The impact of between subject differences in grading standards on composite measures of student ability in the UK

Timothy Stratton

A Research & Development Project

Submitted for the MSc Educational Assessment 2020

DEPOSIT AND CONSULTATION OF THESIS

One copy of your dissertation will be deposited in the Department of Education Library via the WebLearn site where it is intended to be available for consultation by all Library users. In order to facilitate this, the following form should be completed which will be inserted in the library copy your dissertation.

Note that some graphs/tables may be removed in order to comply with copyright restrictions.

Surname	Stratton
First Name	Timothy
Faculty Board	Education
Title of Dissertation	The impact of between subject differences in grading standards on composite measures of student ability in the UK

Declaration by the candidate as author of the dissertation

1. I understand that I am the owner of this dissertation and that the copyright rests with me unless I specifically transfer it to another person.

2. I understand that the Department requires that I shall deposit a copy of my dissertation in the Department of Education Library via the WebLearn site where it shall be available for consultation, and that reproductions of it may be made for other Libraries so that it can be available to those who to consult it elsewhere. I understand that the Library, before allowing my dissertation to be consulted either in the original or in reproduced form, will require each person wishing to consult it to sign a declaration that he or she recognises that the copyright of this thesis belongs to me. I permit limited copying of my dissertation by individuals (no more than 5% or one chapter) for personal research use. No quotation from it and no information derived from it may be published without my prior written consent and I undertake to supply a current address to the Library so this consent can be sought.

3. I agree that my dissertation shall be available for consultation in accordance with paragraph 2 above.

Contents

Abstract 3

Introduction 5

Comparability of subject grades 5

What are exam standards? 7

Conceptions of comparability 8

Measuring subject difficulty 9

Interpreting evidence across subjects 10

Creating a composite measure 12

Issues with composite measures 13

Rasch Equating 14

Composite measures in England 15

Performance Measures 15

Critiques of performance measures in England 18

Evidence of perverse incentives in performance measures in England 19

Aims/Rationale 21

Method 22

Data 22

Analyses 22

1. Differences in statistical subject standards 23
2. Simulate effect on accountability scores 24
3. Imputation of student grades 26
4. Modelling factors influencing accountability scores from empirical data 27
5. Equating scores using a Partial Credit Model 27

Ethical Considerations 29

Results 30

1. Differences in statistical subject standards 30
2. Simulate effect on accountability scores 31

3. Imputation of student grades 36
4. Modelling factors influencing accountability scores from empirical data 37
5. Equating scores using a Partial Credit Model 39
Overall summary of Results 44
Discussion 45
Evidence for incentives driven by the performance measures 47
Subject specific ability and composite scores 49
Rasch equated scores 50
Are Progress 8 and Attainment 8 valid measures of school performance? 52
Limitations 54
Conclusion 56
References 57
Appendix A – Results from individual imputed score combinations 66
Appendix B – Rasch equated subject scores 67

Abstract

Composite scores are created in many education systems to combine the results from multiple assessments into a single measure of student (or school) achievement. In England, two composite scores are used as key school performance measures, Progress 8 (P8) and Attainment 8 (A8). Both of these scores make the implicit assumption that grades received by candidates in different subjects are interchangeable in terms of their relative grading standards or 'difficulty'. No method is currently employed to scale or equate scores across different subjects. This has led to concern that schools can be penalised by these measures by entering students for 'more difficult' subjects. This perceived 'perverse incentive' is potentially driving unwanted behaviour by causing schools to not enter students for certain subjects, most noticeably modern foreign languages (MFL).

This study aimed to evaluate if schools were being unfairly penalised by choosing to enter students to MFL subjects through the methodology currently used to calculate A8 and P8. A secondary aim was to evaluate if this bias could be effectively removed by equating subject scores using a partial credit model.

Analysis showed that using common statistical methods MFL subjects appear among the most 'difficult' at GCSE in terms of statistical grading standards. Utilising a simple score adjustment and a more complex simulation showed that this difference in difficulty could potentially lead to a small negative bias in performance scores for schools entering a large number of subjects to MFL subjects. Although a small effect, this bias could still have a meaningful effect on a schools relative ranking and performance banding. Imputation of student scores indicated that schools could potentially benefit by entering students for alternative subjects instead of MFL, as long as they could be included in the English Baccalaureate 'bucket' of performance measures but not if it was replaced by any other subject. However, despite this estimate negative bias, when real school results were modelled, entering students for MFL subjects appeared to have a positive impact on A8 and P8 scores. This may be due to schools compensating for the existing bias by entering students for MFL with a higher subject specific aptitude, potentially masking any negative effects in the empirical data. Trialling the application of a partial credit model suggested it could be used to effectively compensate for differences in subject grading standards. However, it would add significant complexity to the current method and would potentially introduce additional issues and complications.

It is concluded that the current methodology for calculating A8 and P8 may be driving unwanted behaviour in the system, potentially working against one of the key aims of the

government to promote a “broad and balanced curriculum”. However, as the performance measures are utilised both for school accountability and parental choice through ‘school league tables’, it is unclear which of these should be prioritised and therefore whether introducing a significantly more complex method would be beneficial. Other research suggests there are more pressing issues with these performance measures as they take no account of student background characteristics causing more serious systematic bias.

"*To compress all our information about a single student into a single ranking number is clearly absurd - quite ridiculously irrational. And yet it has to be done*." (Hoffmann & Barzun, 2003, p. 35)

Introduction

Assessments in education can be used for a variety of purposes. In full time education, in most education systems, students take assessments in a variety of subjects. In almost all cases, these assessments are used to give an indication of a student's knowledge, skills and abilities in that specific subject, often in comparison to other students. However, for progression or monitoring purposes often what is desired in addition to these individual subject scores is an overall picture of student ability. To achieve this a process is needed to somehow sum, average or combine results from individual subjects to generate an overall profile of a student, often reduced to a single score. In addition to this in some systems these individual scores of ability are further aggregated to give a school level measure for use for school monitoring or accountability. To aggregate scores in this way requires making a number of assumptions and subjective decisions (Leckie & Goldstein, 2019). Depending on the methodology applied this can lead to a variety of biases or distortions of the scores, which may lead to perverse incentives for students or schools to attempt to 'maximise' their score in a way which can distort the measure (O'Neill, 2013).

In the English education system, one of the key places this composite measure is created is for school performance measures. There is a belief amongst some stakeholders including school leaders and subject lobby groups, that the current methodology being employed to aggregate scores for students and then schools is unfair and introduces bias. As individual schools have some freedom around their policies for entering students to different subjects, there is a belief that schools can be unfairly penalised and disadvantaged by entering students for 'hard' subjects (Education Datalab 2015; Harris & Burn, 2011; Lanvers, 2017a; Titcombe, 2008). To understand the legitimacy of these concerns requires understanding of what is meant by subject difficulty, how different subject scores can be combined, and what these 'composite scores' actually mean.

Comparability of subject grades

Often subjects or qualifications are colloquially referred to as being 'easy' or 'difficult'. What is really meant by this was suggested by Kelly (1975) to be whether a subject is

'leniently graded relative to the consensus standard' or *'severely graded relative to the consensus standard'*. Where the same grading structure is used between different subjects within an assessment system there can be an implication of comparability in 'grading standards' and therefore interchangeability across different subjects when they are awarded the same grade (O'Neill, 2013). In some circumstances the comparability of subject grades is more explicitly assumed, such as school accountability measures indicating a required number of 'passes' or university admissions which require specific grades, but the subject/s required is not specified (Newton, 2010a).

Some have suggested that because of this implication, assessment systems should take all reasonable steps to ensure comparability between subjects on the same grading scale (Coe, 2010). For large entry subjects where students are adequately motivated and teaching quality is similar, like maths and English, some authors have proposed that there is no good reason why on average students shouldn't receive the same grade in both (Nuttall et al., 1974). Coe (2008) goes further to suggest that particularly in situations where grades are used interchangeably or as equivalent currencies, such as admissions to university and as part of school performance measures, then standards need to be explicitly statistically aligned. Coe (2008) argues that if this is not the case it can lead to 'perverse incentives' not intended by the system and potential distortions of the intended measures. Where students have a choice of options across subjects, or even different qualifications, the students, and their schools and parents, have an incentive to select the option where the grade is easier to attain and avoid subjects with higher levels of difficulty (O'Neill, 2013). This may be the case whether these differences are actual or only perceived (Dearing & King, 2007).

There is, underlying these concerns around subject comparability, an implicit assumption that different qualifications are designed to require similar levels of *intellectual demands* on students at the same grade (Schools Council Forum on Comparability, 1979). However, in many cases this assumption may be invalid where different subjects require vastly different skills or knowledge and the grading system does not take this into account (Newton, 2010a). On top of this, it is not clear what this imagined construct of 'intellectual demands' explicitly refers to (Schools Council Forum on Comparability, 1979). In many cases this potential fuzziness of interpretation is not directly acknowledged or addressed by those making these assumptions. How exam standards are conceptualised has potential knock on considerations for how student attainment is measured and particularly for how performance measures are calculated, as discussed further below.

What are exam standards?

Across different subjects or even qualifications we are sometimes interested if scores achieved in those subjects are 'comparable' and therefore if the qualifications are of the same 'standard' (O'Neill, 2013). What this most commonly means is whether the probability of the same student attaining the same grade is equal across different subjects (Newton, 1997). Sometimes expanded to the 'same-students' definition, which states that if the same *group* of students took two subjects they should *on average* get the same grade in each (Cresswell, 1996). However, definitions vary substantially, and subject 'standards' may relate to the value added relationship from a previous assessment, or incorporate other factors such as student motivation, or teaching quality, or refer to future success in a subject (Newbould & Hall, 1981). As highlighted by Baird (2007) what is meant by exam standards can also vary between professionals, academics and the general public adding to the complexity. Fundamentally an assessments 'standard' relates to how we interpret meaning from the grades or scores produced (Wiliam, 1996). Some common methods to quantify if a given standard has been reached involve comparing an individual result to the rest of the cohort taking the assessment (cohort referencing), a standardised reference group of students (norm referencing), or by defining some sort of threshold level of performance in terms of knowledge or skills displayed (criterion referencing; Wiliam, 1996). However, all of these methods have potential weaknesses, and as discussed by Newton (2010a) are more related to operational techniques than theoretical ideas of what standards are.

Cresswell (1996) discusses a 'catch-all' definition of exam standard. He states that two examinations can be considered as having comparable standards if "*two groups of students with the same distributions of ability and prior achievement who attend similar schools with identical entry policies, are taught by equally competent teachers and are equally motivated, receive grades which are identically distributed after studying the respective syllabuses and taking the examinations*" (p. 77). Unfortunately, in practice rarely do we have the exact same cohort taking two assessments and measuring all of theses additional relevant factors is essentially impossible (Baird, 2007). Therefore, the definition of standards applied by individual authors and publications can vary considerably (Bramley, 2005; Baird, 2007; Coe, 2010). Different stakeholders inevitably have different priorities and viewpoints and adopting a single definition of standards may not be practical or desirable (Baird, 2007). Due to this, there have been a number of attempts to categorise and rationalise the diverse definitions of standards and comparability used (Elliott, 2013).

Conceptions of comparability

To attempt to clarify the discussion of exam standards Newton (2010a) categorises different definitions of comparability as 'phenomenal definitions' (focussed on the present), 'causal definitions' (focussed on the past) and 'predictive definitions' (focussed on the future).

From the phenomenal perspective, students who are awarded the same grade must on average be equally as good at knowing, understanding or being able to do something common to both examinations (Newton, 2012). The causal perspective suggests that as long at the 'inputs' into each assessment are similar the grade or result should also be similar. For example inputs such as student prior attainment, teaching quality, quality of resources. Where inputs are equal between students and qualifications, even where the content of the assessments differ, then given this definitions of standards we would expect students to receive the same grade on each assessment (Newton, 2010a). Finally, the predictive definition suggests that students who receive the same grade should be equal in the terms in which their attainment predicts futures success. However, Newton (2010b) emphasises that one definition is unlikely to support all uses of a qualification. Instead he suggests that it may be possible to link qualifications in multiple different ways for different purposes, an idea which he elaborates upon further as the idea of 'construct comparability'.

Construct comparability

Newton (2005) originally described the idea of a 'linking construct' (further expanded on in Newton, 2010b) which can be applied to many of the more general ideas of comparability outlined above. Central to the concept is the idea that it is meaningless to compare two subjects unless there is some dimension in which the two subjects are 'the same' (Coe, 2010). Instead of independently considering that a student achieving below a grade C in psychology, is not as good as a student achieving above a grade C in psychology, and a student receiving below a grade C in sociology is not as good as a student receiving above a grade C in sociology. We could instead interpret both of these on a common scale of 'social research skills', which is potentially more useful for someone attempting to compare students across these qualifications (Newton, 2005). This requires the person evaluating the qualifications to make fewer assumptions and instead they can just consider the quantity of the 'linking construct' displayed in either qualification. This approach also works to make it clear on what basis standards are set and different subjects compared.

