical data. This got withdrawn following a public backlash, awarding each student the maximum of their unmodified or mapped CAGs in 2020, and using teacher-assessed grades (TAGs) in 2021. These events led to unprecedented grade inflation, which in turn raised questions about educational standards and year-to-year comparability.

There is insufficient research to conclude whether the Ofqual algorithm's numerical accuracy was acceptable and how it would compare to the performance of other potential AI models. As such, it is unclear whether grade prediction models could be a viable novel application of AI in education despite the failure of the Ofqual model.

This paper investigates whether AI can be used as an alternative to exam-based grades. Using detailed information about students' past performance, we quantitatively evaluate the efficacy of state-of-the-art machine learning (ML) models in predicting student grades. The following research questions (RQs) are addressed:

> RQ1: What is the efficacy of modern ML models (as well as the 2020 Ofqual model) in predicting exam grades as measured by the accuracy of predictions compared to actual grades awarded?
>
> RQ2: What are the differences in model accuracy

*Correspondence to: The Perse School, Hills Rd, Cambridge CB2 8QF, United Kingdom

Email address: gdenes@perse.co.uk (Gyorgy Denes)

across subjects, and could these be explained by the qualitative difference in teachers' grading practices between STEM (science, technology, engineering, and maths) and non-STEM subjects?

The above investigations are scoped to a case study of a single selective, independent UK school. The sample size is a limitation of this study; however, the findings could still be informative for any other institutions considering ML-based assessments, while results can also serve as a baseline for future, potentially larger-scale studies.

The rest of the paper is structured as follows: Section 2 introduces the background and related literature; Section 3 describes the case study dataset, Section 4 investigates RQ1, Section 5 investigates RQ2. Results are discussed at the end of the manuscript.

2. Background and Related work

2.1. AI in Education

There is an increasing interest in the use of AI for educational purposes Chen et al. (2021, 2022). In particular, smart learning often incorporates an aspect of student performance modelling; however, there are few published studies on the topic of grade predictions as a potential substitute for exams.

Anders et al.2020 have recently used a random forest classifier for A-level predictions. Similar models have been used for university intervention programs (Beaulac and Rosenthal, 2019). Other examples of AI include automated grading and learner-facing AI. For a full review, refer to Zawacki-Richter et al. (2019) and Chen et al. (2022).

The 2020 Ofqual algorithm was a recent attempt to replace exams with an AI model; however, there is insufficient research on the algorithm's accuracy. The Ofqual model is discussed in more detail in Section 2.6. AI prediction models could potentially replace formal exams. This paper contributes to this field by carefully investigating the numerical accuracy of such models.

2.2. Grading

Student performance is normally observed through proxies, such as raw exam marks, which are then transformed into more intuitively interpretable linear grades. Transformation from proxies to grades can be done using Rasch (1993), Thurston V models (Engelhard, 1984; Andrich, 1978), or simpler approximations of these (Pearson, 2016).

Twissel (2011) has shown that student performance correlates with general intelligence (measured through intelligence tests), as well as personality traits, especially conscientiousness or grit (Rimfeld et al., 2019). Intelligence, personality and motivation account for a large portion of grading variance (Kappe and Van Der Flier, 2012), but interestingly Furnham and Monsen (2009) found that the correlation of school grades with IQ scores diminishes with age ($r = 0.6$ down to 0.4 from primary to secondary education). The relative importance of intelligence and personality varies across subjects, IQ scores being better predictors of school grades for STEM, and the reverse for languages. Glaesser and Cooper (2012) found that for GCSEs, high ability leads to high grades for most students. Other required conditions exist on parental education and gender. Correlation models point towards the plausibility of numerical predictions even in a year without a physical exam taking place, which serves as a motivation for this paper.

2.3. Grade prediction

Studies from the last four decades agree that teachers overestimate student performance (Anders et al., 2020). This is not to say that teacher predictions do not have high utility: teacher rankings and predictions correlate strongly with test scores (Rimfeld et al., 2019). Gill (2019) presents how teachers use a variety of methods and consider multiple factors to provide predictions, including student engagement and mock exam results. Yet, Anders et al. (2020) have shown that A-level predictions still indicate that over 40% of the grades are over-predicted (and less than 10% are underpredicted). England's Universities and Colleges Admissions Service (UCAS) claim that over 50% of the students have at least 2 out of 3 grades overpredicted (UCAS, 2016). Some authors including McManus et al. (2020) speculate that up to 3 grades of overpredictions are possible. Interestingly, Attwood et al. (2013) found students' self-assessed grades to have similar accuracy.

Grade prediction has a non-uniform bias across society. Students from disadvantaged parts of the country, from ethnic minorities and those who have lower GCSE grades are more likely to miss their predicted grades for A-levels (UCAS, 2016; Gill, 2019). High grades are more accurate partly due to the *ceiling effect* as described by Anders et al. (2020): it is simply impossible to over-predict when the student is likely to get a top grade (A* or 9). Delap (1994) shows that while a significant difference in prediction for ethnic minorities is present, this is mostly explainable by differences in school type. Independent schools also engage significantly more actively in higher education applications with more thorough GCSE and A-level preparation processes, advising, and help with personal statement preparations and reference writing as pointed out by Dunne et al. (2014). Grade prediction can emphasise or suppress unfairness in the qualification system; as such, average numerical accuracy does not always provide a satisfactory measure of how useful a prediction is. To address this, we also analyse prediction outliers.

2.4. Fairness

When analysing exam results, Coe (2008) and He et al. (2018) found a noticeable difference in *difficulty* across subjects, attributed to teaching time allocation, number of

students who want to take the subject etc. GCSE grades have also been speculated to be affected by short-term high-quality teaching which might benefit those coming from private education (Ogg et al., 2009), which indicates that some unfairness has been present in grading even when students could sit their GCSE exams uninterrupted by the pandemic. Attempts including inspections (Ofsted, 2019) are meant to increase quality and equality. However, studies from Shaw et al. (2003) and Rosenthal (2004) found that inspections can have a negative effect, and they mostly benefit selective schools (i.e. schools which filter their students at the time of application based on academic merit). This can further increase inequality.

Numerical models, even if accurate on average, have been shown to introduce severe injustice, and can act as Weapons of Maths Destruction as coined by O'Neil (2017). The author identifies features of a mathematical system which make it dangerous: opacity, scale, damage and the presence of a self-justifying feedback loop (O'Neil, 2017, p. 33). For GCSE grade prediction, this paper considers the use of less opaque models as well (e.g. random forests).

2.5. Qualifications in the UK

Different organisations across the UK are responsible for qualification regulation across the four nations (Ofqual, 2019). In England, Ofqual is a non-ministerial government department responsible for qualifications, examinations and assessments. Ofqual has a set of legal requirements for awarding qualifications, including "giving a reliable indication of knowledge, skills and understanding", while also "indicating a consistent level of attainment (including over time) between comparable assessments" and "promoting public confidence in qualifications" (UK Parliament, 2009). Such objectives are hard to satisfy concurrently, especially in the uncertainty of no physical exams.