Given this interpretation, we can consider that "*two examinations have comparable standards if performances that correspond to the same level of some linking construct are awarded the same grades in each*" (Coe, 2010, p.278). To link standards across any set of assessments we need to first define what broader construct we are using to compare them. In many cases, what is defaulted to when comparing a wide variety of subjects is a theoretical linking construct of 'general academic ability'. Particularly in terms of GCSEs and A-levels in the UK this construct has regularly been referred to implicitly if not explicitly along the lines of 'general aptitude' (Fitz-Gibbon & Vincent, 1997; Newton, 2005; Coe, 2008). The 'difficulty' of the examination relates to the level at which it corresponds to this linking construct and when the same grades are awarded between subjects they can be interpreted as indicating the same amount of this construct. With this interpretation of standards, individual assessments can be viewed as different instruments for measuring the same construct and evaluated accordingly (Coe, 2010).

Measuring subject difficulty

So how do we go about measuring the 'difficulty' of individual subjects to decide if they are comparable in terms of our linking construct? Coe (2010) outlines the numerous problems that come when attempting to measure subject difficulty. Firstly, it has been argued that students' achievement across different subjects is affected by more factors than just the demand of the subject material (Newbould, 1982). Newton (2010a) breaks these influencing factors down into: student factors including (among others); prior attainment, learning capability, and time committed to learning. Teacher factors including; teaching proficiency, subject expertise and how engaging the teacher is. As well as resource based factors such as course materials and learning tools. Given all these factors, grades between different subjects are unlikely to be a direct reflection of the 'general academic ability' of students. Therefore, if this is the aim, any comparison should take into account all of these additional factors (Newton, 2010a). Coe (2010) also outlines a second problem in looking at relative subject difficulty and that is identifying if there is a suitable level of unidimensionality in what is being compared, essentially the salience of the linking construct in individual assessments (Newton, 2005). To compare assessments there needs to be confirmation that there is a substantial common element or any comparison will be meaningless. The third major issue facing studies of subject difficulty is that the only students available for statistical comparison are those that have chosen to take individual subjects. This becomes a problem if we want to generalise conclusions to a different or wider subgroup (Coe, 2010).

Statistical measures of inter-subject difficulty have been critiqued on a number of occasions (Goldstein & Cresswell, 1996; Newton, 1997). Often the most identifiable problem with these methods is that when the groups of students being used to compare subject difficulty are broken down into smaller subgroups (e.g. male and female) and then the analysis is repeated, a different result is found (e.g. Forrest & Smith, 1972; Nuttall et al., 1974; Kelly, 1976; Backhouse, 1978). For example, for male students it has been found that maths and English were of similar difficulty but for female students, English is more lenient (Newton, 1997). Calling into question the meaning of results. Newton (1997) also argues that all statistical measures have the same underlying assumptions. Namely that 'general academic ability' is a valid construct and that students' grades should be equated across subjects despite domain specific factors. Newton (1997) concludes that all statistical methods are inadequate for measuring 'errors' in grading standards by identifying subjects which are 'too harsh or 'too lenient', as the results are always clouded by other factors which are not taken into account.

Cresswell (1996) ultimately rejected the various statistical definitions of standards, instead settling on a social rather than technical definition. This definition relies on the value judgements of experts and their decisions on comparability. Judgemental methods involve comparing observable examination performances, with the aim to holistically link unobservable latent constructs. Baird (2007) discussed this as the 'conferred power' definition when the responsibility to determine the standards of assessments is placed in the hands of a limited number of experts. Judgemental methods do not need to make the same assumptions as statistical methods (Cresswell, 2000). Two qualifications are comparable only to the extent to which those who use the results of qualifications are prepared to believe they are equivalent (Wiliam, 1996). Given this 'value-based' definition of comparability there is no real need to confirm the existence of a unidimensional underlying trait, rather accepting that qualifications have been qualitatively judged to be comparable. However, these methods have their own limitations as they require judges to observe student 'attainment' through the 'foggy lens' of exam performance, which is not always a straightforward task (Cresswell, 2000). It also provides no guarantees over what criteria is being used to judge the relative standards of qualifications (Cresswell, 1996).

Interpreting evidence across subjects

So why is it important to decide whether grading standards are comparable between subjects? Most notably in cases when it is important to the person interpreting results from a diverse array of subjects to gain an idea of how well a student may perform in

general, rather than in specific subjects. Coe (2007) gives the example of a university admissions tutor assessing students on a diverse array of subject results, most of which are not the subject the student intends to study at university. In these situations, higher education personnel are often forced to make assumptions about subject standards when choosing students with a wide variety of subject outcomes. If these assumptions are made wrongly it risks advantaging or disadvantaging students based on the subjects which they took. Another situation in which this is important, which is the focus of this study, is when government performance measures aim to gain a holistic view of student ability and/or progress as a measure of school teaching quality. For this purpose, first measures of individual student ability are calculated, then in a further step these scores are aggregated across a school. In this circumstance again there is a risk that if scores across different qualifications are not equivalent, then a school risks being advantaged or disadvantaged based on the composition of qualifications taken by its students.

In both these cases, it would be useful to be able to accurately interpret the results from diverse assessments on a consistent scale. Unfortunately ensuring comparable standards during awarding is practically complex (Alton & Pearson, 1996) and can lead to individual subjects not discriminating well between students and therefore not meeting their primary objective (Coe, 2008). Therefore a more practical solution is through post-award scaling of results to take account of any potential differences in subject or qualification standard to give a fair representation of different schools or students. Utilising a linking construct such as 'general academic ability' as discussed by Newton (2010b) is one approach to achieve this. Spearman (1904) originally suggested that a general factor taken from a battery of tests based on school performance would be almost perfectly correlated with a scale of general intelligence. Recent studies have found evidence to support this; for example, Coe (2008) found that across 34 different and varied GCSE subjects a single latent trait accounted for 83% of the observed variation in results. However, the target construct can vary depending on the scaling method used (Coe, 2010). For example, subjects can potentially be compared based on a variety of dimensions such as: the demand of the assessments, the content and domain coverage, the performance of students, grading standards or the predictive power of exam results (Elliott, 2013). Although importantly these scales should take account of the 'amount' or 'level' of the linking construct displayed in the results from each of the different subjects included.

Once scaled, individual subject scores can be combined to give an overall measure of the target construct. 'Composite scores' such as these are produced in a number of assessment systems around the world and are primarily used for university entrance or

accountability purposes (Lamrianou, 2009). There are additional potential benefits to using composite measures over utilising a single test or score as general measures of academic ability. Taking scores from multiple testing can minimise the error variance in scores (Watkins et al., 2007). It can also minimise variance that could be caused from a single test related to motivation, comprehension or attention (Kaufman, 1979).

However, these combined scores tend to be second order purposes of assessments (O'Neill, 2013). The primary purpose relating to demonstrating ability in a subject, and only a secondary consideration is given to how these different measures might be combined. Unfortunately, issues often arise when assessments have an unclear purpose or are used for purposes for which they are not designed (Kelly & Downey, 2011). There is a risk that the scores and interpretation may be adjusted to fit the methodology employed for scaling and combining scores rather than finding appropriate methods that fit the appropriate purpose (Kelly & Downey, 2011). For example, ideas of 'general academic ability' fit most easily with statistical methods to assess standards, potentially explaining their proliferation.

Creating a composite measure

So given that guaranteeing inter-subject standards is difficult, if not impossible. How then do we go about reliably producing a single 'composite measure' of student ability? For any sort of composite measure to make sense there must be enough in common between different subjects for them to be meaningfully compared (Allen, 2014). If we believe all our subjects display our linking construct to the same degree then combining scores across subjects is straightforward, we could take an average or add up scores converted to grades. However if there is evidence of a difference in 'standard' relating to the construct we are trying to measure then more complex methods may need to be employed. As discussed above, most forms of composite measure apply some sort of scaling to account for the non-equivalence of different subjects (Lamprianou, 2009). The intended purpose of such scaling is usually to prevent there being any systematic advantage or disadvantage to different students based on their subject selection and to avoid the unfairness from just summing scores from different subjects (Allen, 2014). In this way, scaling methods are designed to improve fairness and to create a level playing field (Allen, 2014).

Coe (2010) suggests that if we are combining scores across assessments for a particular purpose, then it should be considered a 'variable conversion' problem akin to converting currency onto a consistent scale. The conversion rate is therefore variable

between subjects and depends on what scale it is being converted to and the interpretation for which it is being used. An important consideration, therefore, when forming composite measures is whether it produces a valid representation of the target construct, usually a measure of student (or school) achievement. Given modern validity theory, it is the adequacy and appropriateness of the inferences and actions based on the test data that is key to evaluating validity (Messick, 1989). Wiliam (1996) asserts that it is not necessarily that the measured construct exists within a formal theory that is important, but that there is enough agreement about the ratings for the purpose for which they are being used. Therefore, it is important that the interpretation of composite scores by end users is appropriate and that any issues or biases created by the scores is acknowledged.

Issues with composite measures

If composite measures are not calculated in such a way as to account for differences in the probability of achieving certain grades in different subjects then this can lead to distortions of the scale and potentially drive unintended behaviour in the system. Allen (2014) highlights a common issue by suggesting that when creating composite scores no one should assume that students taking different courses are essentially similar and argues that this is why some form of scaling is essential. Combining grades through simple averaging to obtain grade averages results in systematic bias against students enrolled in more difficult subjects (Johnson, 1997). Unfortunately, this is often exactly what is done to create composite scores, particularly in the UK (discussed in detail below).

Lamprianou (2007) indicates that this issue is not restricted to the UK and in education systems across the world where exam scores are converted to a single ‘currency’ then perverse incentives, either real or perceived, develop. Composite scores are often used at the level of entry to tertiary education, where students perceive that taking ‘difficult’ subjects potentially limits their chances. In many systems, therefore, statistical adjustments or models are used to reassure the public and compensate for differences in difficulty. However, this does not mean that these systems of scaling are problem free and often produce perverse incentives of their own (see Lamprianou, 2007 for examples). Due to the investment in results by multiple stakeholders, issues with these types of composite scores inherently become political issues (Lamprianou, 2007). Therefore any method will fail without first reaching the consensus of stakeholders (Lamprianou, 2009). Lambrianou (2009) suggests that simple and transparent solutions are often more beneficial than complex statistical models in terms of stakeholder

engagement. Unfortunately, the majority of methods of scaling are rather opaque and difficult for the non-specialist to understand, which leads to a variety of beliefs, either supported or not, about unfair effects (Lamprianou, 2009). Due to this, Lampianou (2009) asserts that the general public find it easier to buy in to the idea that all subjects are of some comparable level of difficulty than to trust in complex statistical fixes.

Rasch Equating

A number of more recent methods of creating composite scores have been based on item response theory (for example Australia, Allen, 2014; Tasmania, TQA, 2007), first proposed by Tognolini and Andrich (1996). The aim of these methods is to compensate for the relative difficulty of different courses or exams and the different ability levels of students when aggregating scores into a composite measure of student ability (Johnson, 1997; Young, 1990, 1991). The idea of using a Rasch model to scale different subjects to account for assessment difficulty has been proposed on a number of occasions as a possibility in the UK (Coe et al., 2008; Coe, 2010).

To apply the Rasch model to subject difficulty each subject score is treated as a polytomous item. Each grade is treated as a score on that item, representing ordered response categories. All of the subjects are then treated as items within a single test (He et al., 2018). This way a scaled score can be given to each student taking account of which subjects they took and their relative difficulty. Key to the use of the Rasch model in this context is the assumption that the data are unidimensional, there is a single trait underlying the different difficulty of items, or in this case subjects. However, Tognolini and Andrich (1996) argue that the assumption of unidimensionality is already inherently being made in most systems every time scores are aggregated, the Rasch model just makes it explicit. Daley (1989) asserted that any moderately correlated set of scores in different subjects will be dominated by a single factor which will be aligned with the first principal component of these scores. The latent trait specified by the Rasch model is therefore operationally defined by the subjects included in the analysis (He et al., 2018). After carrying out a Rasch model of GCSE results, Coe (2008) interpreted the trait as a measure of ‘general academic ability’ of examinees included and the subsequent difficulty of each assessment on this scale.

Critique of the Rasch for creating composite scores

There is particular contention about the impact of missing data on the results of the Rasch analysis when applied to create composite scores, as most systems have some element of subject choice. As all students will only take a subset of subjects there will

inevitably be ‘missing results’ included in the data used for analysis. Korobko et al. (2008) states that this violates the ignorability principle (Rubin, 1976) and biases the results. Supporting this, Bramley (2016) used simulations to show that subject selection, i.e. the non-random missingness of data, can produce biased estimates for statistical measures of subject difficulty. Bramley’s analysis showed that if students choose to study subjects they are best at then it becomes difficult or impossible to get unbiased estimators of subject difficulty. Bramley (2016) found that subjects that are less correlated with other subjects appear easier, whereas more highly correlated subjects are evaluated as being closer to the mean subject difficulty. This is due to students optionally choosing subjects that they have an aptitude for beyond the general academic aptitude accounted for by the model.

However He et al. (2018) critique this analysis by suggesting the nature of Bramley’s simulated data set did not accurately represent the true picture as it involved non-randomly excluding data points of known value. He et al. (2018) argue that for exam results where students have a choice of assessment, missing data can be considered as missing by design from a data collection perspective, making it similar to a test equating design that involves the use of multiple common item non-equivalent groups (CINE), such as that presented by Kolen and Brennan (2014). Others have argued that it is exactly this ability to deal with missing data that makes the Rasch model so valuable to comparing subject difficulty (Ofqual, 2015). Therefore, the effect of the missing items would be minimal to the estimation of Rasch difficulty.

Composite measures in England

There are only a few practical instances within the English educational system where composite measures are used. However in all of these cases the, likely erroneous, assumption of between subject comparability in statistical difficulty is incorporated. The focus of this study is on the use of composite scores as measures of school performance.

Performance Measures

Performance measures, used as indicators to hold schools to account of their educational value, are controversial in their proposed value. There is some evidence that school performance measures improve pupil outcomes in terms of measured test scores (e.g. Burgess et al., 2007; Chiang, 2009). Although evidence is the UK is limited (NFER, 2018). However, there is also evidence that performance measures can also lead to potentially undesirable responses from schools, such as teaching to the rating or trying

to game the system (Ooghe & Schokkaert, 2009). This can be particularly exacerbated when the key aim of performance measures is unclear; in the UK performance measures can be considered to have three main aims, to aid parents in making informed school choices, so schools can be held accountable by the government, and to help shape behaviour across the system (Menzies & Jerrim, 2020). Schools also increasingly use the same measures for internal purposes such as resource allocation, self-evaluation and target setting (Leckie & Goldstein, 2019).

How to scale and combine scores to create composite measures for accountability can have a significant impact on school ratings and therefore school behaviour and so should not be left as a secondary concern (Leng & Koretz, 2015). Studies that have looked at different measures of combining schools for performance measures have found that although they are often highly correlated, they can result in significantly different rank orders (Briggs & Week, 2009; Leng & Koretz, 2015). Policy makers have to make a series of non-trivial decisions when deciding on how they will be computed (Guarino et al., 2019). Many of these decisions are subjective and are based on assumptions, which require robust evaluation (Leckie & Goldstein, 2019). Significant consideration needs to be given to what behaviours are being driven, then the measures should be adapted to meet these objectives (Neal, 2008). Particularly important is that schools should only be held responsible for the factors that are under their control (Ooghe & Schokkaert, 2009).

Progress 8 and Attainment 8

In England performance measures are published in 'school league tables' for all government funded secondary schools since 1992 (DfE, 2016a). These measures are used to both inform inspections and provide public information to aid parent and student school choice (Leckie & Goldstein, 2017). Where schools fall in league tables of school performance measures has been shown to affect parents' choice of school (West, 2010) and the likelihood that teachers will apply for a job (Noden, 2001). Unfortunately the 'high stakes' attached to these performance measures can inevitably lead to a distortion of the behaviours seen in the system, both intended and unintended (West, 2010).

Current school performance measures at GCSE level (typically 16 year-old students) in England include two headline measures of 'Attainment 8' (A8) and 'Progress 8' (P8) scores used for accountability. A8 and P8 were introduced as key performance measures for schools in England in 2016 (DfE, 2020a). Compared to other potential options (such as value-added multi-level models; Leckie & Goldstein, 2011), these are relatively simple measures. A8 is essentially a total score across eight traditionally

academic subjects, giving an indication of pupil performance at the end of secondary education, age 16. These subjects need to fill eight subject 'slots' split into four 'buckets' which can only be filled by certain subjects, one for maths, one for English, one for 'Ebacc' subjects (typically academic subjects included in the English Baccalaureate; DfE, 2020a) and one 'open group' which can be filled by Ebacc qualifications or other approved subjects (Figure 1). The maths score is double weighted and if both English language and English literature are taken, the best of these scores is also double weighted. If one of the slots is left empty it is counted as a zero score, creating a significant penalty. Subject scores within a bucket are therefore considered interchangeable in terms of the level of attainment they represent.

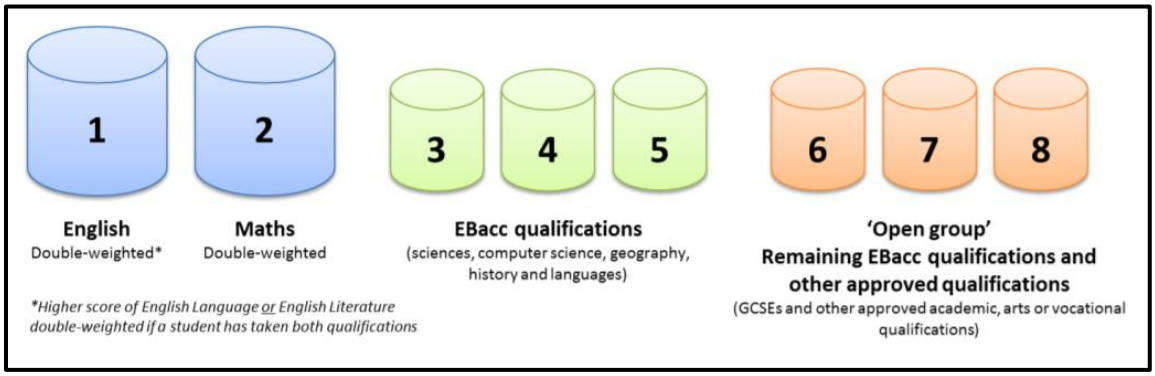


Figure 1. Diagram indicating subjects included in the calculation of the Attainment 8 measure. Image reproduced from DfE (2014).

P8 takes a value added approach, with an aim to evaluate student improvement between the end of primary education (KS2) and GCSE examinations. Each pupil's score is their A8 score minus the average A8 score for pupils nationally with a similar prior attainment score from key stage 2 (KS2) assessments taken 5 years earlier at age 11 (categorised into 34 distinct prior attainment groups). Schools level scores are simply calculated as the average of all pupils' A8 or P8 scores (DfE, 2016b). Although with some slight adjustment to avoid extreme negative student scores overly impacting a school's score (DfE, 2020a). P8 was designed to be a fair measure for schools, in that it would take into account the prior attainment of pupils and schools would not be penalised based on the prior attainment of their intake (Burgess & Thomson, 2013).

The stated aim of these measures is to "*to encourage schools to offer a broad and balanced curriculum with a focus on an academic core at key stage 4, and reward*

schools for the teaching of all their pupils" (DfE, 2016b, p2). Schools are labelled as underperforming if their school falls below a certain minimum "floor standard" (DfE, 2020a). This prompts increased scrutiny by Ofsted, the national schools inspectorate and local authorities. Whereas schools with the highest P8 scores are exempt from routine inspection in the following year (DfE, 2020a). Both of these are strong motivators for schools to do well in these measures.

Critiques of performance measures in England

Value added measures such as P8 have met with much controversy in England over time (Gorard et al., 2012). This has stemmed from questions around whether they meaningfully reflect interschool differences. These concerns have come from a variety of places including whether they appropriately account for confounding variables including socio-economic status (Wilson & Piebalga, 2008). Wilkinson et al. (2018) showed that the variance in value added scores due to schools is only around 9%. A large proportion of the rest of the variance is due to uncontrollable factors including pupil characteristics such as demographics and socioeconomic status (Reynolds et al., 2014; Teddlie & Reynolds, 2000). This results in the effect that schools with more 'educationally advantaged' intakes generally show higher P8 scores than those with more disadvantaged intakes (Leckie & Goldstein, 2019).

O'Neill (2013) produced a detailed discussion of school performance measures in the UK and the potential pitfalls to be avoided. He likens poor systems of accountability to 'Goodhart's Law', originally used to discuss monetary policy, which states *'Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes*' (Goodhart, 1984, p. 96). Essentially that as soon as we start to measure something, we start to distort it. O'Neill (2013) calls for a scheme of 'intelligent accountability' where educational objectives are made explicit in any measures created. In relation to current practices, O'Neill (2013) is adamant that we cannot have a combination of optionality and a fiction that all grades are interchangeable. O'Neill claims that all schools, students and universities are aware that grades in GCSEs and A-levels are not interchangeable, but when points are summed this is more due to a deference to performance measures than a belief that they are equivalent. In a system where it is 'riskier' to take some subjects over others then it is not surprising that some students or schools will select the 'easier' and safer option, even potentially at the expense of less educationally valuable subjects and educational detriment (O'Neill, 2013).

The idea of summing scores without any sort of scaling, as is done for A8 and P8, has been critiqued by other nations, for example the Cyprus Ministry of Education has said that aggregating raw scores would lead to worse side-effects and problems than using scaled scores (Cyprus Minister of Education, 2006). However, as a counterpoint Kelly and Downey (2011) argued that any measure used for accountability purposes needs to be understandable by those in the system without requiring outside expertise, limiting the level of complexity possible.

Evidence of perverse incentives in performance measures in England

In the UK the concern of the comparability of grading standards across different subjects has been ongoing for a number of years. In 1996 Sir Ron Dearing published a review highlighting concerns that GCE (A-level) subjects, typically taken by 18 year old students for university entrance, did not have a consistent standard and some subjects (including individual sciences, French and history), were more difficult (Dearing, 1996). O'Neill (2013) subsequently stated that it is '*a fiction that treats scores on the assessments of different options as using a common unit of educational quality*' and that '*it is not in the least plausible to assume that all instances of a given score at GCSE or at A-level have the same educational weight or significance*' (p. 15). He goes on to state that schools and students '*take for granted that different subjects make demands of varied difficulty, and that subjects will be chosen partly with this in mind*' (p. 15). In addition to this, Ofqual, the qualifications regulator in England have publicly acknowledged that there are differences between subjects in grading standards as measured using statistical methods (Ofqual, 2019). Likely to have occurred as the comparability within subjects over time has been operationally prioritised in England over comparability between subjects (Baird, 2007). Despite this, performance measures in England do not take account of between subject differences in grading standards in the way they are calculated.

The argument about inter-subject comparability has been recently further pushed forward by subject lobby groups, particularly around languages (e.g. Myers 2006; Dearing & King 2007), mathematics and the sciences (e.g. Coe et al., 2008; Royal Society, 2008). O'Neill (2013) and others have stated their belief that in England the decline in entry to modern foreign language (MFL) subjects is, at least in part, due to the non-comparability of subjects and the potential 'harsh grading' of MFL. MFL subjects became non-compulsory for students at GCSE in 2002 and since then there has been a steady decline (until recent years; Tinsley, 2019). This has been referred to a 'language learning crisis' in English schools (Lanvers & Coleman 2017; Tinsley & Board 2017a,

2017b). Due to pressure on schools to improve their position on 'school league tables' there is a believed disincentive to enter students for MFL subjects (Education Datalab, 2015; Harris & Burn, 2011; Lanvers, 2017a; Titcombe, 2008).

There is some statistical evidence supporting this belief, suggesting that obtaining a 'C' grade in MFL subjects is more difficult than all other subjects except the individual sciences and statistics (Coe, 2008; He at al., 2018). Due to this potential difficulty Coe (2008) concludes '*it is hard to believe that such considerations will play no part in the choices that are made. For example, for the majority of students, taking media studies GCSE instead of German is associated with a result that is a whole grade higher*' (p. 632). Tangential evidence of the effect of the P8 performance measure on uptake of MFL is provided by Gill (2018) who shows that there is a decline in MFL uptake and that it is usually only the MFL 'pillar', which is empty in the 'English Baccalaureate' (Ebacc) performance measure. The Ebacc is a simpler measure of school accountability, which is calculated as the proportion of students in a school entered for all of the Ebacc subject 'pillars', consisting of English, mathematics, science, humanities and languages (DfE, 2020a). This suggests that although there are incentives through the Ebacc performance measure for schools to enter students for MFL subjects, there is enough of a belief in the 'difficulty' of MFL subjects and the potential disadvantage this leads to, for schools to leave one of these 'pillars' empty (Gill, 2018).

Importantly for the consideration of performance measures, the entrance policy for language qualifications is left in the hands of individual schools. There is no requirement in England to teach a language to children over the age of 14 (Parrish & Lanvers, 2019). This results in schools having different policies, such as some requiring all students to take a language aged 14-16, whereas other may have it as an option or split students into different 'streams' taking different combinations of subjects based on ability. In only around 50% of schools are all students given a free choice whether to study a language subject (Parrish & Lanvers, 2019). Higher ability students are often those encouraged to take a language (Education Datalab, 2015; Lanvers, 2017b). Evidence suggests that independent schools and those with a better academic record are more likely to enter students for language subjects (Lanvers, 2017a). Due to this particular focus and evidence around MFL subjects, it serves as a useful case study to evaluate whether this behaviour is justified and driven by the method currently used to calculate A8 and P8.