GCSE exams in England are taken at the end of year 11 at age 16 and are of considerable importance for students, teachers and schools. Since 2017, grading has been discrete numerical (9–1), with 9 being a top grade, 1 being a bottom grade. There is a non-trivial transformation from the old ordinal A*–G scale (Harwell and Gatti, 2001; Wetzler, 2019; Pearson, 2016). Grades are considered linear.

2.6. The Ofqual algorithm

With no exams in 2020, Ofqual proposed their Direct Centre Performance model, which is a hybrid pipeline designed to replace formal assessments (2020b). The model can be considered a form of white-box AI.

First, teachers across England estimated plausible CAGs for each student for each subject alongside a strict ordering (ranking) of students within each subject. In the meantime, Ofqual produced a predicted grade distribution for each centre for each subject, derived from historical performance of the exam centre (school) at previous GCSE exams as well as the prior attainment profiles of candidates (i.e. previous exam results such as the Standard Assessment Tests taken at age 11). Finally in the *standardisation* step, Ofqual performed distribution mapping from the CAGs to the expected grade distributions, producing the so-called *calculated grades*. Rankings were used to resolve candidates around grade boundaries. Exceptions were made for small exam centres, where predicted grade distributions were deemed less reliable (Ofqual, 2020a,b).

Ofqual argued that the standardisation step is required, as previous studies indicate that teachers are likely to over-predict (see Section 2.3), which could result in undesirable grade inflation. Ofqual also explored alternative models of standardisation (linear and logistic regression) and picked the distribution mapping approach based on historical data (Ofqual, 2020a).

One major criticism is that Ofqual only released details of the algorithm in August (Ofqual, 2020a), after results were announced, following a direct request from the House of Commons Education Committee (2020).

Furthermore, AI prediction models in the literature are often analysed through their numerical accuracy, and while the Ofqual model can be considered an AI model, no such studies have been published by Ofqual.

2.7. Results of the Ofqual algorithm

CAGs are produced through a sequence of rating and ranking steps, following a more formal process than other teacher-predicted grades, such as the ones used for UCAS university applications (Ofqual, 2020b, p.7). However, it is reasonable to suspect that CAGs can reproduce the positive bias of teacher-predicted grades (Section 2.3), acknowledged by Benton (2021), and the House of Commons Education Committee report (2020). Hence, it is expected that the algorithm lowers the majority of CAGs to match the expected distributions. This correction affected state schools and minorities more severely, resulting in a public backlash (Piwowarski, A, 2020; The Guardian, 2020; BBC News, 2020). Eventually, Ofqual awarded the *maximum* of CAGs and the *calculated* grades (Cambridge Assessment, 2020).

Lee and Newton (2021) argue that there is no statistical evidence for systematic disadvantage in *calculated grades* based on socioeconomic status, mixed findings for ethnicity, and only a small bias in gender. Kelly (2021) provides an excellent description of the initial concerns about rogue results, through social and political challenges to the repeal of CAGs standardisation. One possible interpretation of the events is that the algorithm tried to solve too many problems and failed to secure public confidence and trust. Kippin and Cairney (2021) argue that the fiascos (which occurred in all four nations of the UK) are the results of the order and timing of political events, misjudged political feasibility, and the lack of inclusion of the educational communities in the decision-making. The current consensus is that the technical limitations of the algorithm did not play a significant part. Our quantitative comparison of

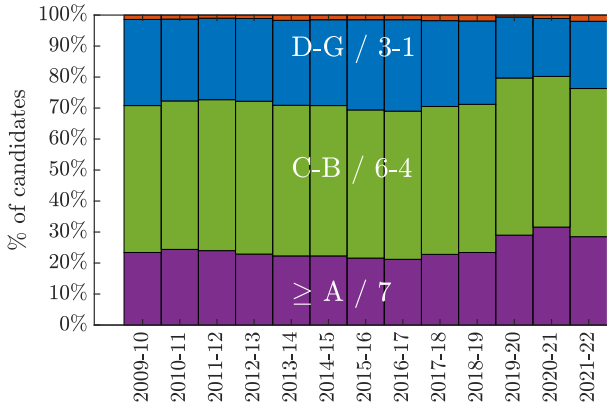

Figure 1: There is visible inflation in the top and middle grades in 2019-2020 when the Ofqual algorithm was repealed. Similar trends can be observed for 2020-2021 when evidence-based teacher-assessed grades were used. Data from UK Government (2022)

the Ofqual model with widely-used ML models contributes further to this discussion.

The 2020 exam resulted in unprecedented grade inflation (Wei Lee et al., 2020) and created new debates on potential unfairness against high achievers. McManus et al. (2021) argue that there will be a long-lasting negative impact on fields such as medicine; holistic judgement and teachers' expert knowledge cannot replace rigorous external assessment. TAGs in 2021 have reproduced the 2020 trends (see Figure 1) and hence raise similar questions for the 2021 cohort.

2.8. ML

ML is a branch of AI that has been used successfully in data-rich fields. The input of an ML model is raw data or a *feature vector*, which the ML model turns into an output. Such prediction models have been hand-crafted in the past, but in *supervised learning*, the model is fitted by the computer itself based on historical data. In this paper, the input data are proxies of student ability such as previous grades, raw test scores, and general intelligence. The output is a predicted grade.

Models can be further categorised into regression and classification models. Regression models output continuous numerical values, while classification models output nominal *labels*. As GCSE grades are discrete numbers (1–9), it is not trivial to decide which class of models to use. The outputs of regression models need to be rounded and clamped, while classification models are also suboptimal, as they are unaware of the numerical differences between grades (and focus on getting a grade exactly right or wrong). Anders et al. (2020) have used a classification model for the similar problem of predicting A-level results in England. In Section 4.1, we explore both approaches.

ML models have two notable limitations: they can overfit to the training data, and their behaviour can be hard to explain. To tackle and quantify overfitting, it is best to split the available data to *train* and *test* datasets. The

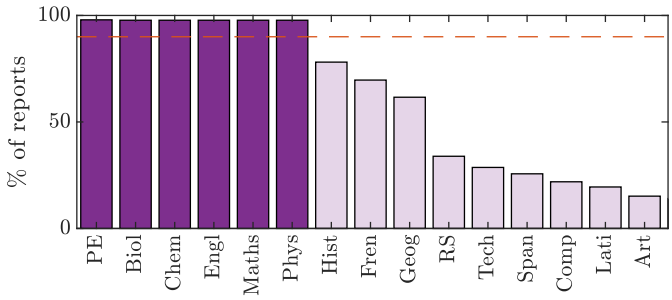

Figure 2: % of reports which contain entries for each subject. Only subjects with > 90% coverage were used as features in the analysis.

model learns in a supervised manner on the training data, but its performance is evaluated on the test data (data it has never seen before). Regularisation parameters can also help the model generalise better to unseen data.