Aims/Rationale

The aim of this study is to evaluate if there is evidence for schools being disadvantaged by performance measures when entering students for certain subjects, essentially representing a 'perverse incentive' created by the current accountability system. The focus of this study is on GCSE level performance measures used to provide evidence of relative school quality (P8 and A8). Analysis is focussed on modern foreign language (MFL) subjects (namely; French, German, Spanish) as these are often the focus of discussions by stakeholders around GCSE subject difficulty and preliminary analysis identified these subjects as being among the most difficult in terms of relative statistical grading standards (see Results).

The first objective is to identify subjects that differ substantially in their grading standards and therefore can be considered 'more difficult' using common statistical approaches. The second objective is to identify what impact entering students to these statistically more difficult subjects is likely to have on school accountability scores, both through simulation and modelling of real data. The final objective is to examine if applying the Rasch equating method can be used as an effective method to counter any negative effects of subject difficulty on the calculation of accountability scores.

Method

Data

The key dataset used for the analysis was the National Pupil Database (NPD) from 2017 and 2018. The NPD is a dataset collected by the Department for Education yearly and contains details of student results at A-level (typically aged 18), GCSE (typically aged 16) and key-stage 2 (KS2; typically aged 11), as well as containing other student characteristics such as age, gender and indicators of socio-economic status. The NPD also contains school-level information such as the type of schools (maintained, independent etc.) or entry policy (selective or not).

Data for all analyses was restricted to students that were included in the performance measure calculations in the year of the dataset (2017 or 2018) and were in maintained, mainstream schools, and so excluded special and independent schools. Only full GCSE qualifications that were included on government performance tables and were awarded in summer 2018 were included in the analysis. Double awards (e.g. combined science) were counted throughout as two qualifications. Where possible, analysis focussed on reformed GCSEs (rolled out since 2017 and are on a 9-1 grading scale) as this allowed consistency in score conversions. Discount rules were also applied, as used in performance tables, so only the first attempt at each qualification was included in analyses (i.e. resits were excluded).

Analyses

Trying to evaluate the effect of subject choice on performance measures is complex because it requires attempting to evaluate the 'counterfactual' situation where individual students had been entered for a different combination of subjects. Given the data available it is impossible to account for all of the variables which lead a student to take a certain selection of subjects, particularly whether to take an MFL subject or not, and subsequently what grade would they have received if they had taken a different subject.

To try to unpick these complexities five strands of work were carried out as part of this study. Strand 1 estimates the relative difficulty of different GCSE subjects to confirm the effects seen in previous studies using similar statistical methods and provide an estimated magnitude of the difference in difficulty between different subjects. Strands 2 and 3 represent different approaches to estimate the likely effect of these differential subject difficulties on performance measures. Strand 2 uses the estimated grade difficulties to create a simple and a more complex simulation of the impact of adjusting

MFL scores to be similar to an 'average' subject. Whereas strand 3 utilises score imputation to estimate what grade individual students would have got had they taken a different subject instead of MFL, utilising data on grades achieved in other subjects and other student and school characteristics. Strand 4 models the empirical data on performance measures from 2018 to calculate the effect of school subject entry policy on accountability scores, particularly if schools enter a larger number of students for MFL subjects or are less selective about which students they enter in terms of their prior performance, while controlling for other school characteristics. Finally, strand 5 evaluates whether equating subjects via Rasch modelling is an effective way to circumvent the issues identified in the current method for calculating performance measures. All analysis was carried out in R version 3.6.1 (R Core Team, 2019).

1. Differences in statistical subject standards

A series of statistical measures were applied to evaluate the relative difficulty of different subjects. By difficulty in this case we refer to the likelihood of the same student gaining a particular grade in different subjects in the same year. Data on GCSE results from the NPD 2018 were used to create three measures of relative statistical difficulty. For each of the measures all full GCSE qualifications within a subject area were combined, to give a difficulty measure at subject level rather than for individual qualifications. Subjects are defined using the DfE performance table discount codes included in the NPD, as these qualifications are considered equivalent for performance measures.

The three measures calculated were:

Absolute difference in mean grade – for this measure a point score was applied to each GCSE grade (as is used in performance measures), for 9-1 grades this is simply the grade taken as the numeric value. A simple mean score for each subject was then taken across all students taking each subject.

Relative difference in mean grade – A point score is applied for each grade as above, then for each grade a student achieves the difference between that score and a mean of all other GCSE scores that student achieves was calculated. The mean of these differences was then taken for each subject across all students.

Kelly's method – The above method has previously been criticised as not accounting for subject selection. If students who take a hard subject also tend to take other hard subjects then this will skew the measure. To counteract this, Kelly's method (Kelly, 1976) takes an iterative approach. Starting by calculating the 'relative difference in mean grade' measure above. This mean difference from other subjects is then deducted from each

score in that subject and the analysis is run again with the new scores. This is repeated until the relative different scores converge and cease changing with each iteration; these scores are then used as the difficulty measures.

2. Simulate effect on accountability scores

Two approaches are used to simulate the effect of relative subject difficulty on A8 and P8 scores.

Simple Score Adjustment

Initially a simple score adjustment was carried out by adding 0.5 or 1 grade points to each MFL (French, German and Spanish) subject result to see the impact of this crude grade alignment based on the size of the difference in difficulty between MFL subjects and 'average' subject difficulty. This is similar to an adjustment to grade scores previously proposed to 'align the difficulty' of MFL subjects with other subjects (Education Datalab, 2019). The original and adjusted scores in each case were then used to calculate A8 and P8 scores at a student and school level. Due to computation restrictions, this analysis was carried out on a random sample of 500 schools.

Summary results of the above analysis are presented at student and school level. For students this is divided between students who did and did not take an MFL subject. At school level schools are split into three bands: those with a low entry to MFL subjects (less than 20% of students take an MFL subject), medium entry (20-80%) and high entry (greater than 80% of students entered.

Full Simulation

A second more complex simulation approach was utilised based on empirical data from GCSE results in NPD 2018. Due to the interrelatedness of some of the school and student level factors, which may have an impact on successfully interpreting the results from the simple analysis above, more variables are included here. Simulation also has the benefit of being able to adjust some of the input variables to identify what impact these have on schools scores, such as the size of the school or the mean prior ability of pupils (ability level on intake). Data for simulated distributions was taken from NPD 2018. Summary outcome data was created based on key-stage 2 result groups, which are used to calculate P8. For each group summary statistics were calculated including the number of qualifications in each subject taken and average outcomes in those subjects. The simulation was carried out for 1000 simulated schools. The simulation is calculated as follows:

i. For each simulated school a school size is selected from a truncated normal distribution (mean-150, sd-60, min-5, max-550), along with a mean prior attainment score (KS2 score; mean-4.75, sd-0.16, min-0, max-6) and a standard deviation of KS2 scores (mean-0.65, sd-0.08, min-0.1, max-1.5).
ii. For each simulated student within each school a random KS2 prior attainment band was selected (from the 34 available bands used of the calculation of P8), using the school mean and SD to produce a truncated normal distribution (min-0, max-6).
iii. For each prior attainment band, values were drawn based on empirical data used to create discrete probability distributions of number of qualifications taken in each progress measure 'bucket' (maths, English, Ebacc (excluding MFL), other (excluding MFL). The score each student achieved in qualifications in that bucket is then drawn from a normal distribution based on mean and SD for that prior attainment group from the NPD. MFL scores were then calculated for each student by sampling from a distribution including the mean and standard deviation of other grade scores, then reducing that score by 0.5 or 1 grade, simulating how much more difficult they are estimated to be than average.
iv. A8 and P8 scores were calculated for each student either including the calculated MFL grade scores (simulating if the student 'took' an 'MFL' subject which was known to be more difficult than other subjects) or replacing MFL grade scores with a mean of other grade scores (simulating if the student took an alternative subject of average difficulty).

Analysis of the simulation used a series of descriptive statistics as well as linear models to evaluate the relationship between various student and school level variables and A8 and P8 scores. The focus was on the difference in scores at school level when MFL grades are included or not. Analysis looked at whether including MFL subjects produces a positive or negative change in scores and how these changes correlated with factors such as school size, school mean prior attainment, average number of qualifications taken, the proportion of students taking MFL within a school and the prior attainment of those taking MFL compared to the rest of the students in each school.

The above analyses were repeated after restricting students simulated to only those filling all the A8 'buckets'. This is both because the effect of taking a harder subject may differ depending on whether the buckets are all filled, but also for a technical reason as the analysis is skewed towards those not filling all the buckets as no control is put in place to represent realistic subject selection at an individual level.

3. Imputation of student grades

Another approach was taken to evaluate the impact of students taking or not taking MFL subjects by attempting to impute the grades students might have achieved had they taken alternate subjects. To do this, students were first identified in NPD 2018 who took similar combinations of subjects but who either took an MFL subject or an alternate subject. Analysis was restricted to those that took at least 8 subjects and were entered for English language, English literature and maths. This is to avoid issues where students did not fill all of the buckets available, which significantly lowers student scores and is not the primary focus of this study.

For these two sets of students, a multinomial logistic regression model was used to estimate the grade they would have received in the alternate subject, while controlling for a wide selection of student and school level features. Finally, the difference between these grades was calculated and the impact on progress measures estimated at the student level. The process is outlined in more detail below:

i. For students taking an MFL subject and those that did not take an MFL subject in NPD 2018 the most popular subject combinations were ranked from most common to least. For this analysis only full GCSEs taken in summer 2018 and included in performance tables were used.
ii. For each MFL combination starting from the most frequent, the highest ranking alternate subject combination not including MFL was identified. Only combinations taken by over 500 students were used to ensure estimation accuracy.
iii. For both the MFL combination and matching non-MFL combination a multinomial logistic regression model was built predicting the grade in the MFL subject or matching alternate subject achieved by the student. A large number of predictor variables were included to attempt to control for student and school characteristics. Variables included in the model at student level; grade in each other subject taken, number of GCSEs taken, number of other level 3 qualifications taken, KS2 score, DOB (by quarter of a year), free school meal eligibility, EAL (English as an additional language) group, Ethnic group, SEN (special education needs) status, gender, IDACI (Income deprivation affecting children index) score. School level; school type, number MFL subjects offered (1-3), whether Latin is offered, school size, mean school KS2 score, school intake age, max pupil age, proportion of students taking an MFL subject, region of country, urban/rural, mean KS2 score of students taking MFL, proportion of students entered for all Ebacc subjects, mean IDACI score.

iv. Each model was then applied to the alternate group to predict MFL/alternate subject grade scores.
v. Model fitting measures were calculated (percentage prediction accuracy) and the difference between MFL and alternate scores are calculated along with the impact that this would have on P8 and A8 scores at an individual level.

4. Modelling factors influencing accountability scores from empirical data

All of the above approaches provide a theoretical impact of taking MFL on accountability scores. The following models were used to identify if there is a direct identifiable impact of schools entering students to more difficult MFL subjects on A8 and P8 scores in empirical data. This is approached in three ways:

1) A multi-level linear model of P8 and A8 scores at student level (with a random school level factor) was created to attempt to identify if entering an MFL subject on average leads to lower scores once other available student and school level features are controlled for. Variables included are the same as outlined above for the imputation approach.
2) A linear model of P8 and A8 scores at school level including a large series of school level variables, focussing on the impact of the proportion of students entered for MFL subjects, the prior attainment of students entered for MFL subjects and the prior attainment of students entered to MFL relative to the rest of the school population. Other variables included: school size, urban/rural, oldest year group in the school, youngest year group in the school, whether the school offers Latin, school type, number of MFL subjects offered, mean prior attainment, proportion of students filling all their Ebacc slots, proportion of students filling all their 'other' slots, mean student age, mean IDACI score, proportion FSM students, proportion EAL students, proportion SEN students.
3) A linear model similar to the above but variables were calculated as the change in school level features between two years (2017 and 2018) to see if these were related to a change in P8 and A8 scores between those two years. This analysis was focussed on whether a change in MFL entry policy in terms of proportion of students entered or prior attainment of student entered had a noticeable effect on accountability scores once other school level factors are accounted for.

5. Equating scores using a Partial Credit Model

The key issue with the A8 and P8 performance measures is the assumption that the same grade attained in different subjects are equivalent. This is operationalised through

the point scores applied to each grade, which are then summed together. An alternative to this approach has been suggested on a number of occasions (He et al., 2018; Coe, 2008). This alternative involves equating subject difficulties using a Partial Credit model, a variant on Rasch modelling which allows multi-mark questions. In this instance, each level of response represents a grade and each question represents a subject. These difficulty estimates can then be converted to point scores, which can be used to allow different subjects to be of different difficulty and for that difficulty to vary across the grade range. Rasch modelling assumes that the difficulty of items (in this case grades within each subject) can be measured on the same scale as student ability. The probability of a student achieving each grade is therefore the difference between the difficulty of the grade and their person ability estimate (Coe, 2008).

To operationalise that here a partial credit model (PCM) was applied to GCSE results data from 2018 to obtain difficulty estimates. The model used here closely follows that used by He et al. (2018) and Coe (2008). Each grade in each subject was converted to a point score with the lowest grade being given a 1 and each grade above being given a higher value by 1. Grade 'U' (the lowest grade) was not included in the model, partially as regardless of estimated difficulty most scoring methods give students 0 when attaining a 'U' and partially because previous analyses have found trouble with fitting a PCM to U grades (He et al., 2018; Coe, 2008). Following He et al. (2018), subjects with less than 1000 students and students taking less than 3 subjects were removed from analysis. The R package 'TAM' was used to run the PCM and produce model fit statistics. The fit of the data to the PCM was evaluated by looking at infit scores for each subject and a PCA was ran on the residuals of the analysis to evaluate if different subject scores could be considered to be unidimensional.

Person ability estimates could potentially be calculated directly from the model as a measure of student ability. However as not all students are included in the model and this would be a large divergence from current procedure, instead the item difficulty estimates were used to apply a point score on a similar scale to that used currently, to each grade. Difficulty estimates were converted to a standard point scale by standardising to a mean of 5 and standard deviation of 2.5 and rounding scores to one decimal place. Grades with a score lower than 0.5 were arbitrarily rounded up to 0.5 for face validity reasons. The potential impact of this score conversion was then evaluated by repeating the first part of this analysis and looking at subject level difficulty estimates to see if this had reduced the difference in difficulty between subjects.

Ethical Considerations

There are minimal ethical concerns for this research as it was carried out using entirely secondary data. All data sources were anonymised prior to use so students were not identifiable within the data sets and no student or school level data is reported.

Results

The results of each of the five strands of analysis are presented below, followed by a brief summary of the overall analysis.

1. Differences in statistical subject standards

The results of the four statistical methods of calculating subject difficulty are shown in Table 1. The analysis shows some variation in the rank order of subject difficulties using the different methods. Taking just mean grades, classical Greek and Latin have the highest mean grade, which may suggest these are the easiest subjects. However, this is unlikely to be the case as these subjects tend to be taken by high performing students, suggesting this measure is affected by subject selection. The difference from mean and Kelly's method give a much more consistent and intuitive result, which aligns with other sources of evidence (e.g. He et al., 2018; Tinsley, 2019). Both these methods suggest that the statistically hardest subjects are French, German, Spanish, Computer Science and classical Greek (although the sample size is very small for Greek). The analysis suggests that students receive a grade around 0.7-1 grade lower in these subjects than the average of other subjects.

Table 1. Results of subject difficulty calculations. Sorted by Kelly's method estimates from most difficult to least.

Subject	N Students	Mean Grade	Diff from mean	Kelly's Method
Greek (Classic)	114	6.92	-0.86	-1.07
German	38309	4.84	-0.91	-1.00
French	103493	4.68	-0.79	-0.86
Computer Science	63431	4.36	-0.72	-0.81
Spanish	76379	4.73	-0.69	-0.74
Latin	2819	6.75	-0.45	-0.56
Music Studies	29161	5.13	-0.24	-0.28
History	223460	4.58	-0.27	-0.28
Geography	212446	4.54	-0.24	-0.25
Chemistry	138828	6.01	-0.02	-0.08
Physics	137986	6.03	-0.02	-0.08
Biology	139739	6.07	0.06	0.01
Sports Studies	75614	4.81	0.07	0.08
Science Double Award	335235	4.07	0.05	0.08
Maths	477886	4.66	0.08	0.09
Community Development	12725	4.51	0.07	0.10
Speech & Drama	48847	4.90	0.09	0.11
Dance	7560	4.83	0.09	0.12
Religious Studies	190032	5.01	0.13	0.14
English Language	467178	4.74	0.15	0.16
English Literature	456781	4.78	0.17	0.18
Food Technology	42102	4.40	0.27	0.33
Art & Design	111839	4.96	0.39	0.46
Photography	22915	4.95	0.85	1.01

2. Simulate effect on accountability scores

Simple score adjustment

Table 2 shows the effect of the simple statistical adjustments on A8 and P8 scores at a student and school level. The adjustments included adding 1 or 0.5 to all MFL scores as a crude statistical adjustment to align MFL difficulty with other subjects. Student level results are split between those that did and did not take an MFL subject. School level results are divided based on the proportion of students entered for MFL subjects in those schools.

Table 2. Results of simple statistical modifications on Attainment 8 and Progress 8 scores showing mean scores and standard deviations in brackets.

	Student Level		**School Level**		
	Taking MFL	**Not-taking MFL**	**<20% MFL (N=60)**	**20-80% MFL (N=344)**	**>80% MFL (N=57)**
Original A8	60.0 (16.3)	42.7 (17.4)	42.65 (5.92)	49.59 (6.90)	60.76 (12.34)
+0.5 MFL adjusted A8	60.3 (16.2)	-	42.70 (5.93)	49.75 (6.90)	61.04 (12.29)
+1 MFL adjusted A8	60.7 (16.1)	-	42.76 (5.93)	49.92 (6.90)	61.35 (12.23)
Original P8	0.36 (1.13)	-0.31 (1.22)	-0.44 (0.36)	-0.03 (0.40)	0.42 (0.36)
+0.5 MFL adjusted P8	0.38 (1.13)	-0.32 (1.22)	-0.45 (0.36)	-0.03 (0.40)	0.43 (0.35)
+1 MFL adjusted P8	0.40 (1.13)	-0.34 (1.22)	-0.46 (0.36)	-0.03 (0.40)	0.44 (0.35)

The smaller adjustment adding 0.5 to MFL subject scores, for students taking at least one MFL subject, on average only increased A8 scores by 0.3, the larger adjustment of adding 1 increased scores by 0.7. This figure is lower than the absolute adjustment due to the score not counting for all students if they took more than eight subjects and had better grades in other subjects than MFL. At the school level, where the majority of students are taking MFL (over 80% entry group) this translates to a 0.3 point increase in A8 scores when 0.5 is added to MFL scores and 0.6 point increase when 1 is added to MFL scores. For more average schools with between 20% and 80% of students taking MFL this equates to around a 0.2 increase when 0.5 of a grade is added to MFL scores and a 0.3 increase when 1 is added to MFL scores.

For P8 scores the pattern is more complex. Due to the way P8 is calculated an increase in scores for those taking MFL results in a decrease in scores for those not taking MFL. Even increasing MFL scores by 1 to schools where the majority of students take MFL only leads to a 0.02 increase in P8 scores, much smaller than the standard deviation in scores of 0.35.

Full simulation

To evaluate how realistic the output of the full simulation was to empirical data, distributions of the key variables, particularly A8 scores were evaluated (see Table 3). When including all students from the simulation (not restricting to students filling all A8 'buckets'), simulated students taking MFL had a mean score of approximately 47 and

those not taking MFL a score of approximately 36. Which are significantly lower than empirical scores. This is likely to be because the simulation did not take account of combinations of subjects taken or ensuring that students aimed to fill all the A8 'buckets'. This is likely to have led to many empty slots and unrealistically low scores.

Restricting the simulated students to only those filling all A8 slots gave a mean score for those taking MFL subjects of around 50 and those not taking MFL at around 44, which is much closer to empirical scores, although still lower for those taking MFL subjects. Without making a more complex model including interactions of subject choices with variations in student ability this is likely to be the most valid model. Therefore, analysis focussed on the model produced after restricting to those filling all eight A8 slots. Table 3 gives summary data from the restricted simulation compared to empirical data.

Table 3. Summary data from simulation compared to empirical data.

Variable	**Model**	**Empirical data**
Mean Attainment 8 score	47.79 (SD=12.78)	47.96 (SD=18.75)
Mean number of subjects taken	9.64 (SD=0.89)	8.48 (SD=1.46)
% taking MFL subjects	63%	44%
Mean KS2 score	4.92 (SD=0.56)	4.78 (SD=0.71)

Student level

The analysis suggested that on average when students were entered for MFL rather than an 'average subject' this resulted in a lower A8 score of on average 0.05 points if MFL subjects are assumed to be half a grade harder than others, and 0.16 points less if MLF subjects are assumed to be 1 grade harder. These scores are lower than the estimates from the simple score adjustments above, likely because in most cases students had other subjects which could fill A8 slots instead. For P8 scores these differences translate to 0.005 and 0.016.

School level

At school level, almost all simulated schools fall into the middle category from the empirical data above of having between 20 and 80% of students taking an MFL subject. For these schools entering students for MFL resulted in an average 0.04 (SD = 0.23, range = -0.99-0.94) point reduction in A8 scores if MFL is half a grade harder and a 0.1

(SD = 0.22, range = -0.95-1.08) reduction in A8 scores if MFL subjects are simulated to be a grade harder than average (See Figure 2 for the distribution). This equates to a 0.004 and 0.01 change to P8 scores. The majority (682 schools) had a negative shift in scores, although 308 had a positive shift, potentially as there was some variance introduced in the calculation of MFL scores.

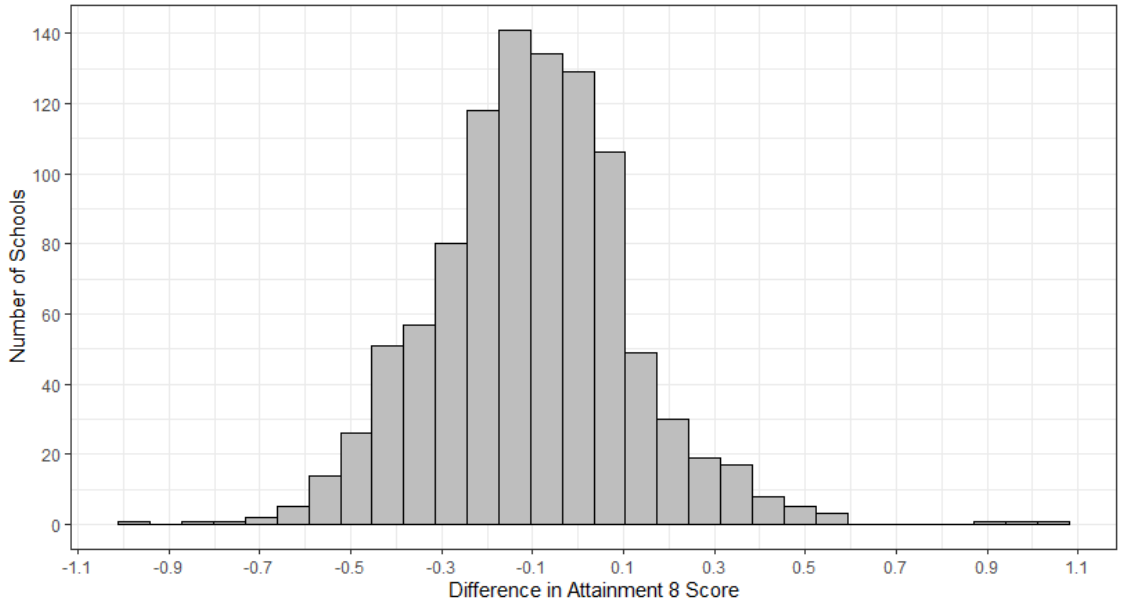


Figure 2. Histogram showing the distribution of the estimated impact on schools of entering students to MFL subjects on Attainment 8 scores based on simulation of MFL grades being on average 1 grade harder than other subjects.

A linear model was used to evaluate which school features predicted A8 and P8 scores in the simulation. All below models included the variables school size, mean school prior attainment, mean number of subjects taken, proportion of students taking MFL subjects and difference in prior attainment between students taking MFL and not taking MFL as independent variables. A8 was higher for larger schools (*ß*=0.001±0.005, *t*=2.44 p<0.001), schools where students took a larger number of subjects on average (*ß*=2.48±0.27, *t*=9.03 p<0.001) and schools which entered students with higher prior attainment compared to the average prior attainment within the school to MFL subjects (*ß*=3.23±0.54, *t*=6.01 p<0.001). In addition schools with higher prior attainment overall had higher A8 scores (*ß*=0.001±0.005, *t*=2.44 p<0.05). The A8 model accounted for a large proportion of variance with an adjusted R-squared of 0.89.

The model for P8 also suggested that larger schools (*ß*=0.0001±0.0005, *t*=2.14 p<0.05) and schools which entered students for more subjects on average (*ß*=0.22±0.025, *t*=8.63 p<0.001) had higher P8 scores. However, it also suggested that schools who enter a larger proportion of students to MFL subjects have a lower P8 score (*ß*=-0.16±0.05, *t*=-3.29 p<0.01). Overall this suggests that the average difference between a school entering no students for MFL subjects and entering all students would be approximately 0.16 P8 points. The model accounted for significantly less of the variance than the A8 model with an adjusted R-squared value of 0.086.