3. Case study dataset

This paper builds on a dataset provided by an independent, selective secondary school in England. All models aim to predict GCSE grades. To evaluate the models, predictions were compared to awarded GCSE grades from 2018 (N=180) and 2019 (N=176).

For model inputs, the following data were considered for each student in the 2018 and 2019 cohorts: (1) Middle Years Information System (MidYIS) results from year 9 which can be considered a proxy of intelligence; (2) yearly number of merits in years 9–11 which are awarded for positive attitude; (3) average yearly report grades for years 9–11 (RepYr); (4) detailed subject grades for common subjects taken by most candidates (see Figure 2): biology, chemistry, English, maths, PE, physics (RepD). There are three reports written each year by each subject teacher with three grade categories in each: standard of work (SoW) capturing the academic results, engagement (E) capturing lesson contributions and effort, and organisation (O) capturing timely homework submissions. For letter-based grades, Pearson's (2016) guidance, in-house expert rules, and linear interpolation were used to scale to a 9–1 scale. Table 1 shows a few rows of example features.

To investigate grading practices across subjects, additional data was collected using a questionnaire. Participants were selected from the teaching staff of the same school that provided the dataset for the AI models.

4. Accuracy of predicted grades

This section explores RQ1, the efficacy of modern ML models and the Ofqual algorithm in the context of GCSE grade predictions. Specifically, we compared model outputs (predicted GCSE grades) with actually awarded grades, and quantified model accuracy as mean absolute error (MAE). Higher MAE means a higher average error in the magnitude of mispredictions, which in turn means poorer model performance. MAE is preferred here over

mean error, so negative and positive mispredictions do not cancel out, while it retains a linear meaning, unlike root-mean-square error (Schneider and Xhafa, 2022, p.59). Accuracy is computed for several subjects to allow for an analysis of model accuracy across subjects.

2-fold cross-validation was used, where the first iteration trained and validated models on 2018 and 2019 data respectively. In the second iteration, the years were swapped. The *test* values in the rest of the paper show the average metrics across the two validations.

4.1. Models

A number of popular models have been considered. Future sections refer to the models by the abbreviated names (square brackets). Except for Ofqual, implementations are based on Scikit Learn (Pedregosa et al., 2011). Regularisation parameters were tuned manually to reduce overfitting to the relatively small dataset. The results of regression models were rounded and clamped to 1..9.

[LinR] Linear ridge regression; $\alpha = 10.0$

[LogR]: logistic regression; $C = 0.01$

[BayR]: Bayesian ridge regression

[GaussR]: Gaussian process regression; $\alpha = 5.0$

[RandF2R]: random forest regression; tree depth 2.

[RandF2C]: random forest classifier, used for public exam grade prediction in England by Anders et al. (2020); tree depth 2.

[MLPC]: multi-layer perceptron classifier; $\alpha = 10.0$ and max_iter = 1500

[Ofqual]: distribution mapping similar to Ofqual's. The predicted test grade distribution is assumed to be identical to the training grade distribution and is described using its cumulative distribution function ($f(grade)$). The model then uses the cumulative distribution function of the input feature ($g(x)$) and computes grade as $f^{-1}(g(x))$ (Figure 3).

4.2. Results

For a visual summary of average model performance across all subjects, see Figure 4 and Figure 5. The confusion matrices in Figure 9 provide a more detailed breakdown (mispredictions highlighted in red).

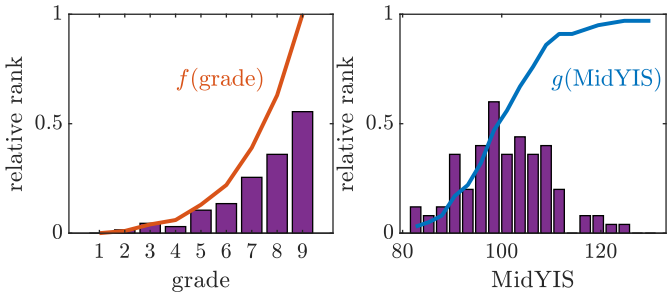

Figure 3: [Ofqual] finds the cumulative distribution function ($f(grade)$) of historical data (left) as well as the cumulative distribution function of the test feature vector ($g(x)$), where x could be report grades, MidYIS results, etc. x is transformed using $f^{-1}(g(x))$.

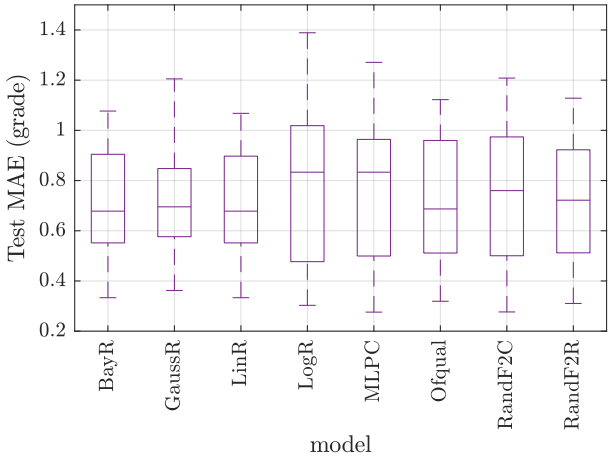

Figure 4: Grade prediction accuracy (test scores, all subjects) with all features. Most predictions are within 1 grade from actual results.

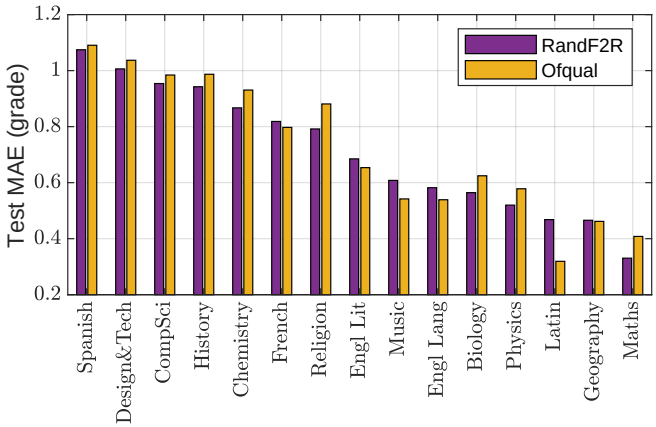

Figure 5: Grade prediction error reasonably low (< 1 grade), but it varies greatly across subjects (plotted for RandF2R vs Ofqual).

MidYIS				RepYr							RepD					Merits		
Vocab	Maths	Non-ver	Overall	9-E	9-O	9-SoW	...	11-E	11-O	11-SoW	9-biology-E	-O	-SoW	...	9-physics-E	9	10	11
119	120	123	123	6.1	6.7	6.9	...	5.7	6.6	7.2	8	6	8	...	8	52	12	8
116	93	103	105	5.2	6.3	6.3	...	5.1	6.9	6.4	6	6	7	...	6	34	21	6

Table 1: Example features in the dataset. Grades are rounded for presentation only. With all report grades, there are 68 inputs in total.