A second set of linear models looked at what school level features were related to the change in A8 and P8 scores when going from not entering to entering students to MFL subjects. For both P8 and A8, the only significant predictor was the schools mean prior attainment (A8: *ß*=0.4±0.06, *t*=6.74 p<0.001; P8: *ß*=0.04±0.006, *t*=6.74 p<0.001), suggesting that schools with students who have on average higher prior attainment were more likely to benefit, or at least see less negative effects, of entering students to MFL subjects. Although both models were weak with adjusted R-squares of 0.044 in both cases.

A correlation of the rank order of P8 scores including simulated MFL grades (with an average of 1 grade below other subjects) and excluding MFL showed that they were highly correlated (*r*=0.967*, N*=1000*, p*<0.0001*);* see Figure 3). However, by including MFL, schools moved down the relative rankings by up to 482 places. 146 schools moved (up or down) the rankings by over 100 places and 19 moving by more than 200 places. Suggesting that if MFL were the same difficulty as an 'average subject' it could have a substantial impact on relative school rankings.

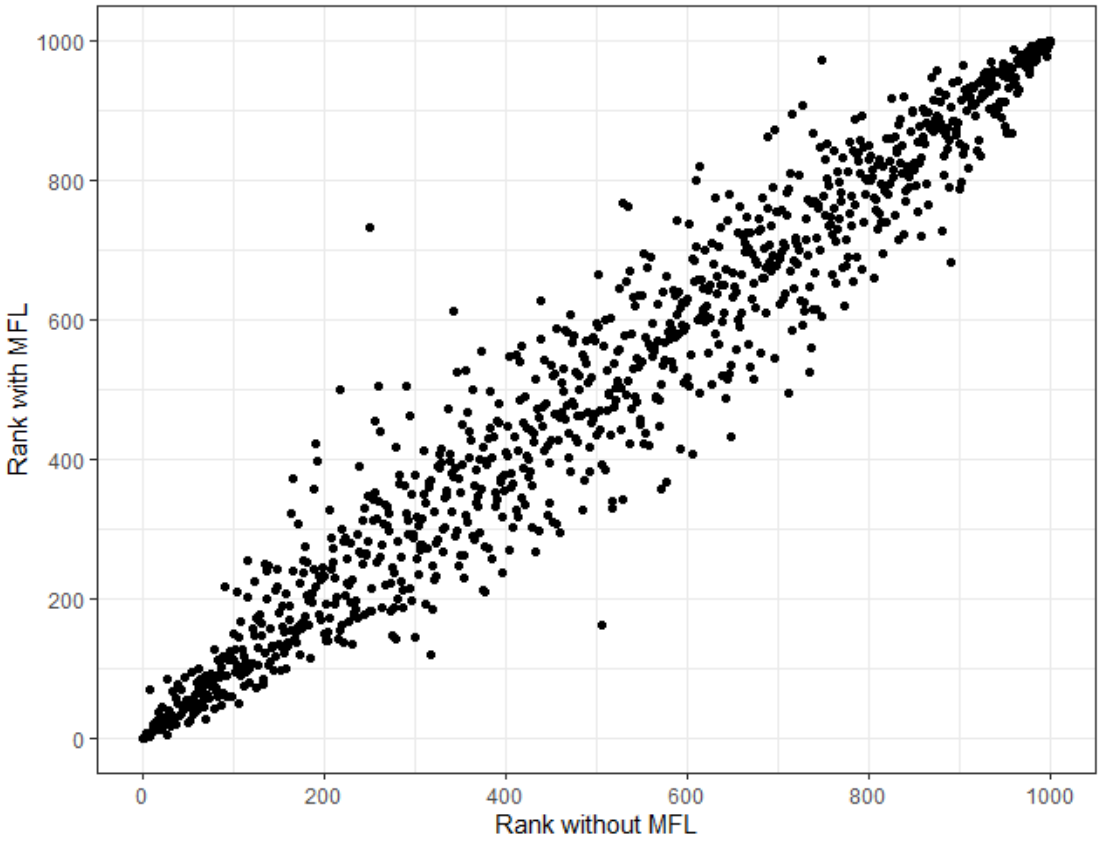


Figure 3. Correlation between school level Progress 8 scores including and excluding MFL subjects (based on a simulated additional difficulty of 1 grade).

3. Imputation of student grades

The most common subjects taken in the place of MFL subjects were Art and Design, Geography and Business Studies. Model accuracy for predicting MFL outcomes was reasonable, ranging from 46% correct prediction to 53%. Prediction accuracy for the alternate subjects taken instead of MFL was more variable, ranging from 37% (for Art and Design) to 71% (for Geography). See Appendix A for results for individual combinations taken by at least 500 students. Only French and Spanish MFL subjects were included in the final analysis as combinations including German were taken by fewer than 500 students.

Taking an average across models and across MFL subjects, MFL subjects were calculated to be more difficult than all other subjects on average. This estimate varied from an average of 0.26 (SD = 0.10) of a grade harder for business studies, to an average of 1.13 (SD = 0.08) of a grade harder than art and design. However, this did not uniformly translate to a lower A8 score.

When the MFL subject was replaced with non-Ebacc subjects (for example food technology, sports studies or business studies) students had better A8 scores with MFL. This is likely due to MFL being able to be used in both the Ebacc and 'other' buckets for calculation of A8; whereas these alternate subjects would only count in the 'other' bucket. Although analysis was restricted to students filling all of their A8 slots, replacing MFL with a non-Ebacc subject reduces the flexibility of how A8 is calculated, reducing the chance of gaining a better grade in each bucket. Replacing MFL with Geography or History, which can also be included in the Ebacc bucket, resulted in a higher A8 score for students by around 0.24 (SD = 0.18) points for History and 0.05 (SD = 0.21) points for Geography. Interestingly, replacing MFL with Art and Design resulted in a negligible effect on A8 scores (a reduction of A8 only by an average of 0.02 points), despite this also being a non-Ebacc subject. This may be because according to the difficulty estimates calculated earlier, Art and Design is one of the easiest subjects. This added score benefit may offset the lack of flexibility of not being able to use it in an Ebacc slot.

4. Modelling factors influencing accountability scores from empirical data

Student level scores

The student level P8 model suggested that even after controlling for a large suite of student and school level traits, taking MFL was still related to a higher, rather than lower, individual P8 score (ß=0.20 ± 0.004, t=51.83, p<0.001). Suggesting that students who took an MFL subject had on average a higher P8 score by 0.2 points. For the A8 model, students who took MFL were predicted to have a score 2.5 points higher than those who did not (ß = 2.49±0.04, t=59.95, p<0.001).

Predict school level scores

Table 4 shows the full model results for the school level A8 and P8 models respectively. The school level model predicting absolute scores of P8 showed that the proportion of students entered for MFL was again positively related to P8 score (*ß*=0.274±0.033, *t*=8.32 p<0.001). Schools with higher mean KS2 score overall had lower P8 scores (*ß*=-0.149±0.043, *t*=-3.42 p<0.01), which may be due to a ceiling effect of value added performance. The most important variables for P8 seemed to be if students had filled all their A8 slots (Ebacc slots: *ß*=0.68±0.067, *t*=10.15, p<0.001; Other slots: *ß*=1.757±0.092, *t*=19.19, p<0.001). A very similar pattern was seen for the model predicting A8 scores, although in this case mean KS2 score was positively related with A8 (*ß*=17.996±0.452, *t*=39.77 p<0.001).

Table 4. Model coefficients from model predicting school level Attainment 8 and Progress 8 scores.

	Attainment 8				Progress 8			
Variable	Coef	SE	T-value	P	Coef	SE	T-value	P
(Intercept)	-294.610	119.939	-2.456	<0.05	-23.898	11.514	-2.076	<0.05
School Capacity	0.000	0.000	-2.230	<0.05	0.000	0.000	-2.559	<0.05
Urban School	-0.002	0.193	-0.013	0.990	-0.004	0.019	-0.206	0.837
Max age over 16	0.440	0.145	3.036	<0.01	0.048	0.014	3.470	<0.01
Offers Latin	1.239	0.293	4.230	<0.001	0.103	0.028	3.675	<0.001
School Type - Modern [Comprehensive]	-0.886	0.328	-2.704	<0.01	-0.039	0.031	-1.244	0.214
School Type - Other Maintained [Comprehensive]	0.093	1.177	0.079	0.937	0.005	0.113	0.044	0.965
Number MFL subjects offered	0.097	0.106	0.915	0.360	0.006	0.010	0.631	0.528
Mean KS2 score	17.996	0.452	39.774	<0.001	-0.149	0.043	-3.424	<0.01
Mean number of Ebacc slots filled	6.511	0.698	9.333	<0.001	0.680	0.067	10.150	<0.001
Mean number of other slots filled	17.319	0.954	18.163	<0.001	1.757	0.092	19.190	<0.001
Mean student age	12.483	7.984	1.563	0.118	1.172	0.766	1.530	0.126
Mean IDACI score	-12.419	1.323	-9.386	<0.001	-1.085	0.127	-8.539	<0.001
Proportion FSM	-1.586	1.216	-1.304	0.192	-0.170	0.117	-1.453	0.146
Proportion EAL	8.073	0.391	20.646	<0.001	0.740	0.038	19.721	<0.001
Proportion taking at least one MFL	3.043	0.343	8.881	<0.001	0.274	0.033	8.322	<0.001
Proportion SEN	-0.441	0.919	-0.480	0.631	-0.094	0.088	-1.062	0.289
KS2 score MFL difference	0.216	0.403	0.536	0.592	-0.003	0.039	-0.068	0.946
Adjusted R-Squared	0.769				0.497			

Predict school level change in scores

The second set of school level models looked at predictors of change in P8 and A8 scores between 2017 and 2018 (full outputs in Table 5). The key predictors of a change in P8 score were an increase in students using all the slots in the Ebacc (*ß*=0.412±0.052, *t*=7.94 p<0.001) and 'other' buckets (*ß=1.483*±0.082, *t*=18.036 p<0.001). Schools which had a decrease in mean KS2 score (*ß=-0.112*±0.01, *t*=-11.81 p<0.001), and a reduction in mean IDACI score (*ß=-0.934*±0.38, *t*=-2.46 p<0.05) also saw an increase in P8 scores. A similar pattern was again seen for A8, but in this case an increase in mean KS2 score was again linked to an increase in A8 score (*ß*=1.933±0.098, *t*=19.773 p<0.001). When analysis was restricted to only those schools which had seen a larger than average (>0.2) change in P8 scores or had seen a larger than average change in proportion of students entered to MFL subjects, the same effects were seen.

Table 5. Model coefficients from model predicting the change in school level Attainment 8 and Progress 8 scores between 2017 and 2018.

	Attainment 8				Progress 8			
Variable	**Coef**	**SE**	**T-value**	**P**	**Coef**	**SE**	**T-value**	**P**
(Intercept)	-0.275	1.172	-0.235	0.815	-0.009	0.114	-0.075	0.940
Mean KS2	1.933	0.098	19.773	<0.001	-0.112	0.010	-11.814	<0.001
Number of students	-0.001	0.003	-0.237	0.813	0.000	0.000	-0.450	0.653
Mean Ebacc slots filled	3.986	0.534	7.469	<0.001	0.412	0.052	7.940	<0.001
Mean other slots filled	14.727	0.846	17.417	<0.001	1.483	0.082	18.036	<0.001
Mean student age	-8.969	6.178	-1.452	0.147	-0.700	0.601	-1.165	0.244
Mean IDACI score	-10.427	3.913	-2.664	<0.01	-0.934	0.380	-2.455	<0.05
Proportion FSM	-0.508	1.164	-0.437	0.662	-0.030	0.113	-0.264	0.792
Proportion EAL	2.032	1.144	1.775	0.076	0.154	0.111	1.380	0.168
Proportion entered for MFL	0.628	0.416	1.511	0.131	0.037	0.040	0.917	0.359
Proportion SEN	-1.082	1.044	-1.036	0.300	-0.181	0.102	-1.784	0.075
KS2 score MFL difference	-0.013	0.054	-0.235	0.814	-0.007	0.005	-1.288	0.198
Adjusted R-Squared	0.385				0.386			

5. Equating scores using a Partial Credit Model

The Partial Credit Model (PCM) included 23 subjects with at least 1000 student grades. EAP reliability for the model was 0.956, giving high confidence in the difficulty estimates calculated. Infit for all grades in all subjects included in analyses ranged from 0.77 to 1.53 (mean 1.06, SD 0.18) suggesting all subjects had a reasonable fit to the model. The upper limit of acceptable infit is slightly contentious, with some authors suggesting 1.3 as an upper bound of acceptability (Linacre, 2002), others slightly higher at 1.4 (Wong, McGrath and King, 2011). However, Linacre (2002) suggested that with values below 2.0 the model results would not be distorted. All subject level infit values (taken as a mean of grade level infit within each subject) were below 1.40, however Dance, Photography and Art & Design had mean infit values between 1.3 and 1.4, suggesting the results may need to be taken with caution for these subjects. However; the MFL subjects of interest all had acceptable subject level infit (French=1.13, German=1.09, Spanish=1.22; See Appendix B for further details).

A correlation matrix of residuals showed some minor pockets of correlations between subjects, particularly sciences (chemistry and biology: 0.37, biology and physics: 0.3, chemistry and physics: 0.44, maths and physics: 0.33) and maths and double science (0.32), art subjects (art and photography; 0.38) and most notably, languages (French and Spanish: 0.36, French and German: 0.39). After multiple imputation of missing

values (using the ‘Amelia’ package), a PCA of the standardized residuals produced one notable component with a mean eigenvalue of 5.21, explaining 22.7% of the variance. The ratio of the variance explained by the first component and the second component was 1.72. Although not conclusive, this may indicate that there is a pattern in the residuals, suggesting that there could be a second dimension to the data, violating the unidimensionality assumption of the PCM. Looking at the factor loadings for this component suggests some pattern to the loadings, with the highest positive loadings for Photography (0.28), Dance (0.25) and Art & Design (0.25) and negative loadings for Combined science (-0.33), Physics (-0.26) and Chemistry (-0.25). Potentially representing a dimension for aptitude at more creative subjects, although these loadings are quite low and further investigation would be required to draw any conclusions.

Grade difficulties and item fit statistics are shown in detail in appendix B. A summary of item difficulties and person abilities is shown through a wright map in Figure 4. The wright map indicates that subject difficulties are well targeted to student ability. Item difficulties are further shown graphically in Figure 5, shown in ‘logits’ as produced by the PCM model. Subjects have been arranged in rank order of difficulty based on their mean difficulty estimate. It is notable that utilising the partial credit methodology unlike other potential equating methods allows grade widths to be variable both within and between subjects. In particular, there seems to be a consistently large gap in difficulty between grade 3 and grade 4 across most subjects and a relatively smaller shift from grade 4 to grade 5. In general, there also seems to be closer clustering of grade difficulties at the top end.

In some cases the difficulty threshold estimates of grades within a subjects is disordered. This reflects a violation of the measurement construct implied by the Rasch model and inappropriate functioning of these items. It is likely that for the subjects where this occurs there is strong selection on students. The individual sciences (Biology, Chemistry, Physics) tend to be taken by stronger students rather than combined science. Students who receive a grade 1 in these subjects are therefore likely to be unusual and may not fit the normal pattern of results. Equally, as discussed previously, there is often high selection for MFL students, so the estimates for the lower grades for French, German and Spanish may also be less reliable.

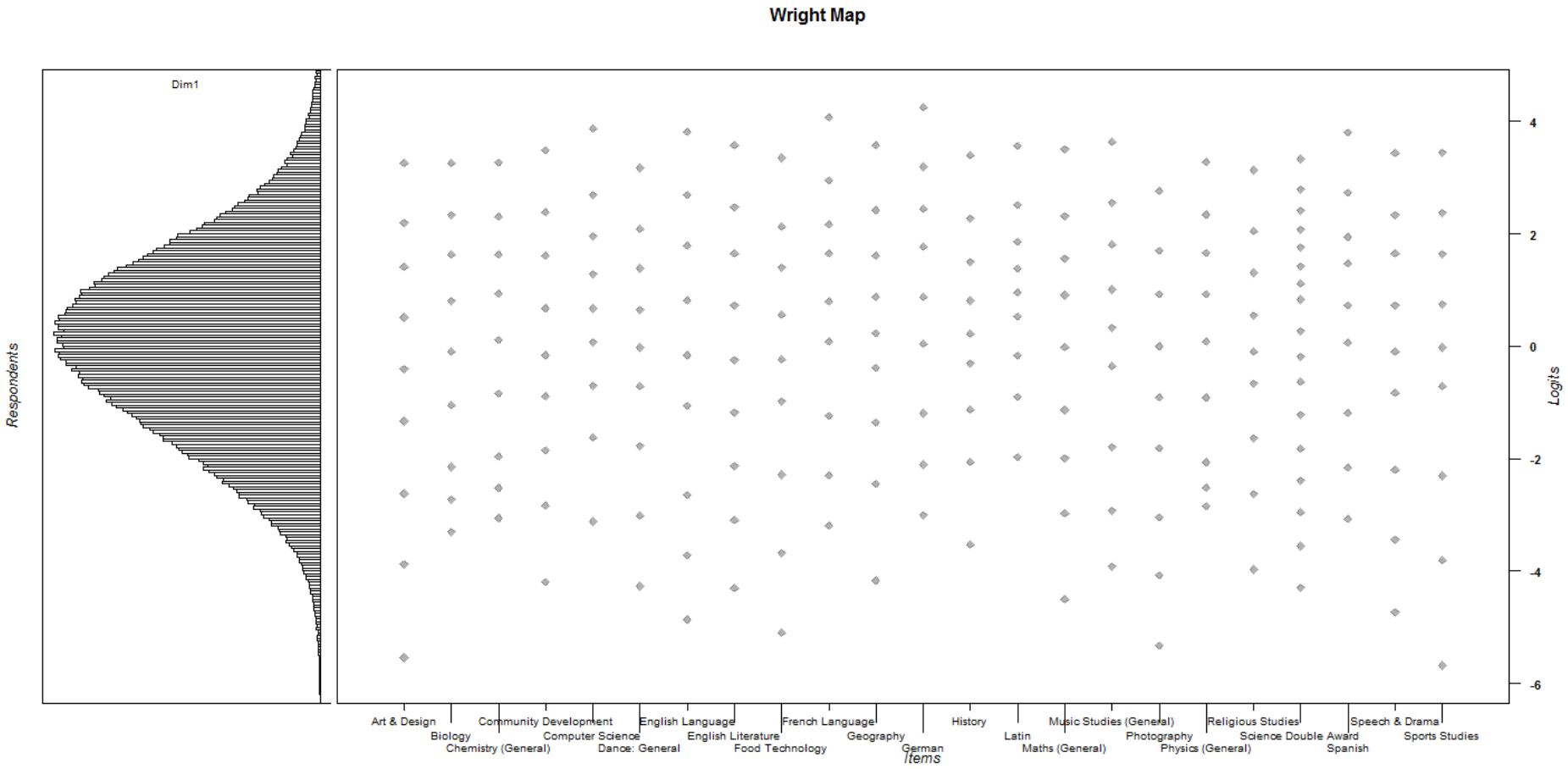


Figure 4. Wright map of subject difficulty estimates and student ability scores from the partial credit model.

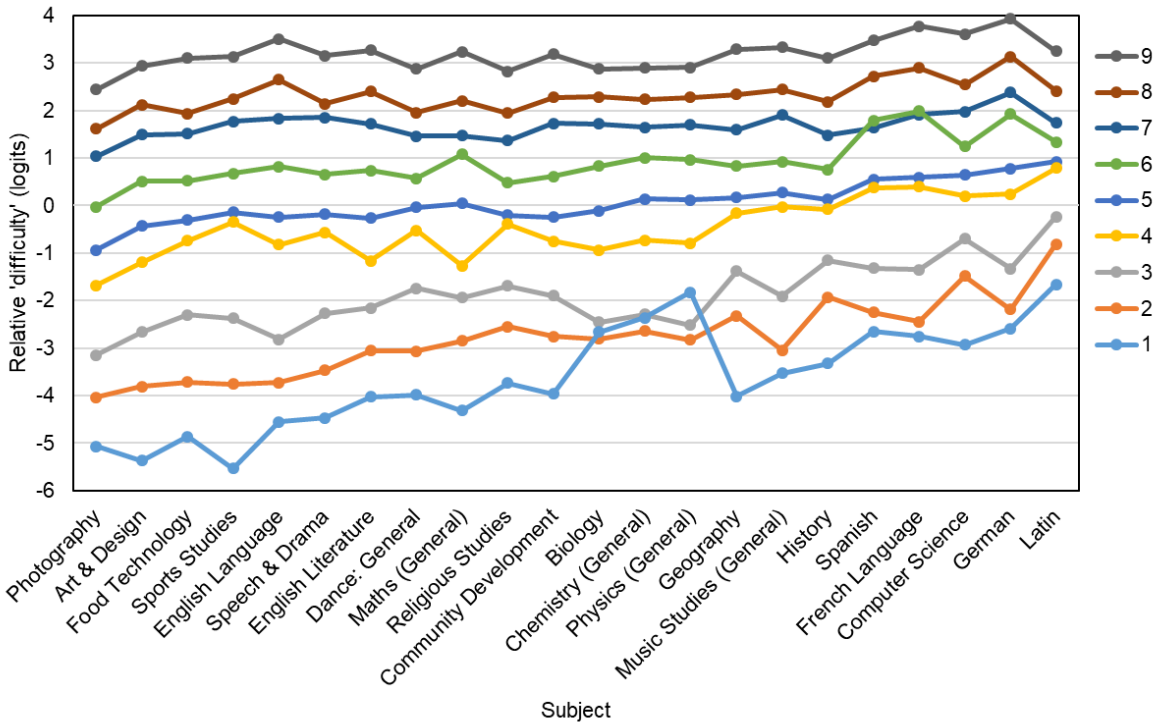


Figure 5. Rasch difficulty estimates from the PCM for each grade (9-1) for each subject included in the analyses. Subjects ordered by mean difficulty.

Creating Rach equated scores

A new score was created by standardising the difficulty estimates to mean of 5 and standard deviation of 2.5 and rounding to 1 decimal place. Scores below 0.5 on this scale were arbitrarily rounded up to 0.5 . This was done to create scores which have a similar overall distribution to the real scores (See Appendix B for subject and grade specific scores). However; if these were to be used practically more consideration might need to be put into the meaning of a 0 score on this scale and how great a penalty should be placed upon not taking a subject, as opposed to doing badly in a subject. As in some subjects (particularly Latin and Physics) a score of over 3 could be achieved by gaining a grade 1, providing substantial incentive to attempt a subject over not taking a subject. When converting to a point score disordered grades were fixed by changing the lower grade to be 0.5 points below the higher grade. Again a potential issue were this to be practically operationalised.

The new ‘Rasch equated’ subject scores were used to rerun the initial analysis evaluating difficulty estimates using the methods outlined previously. As can be seen from Table 6, although the mean grade scores still differ substantially between subjects, Kelly’s method and the difference from students mean score are now substantially

reduced. Although French and Spanish still remain two of the most 'difficult' subjects in statistical terms the difference between them and the average subject is now less than a quarter of a grade (rather than 0.7-1 prior to adjustment). Interestingly the biggest shift in the rank order is 'science: double award' which now comes out as one of the 'easiest' subjects. As discussed previously this may be due to the strong selection for students taking this instead of the individual sciences causing some skew in the difficulty estimates. Unfortunately as this equating was only done for a subset of subjects its direct effect on performance measures could not be assessed, but difficulty estimates suggest it would reduce the potential negative impact of taking MFL over other subjects.

Table 6. Subject difficulty estimates using the 'Rasch equated' subject scores.

Subject	N Students	Mean Grade	Diff from mean	Kelly's Method
French Language	103493	5.36	-0.12	-0.14
Physics	137986	5.90	-0.08	-0.13
Spanish	76379	5.33	-0.10	-0.11
Biology	139739	5.89	-0.05	-0.10
Chemistry	138828	5.92	-0.05	-0.09
Sports Studies	75614	4.63	-0.08	-0.09
English Language	467178	4.48	-0.08	-0.08
Music	29161	5.31	-0.04	-0.05
Speech & Drama	48847	4.73	-0.05	-0.05
German	38309	5.72	-0.03	-0.04
Latin	2819	7.05	0.00	-0.03
Art & Design	111839	4.47	-0.03	-0.01
Dance	7560	4.73	-0.01	0.00
Food Technology	42102	4.04	-0.01	0.00
English Literature	456781	4.58	0.01	0.01
Maths	477886	4.55	0.01	0.01
Photography	22915	3.94	-0.01	0.02
Religious Studies	190032	4.89	0.02	0.03
Geography	212446	4.78	0.03	0.03
Community Development	12725	4.43	0.02	0.04
History	223460	4.88	0.04	0.04
Computer Science	63431	5.17	0.07	0.07
Science Double Award	335235	4.20	0.20	0.23

Overall summary of Results

The statistical estimates of subject difficulty concur with previous analysis that students tend to gain lower grades in MFL subjects (French, German, Spanish) by around half to one grade compared with the average subject. Adjusting students' scores using a simple adjustment had a marginal effect on overall P8 and A8 scores, suggesting a negative effect on schools of entering students to MFL subjects. However, when included in a more complex simulation the estimated effect was reduced, likely to do with selectivity of subjects and whether students are filling all of their subject 'buckets'. This result was borne out through the imputation of students' grades, where results indicated that for students taking a non-Ebacc subject there may be some advantage to taking MFL subjects instead. Potentially as it provides greater flexibility in how A8 is calculated, as MFL subjects can be counted in the Ebacc or 'other' bucket. However, when replacing an Ebacc subject with MFL the effect of taking MFL is almost uniformly negative. The effect size suggested by all of the methods was small but relatively consistent.