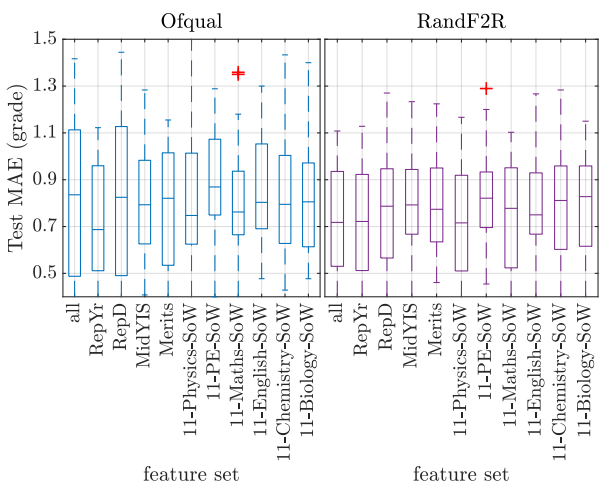

Figure 6: Features have comparable saliency with no single feature that significantly outperforms all others (ANOVA p> 0.05).

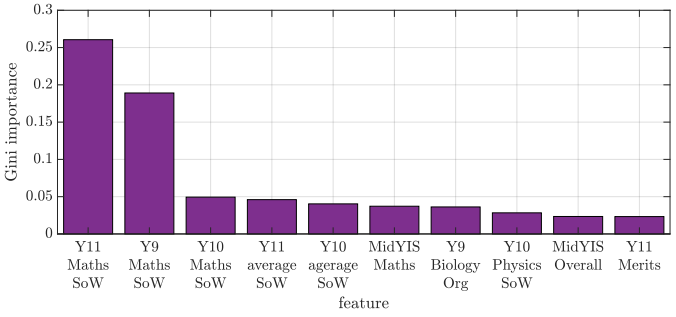

Figure 7: RandF2C for mathematics: Top 10 impurity-based feature importances (Gini importance) show that report-based features are more important than MidYIS scores and merits.

Ablation study on the input data (Figure 6) shows that there are only insignificant differences between parts of the features (ANOVA p> 0.05). Feature importance analysis (Figure 7) can provide further information.

5. STEM vs. non-STEM grading survey

Grade prediction errors vary greatly across subjects. While for some subjects, this can be explained by the relatively small sample sizes, this section explores two other factors that can contribute to the differences: *subjectiveness* and the *perceived purpose* of internal SoW grades.

Machine learning models are best trained on objective ground-truth data. If we can explain the quantitative differences in model accuracies using qualitative differences in teachers' grading practices, that can help us further understand the limitations of such models.

A questionnaire-based survey provides an easy way to investigate subjectivity. The following were investigated:

> [H1]: there is a difference in *subjectiveness* between STEM and non-STEM grading.

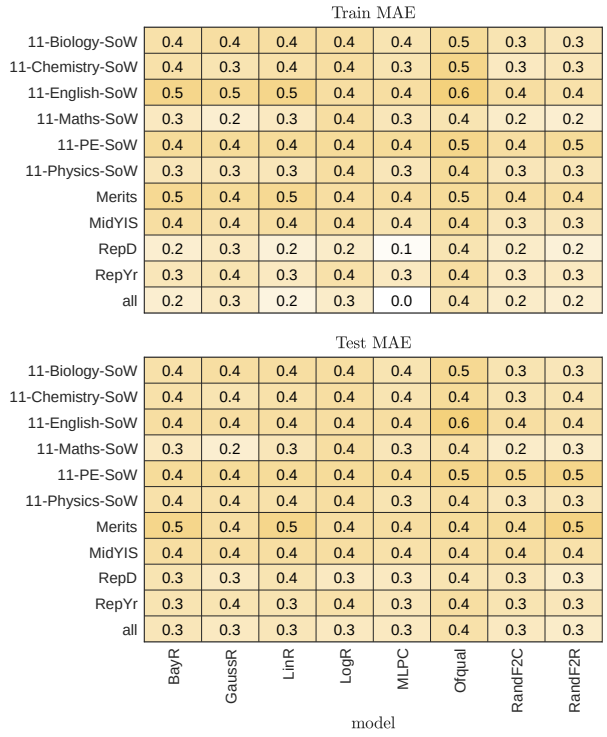

Figure 8: Train vs. test MAE for maths shows some overfitting, but this is only substantial in MLPC

.

> [H2]: there is a difference between STEM and non-STEM teachers whether they consider SoW report grades indicative of future GCSE grades.

Both hypotheses were operationalized as a series of questions. Figures 10 and 11 show the indicator questions for [H1] and [H2] respectively. The study also explored further teacher considerations when awarding a SoW grades.

Responses on the Likert scale were mapped 1–5; N/A responses were excluded. Both 2-sample t-tests ($p \ll 0.05$) and Mann-Whitney U-tests were performed to establish whether a significant difference exists between STEM and non-STEM grading.

5.1. Participants

35 members of the teaching staff participated in the study. Out of this, 34 were included in the analysis, excluding the single respondent who does not teach years 9–11. Each teacher's primary subject was categorised as STEM ($N = 19$ from mathematics, physics, chemistry, biology, geography, computer science) or non-STEM ($N = 15$ from English, modern or classical foreign languages, philosophy, ethics, religion, history, music, art&design)

5.2. Results

As shown in Figure 10 and Figure 11, the majority of the questions support both H1 and H2. STEM grading is perceived to be more objective with stronger teacher consensus on what constitutes a correct answer and greater agreement on marking schemes. Equally, non-STEM teachers

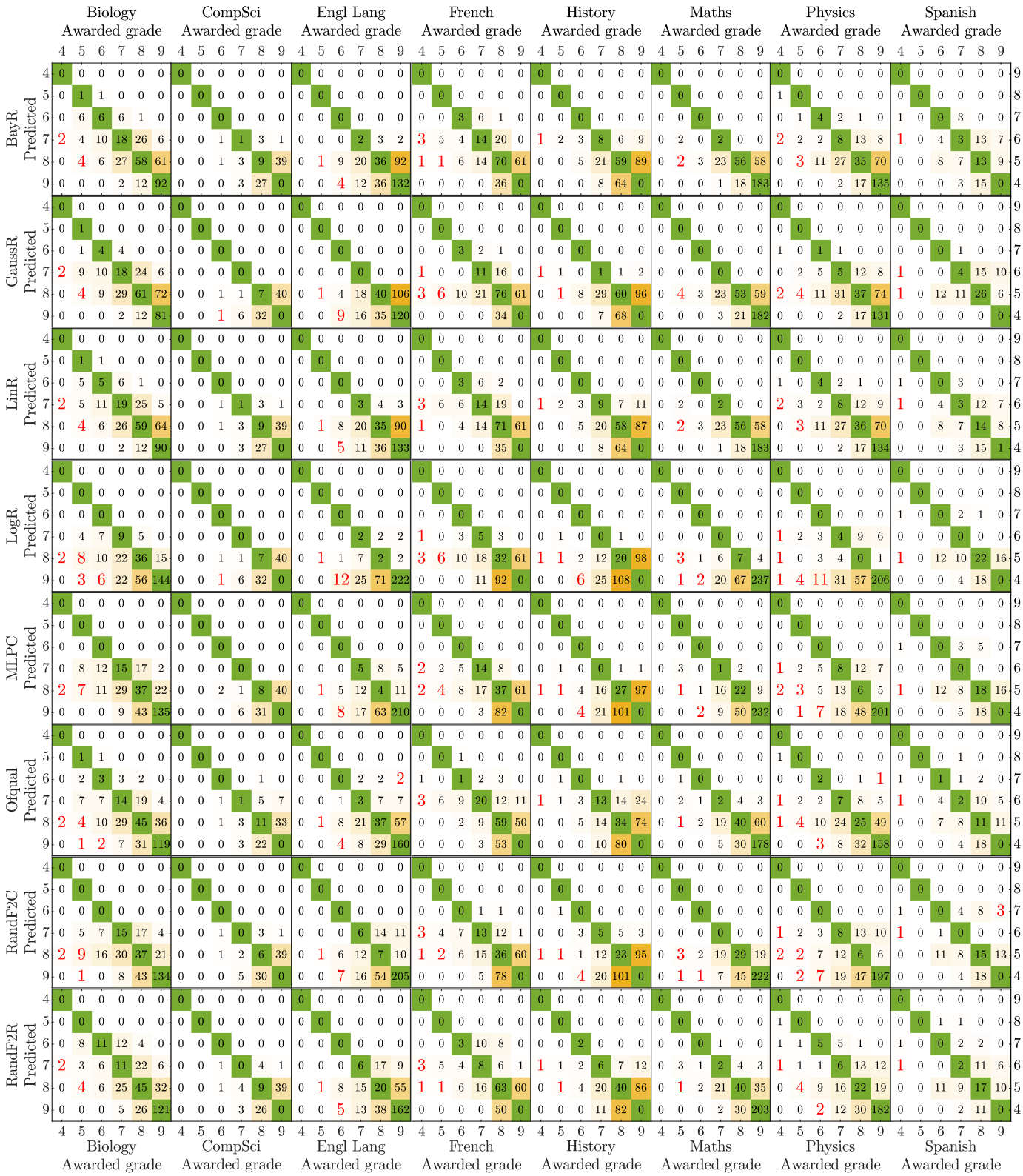

Figure 9: Confusion matrices for model–subject pairs on the RepYr feature set summed over the two test sets. There were no grades below 4. Green diagonals indicate perfect predictions. Most results are either on or very close to the diagonal. Strong mispredictions (underpredicted or overpredicted by more than 2 grades) are highlighted in red. STEM subjects are generally more accurate (proportionally more items on the diagonal); however, STEM subjects are not free from strong mispredictions either.

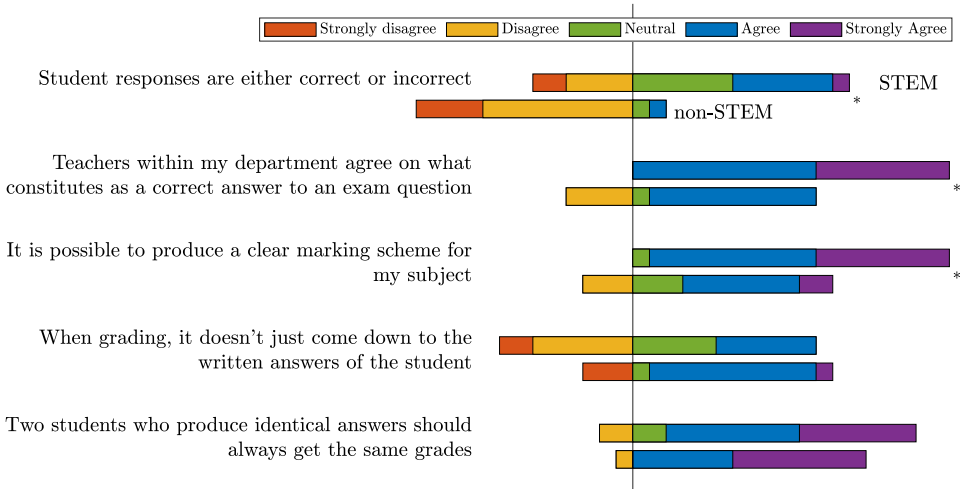

Figure 10: H1: grading subjectivity difference between STEM and non-STEM teachers. For each question, top bars: STEM, bottom bars: non-STEM. Values to the left go from "strongly disagree" to "strongly agree". Asterisks indicate significant differences.

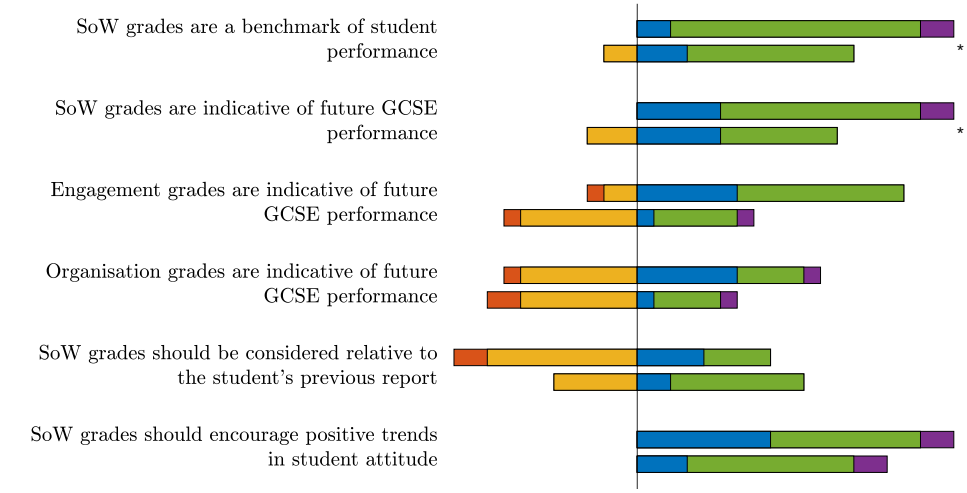

Figure 11: H2: difference in the perception of whether standard of work report grades are indicative of GCSE grades.

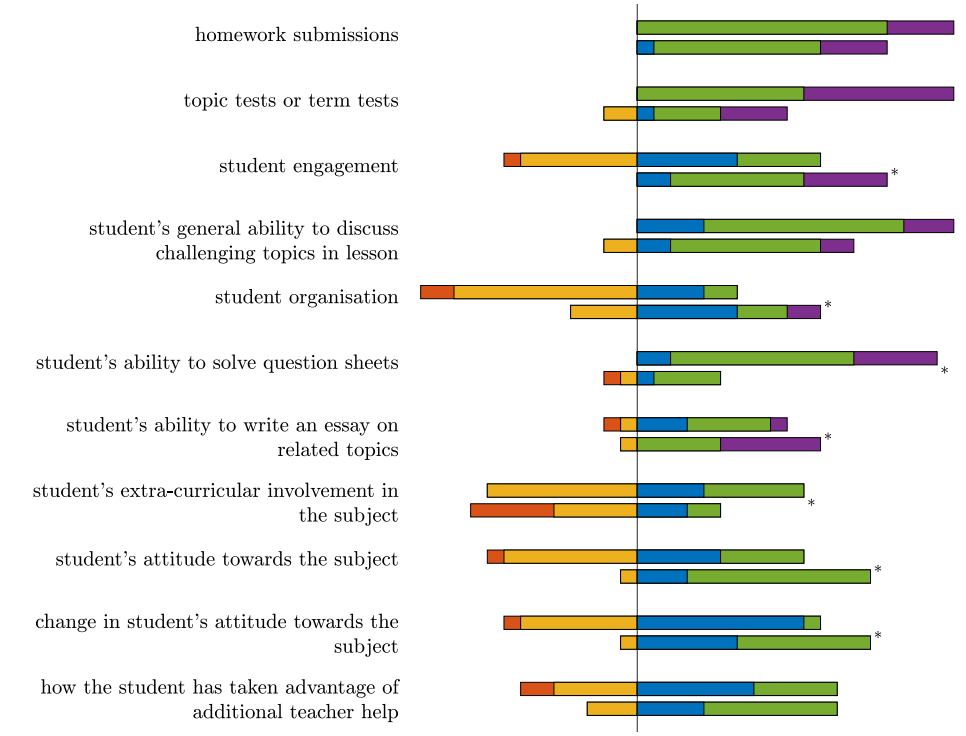

Figure 12: Exploring different factors teachers consider when awarding standard of work report grades.

are less likely to consider end-of-term reports to be indicative of future GCSE performance, and incorporate evidence from student engagement, organisation, essay writing skills, general and positive change in the attitude of students towards their subject.

6. Discussion

6.1. Accuracy of modern ML models (RQ1)

With respect to RQ1, results show that most GCSE predictions ($> 75\%$) are within 1 grade from the actual results. However, the confusion matrices highlight the presence of some strong mispredictions (> 2 grades). Even the most accurate models potentially underpredict awarded GCSE grades by up to 3 grades, or overpredict by up to 4 grades. This indicates that the current models are only suitable for the majority of candidates, and can introduce unfairness if deployed as a substitute for public exams.

6.1.1. Accuracy differences across models

There is no significant difference between classification vs. regression models, and there is no clear difference between competing model architectures (ANOVA $p > 0.05$). The choice of model hence does not play a critical role in the target dataset.

6.1.2. Overfitting

Models suffer from some overfitting (Figure 8); however, the performance difference between the test and the train datasets is small (< 1 grade). This is mostly due to the careful manual choice of model regularisation parameters. Such *hyperparameter tuning* can be automated for larger datasets.

6.1.3. Choice of features

Features have comparable importance when measured across the entire dataset. However, using all features at the same time can introduce higher variance in the results, especially in the Ofqual model. Qualitatively, RepYr seems like an ideal choice, as it performs consistently well, capturing performance across all subjects, while being also low-dimensional, which can reduce overfitting.

When only considering an individual subject, such as maths, Figure 7 indicates that RandF2C finds individual and average grades more salient than proxies of general intelligence or teacher-awarded merit points. Only including the most salient features can reduce training time and can help improve the explainability of the model.

6.2. Differences across subjects (RQ2)

Model accuracy varies across subjects. Overall, more accurate predictions are provided for STEM subjects, however, subjects with smaller cohorts such as Latin and Design & Technology contradict this trend (Figure 5). This is also illustrated in Figure 13: linear correlation coefficients between SoW and GCSE grades for STEM subjects, especially physics, maths and biology, are remarkably high (0.53-0.59) even in year 9. However, in English (language and literature) correlation with exam results remains weak-to-moderate even in year 11. The strength of correlation differs between grading aspects: E and O grades show weak-to-moderate (0.1-0.4) correlation and with no noticeable change over time, while for most subjects, SoW grades show moderate correlation in year 9, then strong correlation in years 10 and 11.

Some of these differences can be also explained by comparing grading practices between STEM and non-STEM subjects. Results in Section 5 indicate that non-STEM subjects use more subjective grades, which introduces more noise both into the input features and the output awarded grades for these subjects. Furthermore, non-STEM teachers also do not consider report grades as indicative of future GCSE performance as their STEM colleagues, which introduces further noise into some of the most salient input features.

With all of these combined, it seems that AI models have a higher chance of success to replace formal exams for STEM subjects.

6.3. Summary

While these exact numbers are specific to the target school, it seems that the presented ML models can predict grades with acceptable accuracy for the majority of the target population. The models might make huge errors for a few candidates. GCSE grades are critical for university applications, and such strong mistakes could prohibit an individual from securing a university place, which makes them in their current shape unsuitable as a substitution for a formal exam. This is especially true for STEM subjects.

7. Conclusion

The COVID-19 pandemic has forced us to re-evaluate many of our assessment practices, including major external exams such as GCSEs. In the UK, Ofqual experimented with using AI to replace formal exams during the pandemic, but they were faced with a public backlash.

This paper has shown a quantitative investigation whether AI, specifically ML models could provide a viable alternative to formal GCSE exams in the context of a selective, independent English school. Results indicate that for the majority of the students the predictions are accurate (MAE$<$1 grade). All explored models, including the Ofqual 2020 model perform comparably. There are some strong mispredictions with some grades underpredicted by 3 or overpredicted by up to 4 grades (on a scale of 9–1). This indicates that numerical models alone are not yet suitable to replace public exams. Future research incorporating an individual appeal processes could help mitigate these limitations.

Prediction performance has been shown to be subject dependent; specifically, predictions are more accurate for STEM subjects and for subjects with more students. In STEM subjects, SoW grades even in year 9 reports are strongly indicative of GCSE performance. In non-STEM subjects such trends cannot be observed. This has been investigated through a qualitative questionnaire, demonstrating that teachers consider non-STEM subject marking more *subjective*. STEM and non-STEM teachers also have a different perception of the goal of awarding standard of work grades – significantly more STEM teachers believe that these grades should be indicative of future GCSE performance. While the same finding might not generalise to other institutions, this is an interesting finding, showing how objective numerical models should be deployed with additional care for non-STEM subjects.

One major limitation of this case-study-based investigation was the small size and specialised (independent) nature of the school dataset; results might not apply to other institutions. Future work could explore unifying databases across state and fee-paying schools to reevaluate the ML models discussed in this paper on a larger scale.

Underperforming students are often spotted late by their teachers. The ML models discussed in this paper could be adapted (especially for large STEM subjects) to flag potential underperforming students as early as year 9, and offer them additional help. A vertical study could test the efficacy of such methods.

Additionally, while this paper described some issues related to the public perception of numerical models, more research is needed on how and whether public trust should be established in AI models in education.

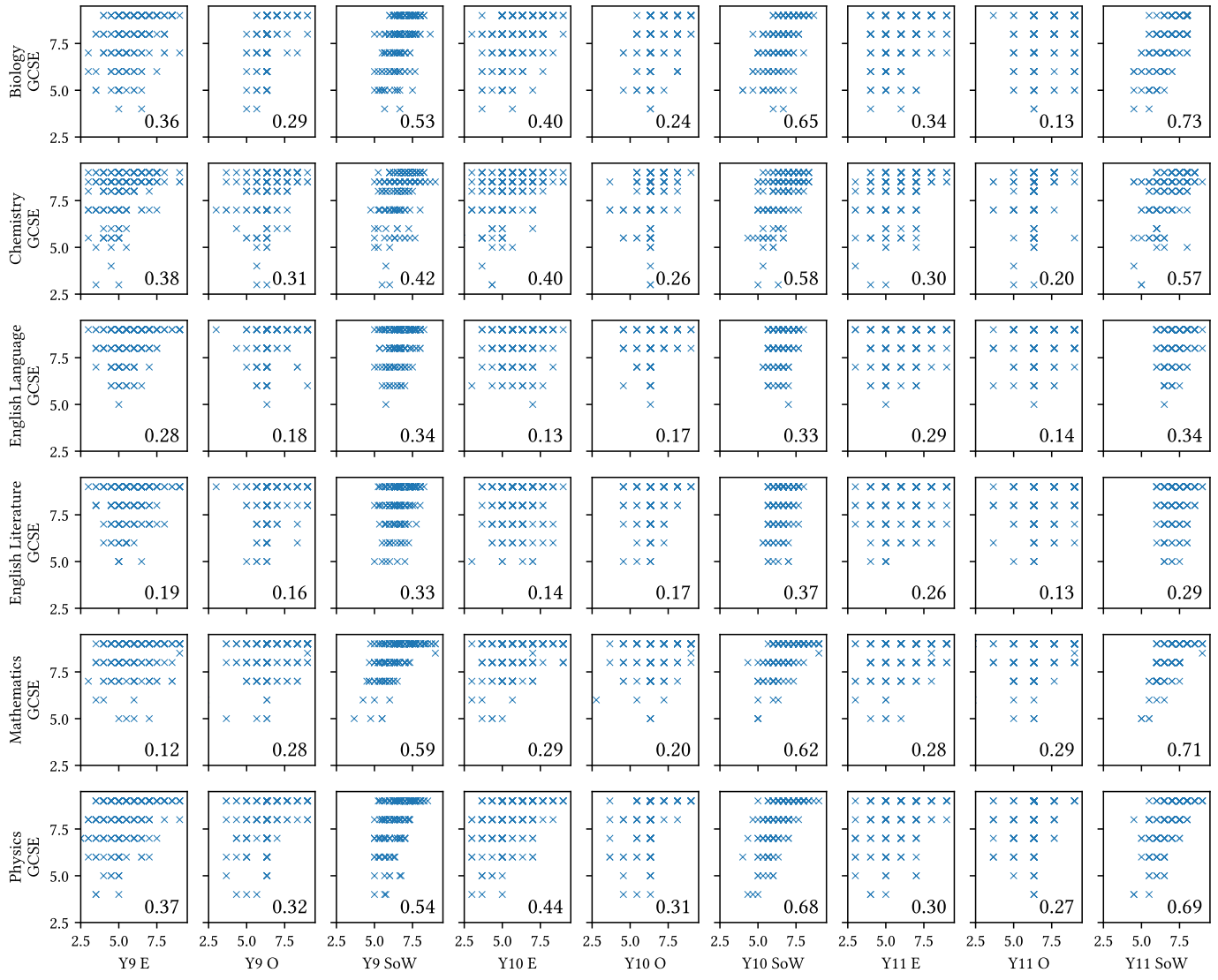

Figure 13: Pearson correlation between GCSE exam grades and mean report grades from years 7–9 (Y7–Y9) for E, O, and SoW. Correlation coefficients (inlined in subplots) highlight substantial differences can be observed across subjects. For STEM, standard of work is the best indicator of future GCSE performance.

Declaration of interest

The author is employed as a teacher at the selective, independent school in England that provided the data for the case study.

Ethical approval

The study received ethical approval from the Head Of Research at The Perse School.

Acknowledgements

The author would like to thank Dr Sue Brindley, Dr Andrea Kocsis, Dr Sylwia Macinska, Dr Aliaksei Mikhailiuk and the anonymous reviewers for their insights and suggestions, and Matt Fox for his support. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Acronyms

CAGs centre-assessed grades
E engagement
GCSE General Certificate of Secondary Education
MAE mean absolute error
MidYIS Middle Years Information System
ML machine learning
O organisation
Ofqual Office of Qualifications and Examinations Regulation
RepYr average yearly report grades for years 9–11
SoW standard of work
STEM science, technology, engineering, and maths
TAGs teacher-assessed grades
UCAS England's Universities and Colleges Admissions Service

References

Anders, J., Dilnot, C., Macmillan, L., Wyness, G., 2020. Grade expectations: How well can we predict future grades based on past performance? CEPEO, UCL .

Andrich, D., 1978. Relationships between the thurstone and rasch approaches to item scaling. Applied Psychological Measurement 2, 451–462.

Attwood, G., Croll, P., Fuller, C., Last, K., 2013. The accuracy of students' predictions of their gcse grades. Educational Studies 39, 444–454. doi:10.1080/03055698.2013.776945.

BBC News, 2020. A-levels and gcses: How did the exam algorithm work? URL: https://www.bbc.co.uk/news/explainers-53807730. accessed: 2021-06-03.

Beaulac, C., Rosenthal, J.S., 2019. Predicting university students' academic success and major using random forests. Research in Higher Education 60, 1048–1064.

Benton, T., 2021. On using generosity to combat unreliability. Research Matters 31, 22–41.

Cambridge Assessment, 2020. Cambridge international update on how we are awarding grades for the june 2020 series. Cambridge Assessment, International Education URL: https://www.cambridgeinternational.org/news/news-details/view/update-on-how-we-are-awarding-grades-for-june-2020-series-20200817/. accessed: 2022-04-26.

Chen, X., Zou, D., Xie, H., Cheng, G., Liu, C., 2022. Two decades of artificial intelligence in education. Educational Technology & Society 25, 28–47.

Chen, X., Zou, D., Xie, H., Wang, F.L., 2021. Past, present, and future of smart learning: a topic-based bibliometric analysis. International Journal of Educational Technology in Higher Education 18, 1–29.

Coe, R., 2008. Comparability of GCSE examinations in different subjects: an application of the Rasch model. Oxford Review of Education 34, 609–636. doi:10.1080/03054980801970312.

Delap, M.R., 1994. An investigation into the accuracy of a-level predicted grades. Educational Research 36, 135–148.

Dunne, M., King, R., Ahrens, J., 2014. Applying to higher education: comparisons of independent and state schools. Studies in Higher Education 39. doi:10.1080/03075079.2013.801433.

Engelhard, G., 1984. Thorndike, Thurstone, and Rasch: A Comparison of Their Methods of Scaling Psychological and Educational Tests. Applied Psychological Measurement 8, 21–38. doi:10.1177/014662168400800104.

Furnham, A., Monsen, J., 2009. Personality traits and intelligence predict academic school grades. Learning and individual differences 19, 28–33.

Gill, T., 2019. Methods used by teachers to predict final a level grades for their students. Research Matters (UCLES) , 33–42.

Glaesser, J., Cooper, B., 2012. Gender, parental education, and ability: their interacting roles in predicting gcse success. Cambridge Journal of Education 42, 463–480.

Harwell, M.R., Gatti, G.G., 2001. Rescaling Ordinal Data to Interval Data in Educational Research. Review of Educational Research 71, 105–131. doi:10.3102/00346543071001105.

He, Q., Stockford, I., Meadows, M., 2018. Inter-subject comparability of examination standards in GCSE and GCE in England. Oxford Review of Education 44, 494–513. doi:10.1080/03054985.2018.1430562.

House of Commons Education Committee et al., 2020. Getting the grades they've earned: Covid-19: The cancellation of exams and 'calculated'grades. URL: https://committees.parliament.uk/publications/1834/documents/17976/default/. accessed: 2022-02-17.

Kappe, R., Van Der Flier, H., 2012. Predicting academic success in higher education: what's more important than being smart? European Journal of Psychology of Education 27, 605–619.

Kelly, A., 2021. A tale of two algorithms: The appeal and repeal of calculated grades systems in england and ireland in 2020. British Educational Research Journal 47, 725–741.

Kippin, S., Cairney, P., 2021. The COVID-19 exams fiasco across the UK: four nations and two windows of opportunity. British Politics doi:10.1057/s41293-021-00162-y.

Lee, M.W., Newton, P., 2021. Research and analysis: Systematic divergence between teacher and test-based assessment: literature review .

McManus, I., Woolf, K., Harrison, D., Tiffin, P.A., Paton, L.W., Cheung, K.Y.F., Smith, D.T., 2021. Predictive validity of a-level grades and teacher-predicted grades in uk medical school applicants: a retrospective analysis of administrative data in a time of covid-19. BMJ open 11, e047354.

McManus, I.C., Woolf, K., Harrison, D., Tiffin, P.A., Paton, L.W., Cheung, K.Y.F., Smith, D.T., 2020. Calculated grades, predicted grades, forecasted grades and actual a-level grades: Reliability, correlations and predictive validity in medical school applicants, undergraduates, and postgraduates in a time of covid-19. Preprint in medRxiv doi:10.1101/2020.06.02.20116830.

Ofqual, 2019. Referencing the Qualifications Frameworks of England and Northern Ireland to the European Qualifications Framework. Ofqual. URL: https:

//assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/840837/Referencing_the_Qualifications_Frameworks_of_England_and_Northern_Ireland_to_the_European_Qualifications_Framework.pdf. accessed: 2022-02-17.

Ofqual, 2020a. Awarding GCSE, AS, A level, advanced extension awards and extended project qualifications in summer 2020: interim repor. Ofqual. URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/909368/6656-1_Awarding_GCSE__AS__A_level__advanced_extension_awards_and_extended_project_qualifications_in_summer_2020_-_interim_report.pdf.

Ofqual, 2020b. Summer 2020 grades for GCSE, AS and A level, Extended Project Qualification and Advanced Extension Award in maths. Ofqual. Accessed: 2022-02-17.

Ofsted, 2019. Education inspection framework. https://www.gov.uk/government/publications/education-inspection-framework. Accessed: 2022-06-22.

Ogg, T., Zimdars, A., Heath, A., 2009. Schooling effects on degree performance: A comparison of the predictive validity of aptitude testing and secondary school grades at oxford university. British Educational Research Journal 35, 781–807.

O'Neil, C., 2017. Weapons of math destruction : how big data increases inequality and threatens democracy. Penguin Books, London.

Pearson, 2016. Gcse awarding from 2017. https://qualifications.pearson.com/content/dam/pdf/Support/results-certification/gcse-awarding-from-2017.pdf. Accessed: 2021-06-03.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830.

Piwowarski, A, 2020. Ofqual's algorithm: The method to madness URL: https://www.if.org.uk/2020/09/03/ofquals-algorithm/. accessed: 2023-01-22.

Rasch, G., 1993. Probabilistic models for some intelligence and attainment tests. ERIC.

Rimfeld, K., Malanchini, M., Hannigan, L.J., Dale, P.S., Allen, R., Hart, S.A., Plomin, R., 2019. Teacher assessments during compulsory education are as reliable, stable and heritable as standardized test scores. Journal of Child Psychology and Psychiatry and Allied Disciplines 60. doi:10.1111/jcpp.13070.

Rosenthal, L., 2004. Do school inspections improve school quality? ofsted inspections and school examination results in the uk. Economics of education review 23, 143–151.

Schneider, P., Xhafa, F., 2022. Anomaly Detection and Complex Event Processing Over IoT Data Streams: With Application to EHealth and Patient Data Monitoring. Elsevier. doi:10.1016/C2020-0-00589-X.

Shaw, I., Newton, D.P., Aitkin, M., Darnell, R., 2003. Do ofsted inspections of secondary schools make a difference to gcse results? British educational research journal 29, 63–75. doi:10.1080/0141192032000057375.

The Guardian, 2020. Ofqual's a-level algorithm: why did it fail to make the grade? URL: https://www.theguardian.com/education/2020/aug/21/ofqual-exams-algorithm-why-did-it-fail-make-grade-a-levels. accessed: 2021-06-03.

Twissel, A., 2011. An Investigation into the Use of Cognitive Ability Tests in the Identification of Gifted Students in Design and Technology. Design and Technology Education 16, 20–32.

UCAS, 2016. Factors associated with predicted and achieved A level attainment. UCAS. URL: https://www.ucas.com/file/71796/download?token=D4uuSzur. accessed: 2022-04-26.

UK Government, 2022. Educational statistics. https://explore-education-statistics.service.gov.uk/find-statistics/key-stage-4-performance-revised. Accessed: 2023-01-03.

UK Parliament, 2009. Apprenticeships, Skills, Children and Learning Act 2009.

Wei Lee, M., Neil, S., Nadir, Z., 2020. Student-level equalities analyses for gcse and a level: Summer 2020 .

Wetzler, E.L., 2019. How Using a Restricted Grading Range Distorts GPAs and Disproportionately Penalizes Low-Performing Students. Frontiers in Education 4. doi:10.3389/feduc.2019.00023.

Zawacki-Richter, O., Marín, V.I., Bond, M., Gouverneur, F., 2019. Systematic review of research on artificial intelligence applications in higher education–where are the educators? International Journal of Educational Technology in Higher Education 16, 1–27.