Despite this estimated effect when modelling the empirical data, even after controlling for a large range of variables that may affect P8 and A8 scores, the effect of schools or students entering MFL subjects was either not significant or positive. This is potentially due to unmeasured effect such as student subject specific ability in languages, motivation or teaching quality, which are not included in the models or simulation as data on these variables was unavailable.

Finally, the Rasch adjusted scores reduced the difference in estimated difficulty between subjects. However, it potentially introduced additional complications and provided some indication of multidimensionality in student grades across subjects.

Discussion

The first part of the analysis presented here aimed to calculate some broad measures of subject 'difficulty' to observe if there were in reality measurable differences in grading standards between GCSE subjects. For these measures, we are generally assuming the common understanding of standards using the 'common candidature' definition, that if the same group of students took a set of subjects, then on average they would receive similar results in each subject (Coe, 2007). The results broadly reproduced what has been observed in other studies previously (e.g. Coe, 2008; He & Black, 2019) that some of the most difficult subjects in these terms are modern foreign languages (French, German, Spanish) and computer science. Each of these subjects was measured to be around a half to one grade more difficult than an average subject. This suggests that we would expect students to gain a grade that is a half to one grade lower on average in these subjects than other subjects they take at GCSE. There is potential criticism of utilising this fairly broad and un-nuanced definition of subject difficulty, as it takes no account of factors such as different teaching time, motivation or teaching quality. It also takes no account of subject specific differences in student ability and instead bases the definition of difficulty on a common scale of ability (similar to the 'linking construct' defined by Newton, 2010b). However, as the aim of this study was to address what impact inter-subject factors had on cumulative scores, which already make all of these assumptions, it was deemed appropriate.

This estimated difference in difficulty calculated in this first step was used to make a crude score adjustment to give an indication of what impact a difference in difficulty of this size is likely to have on Attainment 8 and Progress 8 scores. This simple score adjustment, aiming to broadly equate the statistical grading difficulty of MFL subjects with other subjects, suggested that for an average school, A8 scores would increase by between 0.16 and 0.33 if MFL subjects were 1 grade 'easier'. To put this into context, the standard deviation for A8 scores is 6.9, indicating that, based on this simple simulation, if an individual average school received scores in MFL subjects on average one grade higher for all students they would move up the school rankings by around 1.57%-3.25% of the total number of places. For schools entering the majority of students for MFL this effect increases slightly to 2.75%-5.81%. This simple analysis suggests that by entering the majority of students to MFL these schools are facing an additional disadvantage of around 1.18%-2.56% in terms of school league table position.

The second more complex simulation took account of more school factors such as school size and prior attainment distribution of the entry cohort. This more nuanced

model suggested a more modest effect of entering students for MFL subjects, of around 0.04-0.1 A8 points to an average school. Further linear modelling suggesting that after taking account of other factors included in the model, a school moving from entering no students to all students to an MFL subject would see a reduction in their A8 score of around 0.16 points, similar to the effect size indicated above. As expected, schools tended to see less negative effects of entering students to MFL subjects if they had a generally higher ability entry cohort and if they were entering students for more subjects on average. This is likely to be because in this case an alternative subject may be counted in the calculation of the performance measures if students achieved a better grade in the alternative subject, negating the impact of taking a more difficult subject. An almost identical pattern was seen for P8 measures due to the closed nature of the simulation.

The aim of these analyses was to estimate an approximation for the effect size of the potential disadvantage caused by entering students for MFL subjects. These results suggest that the effect of entering students to MFL subjects is likely to be small, although potentially notable in terms of school league tables. Even relatively small shifts in performance measures can change the relative placement of schools in an area, or push a school into a meaningfully different performance band (Leng & Koretz, 2015). DfE assigns schools to one of five performance bands based on the size of P8 scores and whether they are statistically different from 0, ranging from well below average (a score less than -5) to well above average (a score greater than or equal to 5; DfE, 2020b). Although as noted by Benton (2016) and repeated here accountability scores are still highly correlated before and after adjustment for subject difficulty, a potentially small change in score could push a school into a different category. This is particularly important for the parental choice aim of the measures as there is some evidence that parents focus on this banding rather than the exact scores (Menzies & Jerrim, 2020) and are usually only interested in comparing a small number of local schools.

The imputation analysis gave a more nuanced insight by looking at the impact of replacing an MFL subject with alternative specific subjects commonly taken by students instead. This analysis suggested that schools could see an advantage of entering students for alternative subjects to MFL, but that it depended on which subjects were chosen. Replacing MFL with an alternative subject that could also be used in the ‘Ebacc’ bucket of the performance measures led to an increase in A8 scores at the student level by 0.15-0.5 points, which is towards the top end of the estimates presented above. However; replacing an MFL subject with an alternative ‘non-Ebacc’ subject gave much less of an advantage and in most cases had a negative effect on A8 scores. Again this is

likely to be due to their flexibility in the calculation of performance measures and restrictions around which subjects can be counted in each 'bucket'.

Despite this theoretical advantage to schools of entering students to alternative subjects than MFL, when actual A8 and P8 scores were analysed, both at a student and school level, this advantage was not seen. After controlling for a wide number of student and school level factors, modelling indicated a minor advantage (rather than disadvantage) to schools of entering students to MFL, despite the results of the analyses above. Instead, what the modelling suggested was important for schools to gain higher A8 and P8 scores was that students took enough subjects to fill all of the 'slots' regardless of which subjects they were. There are likely to be a host of unmeasured school level factors not taken into account by these models, which could affect student attainment in different subjects. To compensate for this the final model looked at within school changes in the accountability scores and related it to changes in various other within school factors. By modelling it this way, we are inherently taking account of any other factors that differ between schools, which could be affecting results. However, again this model showed no significant impact on accountability scores of either an increase or decrease in the proportion of students entered for MFL subjects within a school. Instead, it again showed that schools that increased the proportion of students filling all of the 'Ebacc' slots in the measures had an increase in overall scores. There is the possibility that increases in performance measures is linked to other unmeasured changes within the school between those two years, which have not been accounted for here, proving potential for further investigation.

Evidence for incentives driven by the performance measures

A possible reason for the difference seen between the simulated impact of taking these more difficult subjects and the realised scores is likely due to subject selection. Schools in England are free to set their own entrance policy for MFL subjects, and as noted previously, schools tend to enter higher ability students for MFL subjects (Parrish & Lanvers, 2019), who in general perform better at all subjects taken. In addition to this there is likely to also be a subject specific selection based on student aptitude (Parrish & Lanvers, 2019). Students who have an aptitude for languages, potentially above and slightly disconnected from their 'general academic ability', are more likely to enter these subjects (Parrish & Lanvers, 2019). This may mean that as the simulations suggest there is a disadvantage for the average student in taking an MFL subject, based on their general aptitude in other subjects. However, due to this subject specific selection effect this disadvantage may be masked by schools only entering students for these subjects

with a high subject specific aptitude, not accounted for by the statistical measures calculated here. If this is the case, it may be that students are still relatively disadvantaged by taking MFL based on the difficulty of the subject, as potentially they should be doing even better in the subject for which they have a high aptitude, but due to differences in grading standards their scores are lower than they should be.

If this is true, this would suggest that the 'perverse incentives' which have been perceived may indeed be real and schools are simply responding to the pressures applied to them by being more selective and entering less students to MFL subjects. This unfortunately would mean that schools with a less selective policy for MFL subjects, who could be considered to be 'gaming the system' less, are likely to be those with the most visible disadvantage when it comes to performance measures. If this is undesirable then policy makers may need to rethink how these scores are calculated and the intended educational objectives (Neal, 2008). An additional concern is that the results of the analysis suggest that the relative disadvantage is higher for schools with a generally lower ability candidature, who also tend to take less subjects. This type of disadvantage is beyond the school's control and may distort the desired measures.

There is an additional risk that these incentives are reducing the choice available to students with lower prior ability. If schools are only entering students with higher prior ability scores, they are essentially restricting choice to part of the curriculum to certain students. An effect identified by Harris et al. (2020) for other Ebacc subjects. A potential source of inequality, particularly as it is more likely to impact students from areas of low socio-economic status (Harris et al., 2020). Reducing this access to certain subjects and knowledge has the potential to reduce social mobility (Griggs & Walker, 2008). There is also the possibility that this restriction has the additional impact of affecting student's perceived value of foreign language learning (Parrish & Lanvers, 2019). It is concerning if school decisions for entering students for certain subjects is being motivated not by what is best for the student but by increasing performance scores. Evidence presented here suggests that even if the effect is relatively minor it still demonstrates a potentially powerful lever in determining school behaviour and which students gain access to certain subjects. Potentially representing an unintended consequence, directly opposed to the stated intention of the measures to encourage a 'broad and balanced curriculum' (DfE, 2020a).

Although the analysis suggests incentives may exist for schools to be more selective in which students they enter for MFL subjects, this effect of entering students for MFL subjects over other subjects on schools accountability scores is relatively modest. It

would be surprising if this was the main cause of a consistent decline in entries to language subjects in the UK, although it may be one in a series of contributing factors (Parrish & Lanvers, 2019). As a potential counter-argument to the suggestion that current performance measures are the cause of this decline, many other English speaking nations have also seen a decline into language subjects, despite having a wide variety of different performance measures (Berman, 2011).

One incentive which is clearly being driven by the current calculation of performance measures is for schools to fill all of the 8 subject slots which count towards the P8 and A8 measures. All of the analyses suggested a strong disadvantage to schools of leaving an 'empty slot'. Due to the way the measure is calculated an empty slot counts as a 0 in the total score (DfE, 2016b) and can reduce a candidates potential score by around $1/10^{th}$. Rather than a 'perverse' incentive, this is an intentional design decision as part of the measure to encourage students to take a "broad and balanced curriculum", with a particular focus on core academic subjects (DfE, 2020a). It would seem that it has been successful in driving this behaviour as schools have generally seen an increase in the number of students taking the required number of each qualification since these performance measures were introduced (Gill, 2018). Although it is also perhaps notable that there has also been almost no increase in the uptake of MFL subjects over this period (Gill, 2018).

Subject specific ability and composite scores

As discussed above, the reason for this potential high measured 'difficulty' in language subjects in particular, may be due to aptitude in languages being somewhat divorced from ability in other subjects. Research suggests that the link between general academic ability and language ability may be less clear than for more academic subjects. Laufer (1992) found that there was only a correlation between general academic ability and second language reading performance once students were above a certain threshold, which required a reasonable vocabulary in the target language. This suggests that some sort of basic language learning ceiling needs to be reached before ability in the first language can begin to transfer (Perkins et al., 1989). Other evidence suggests that language subject results tend to be less well correlated with other subjects (Benton & Sutch, 2013), suggesting the underlying construct across different subjects may not be entirely unidimensional (Korobko et al., 2008). In particular, there is a weaker correlation between MFL subjects with maths and English than other 'core' subjects included in the performance measures included in the Ebacc bucket (Benton & Sutch, 2013). This was partially supported in this study as the PCM of the Rasch model residuals gave some

indication of a second dimension and there was a small but consistent correlation between the residuals of language subjects.

This relationship between languages and other subjects is further notable when it's considered that the P8 measure is based on the value added relationship between key stage 2 results and GCSEs. The key stage 2 score consists of an average of results from only maths and English assessments. He and Black (2019) found a greater deviance in value added from KS2 tests, particularly for design, drama and languages when compared to more academic subjects. Between the sitting of key stage 2 assessments at age 11 and taking GCSEs there is the opportunity for significant subject selection based on individual differences in educational progression. This subject specific filtering may lead to apparent different patterns of difficulty across subjects, particularly subjects where the core construct may differ significantly from the 'core academic subjects'. This effect is most notable when they are treated as interchangeable with other subjects that are potentially more closely related to 'general academic aptitude'. If composite accountability scores are intentionally measuring a 'linking construct' along the lines of general academic ability, then the difference in how strongly this construct is represented in different subjects should be explicitly accounted for or at least evaluated. Therefore, any method that makes some of the assumptions of measuring an underlying latent trait between subjects such as 'general academic ability' is best served by applying more complex statistical measures, such as a Rasch model (Lamprianou, 2009).

Rasch equated scores

The last part of this study attempted to account for subject specific difficulty in the creation of a composite score by applying a partial credit model to results data. The aim was to calculate relative subject scores as a trial of one proposed method to counteract this subject inequality created by the current methodology. To do this, a partial credit model was ran to create new 'Rasch adjusted scores' which could be used in the calculation of performance measures. Important to note is that the proposal is not to change the initial grading of qualifications, but to ensure that they are converted onto the same scale for the purposes of performance measures. These scores are explicitly aligned along a scale created by the model, which has previously been interpreted as 'general academic ability' (Coe, 2008).

The data reasonably fitted the model, an important criteria for Rasch based models, allowing further interpretation. The equated scores when subsequently used in the

analysis of subject difficulty did go some way to balancing out different subjects. Although MFL subjects still appeared more difficult, the effect was greatly reduced. However, this method comes with its own complications. In some cases some grades were disordered, which draws questions around the validity of the model as a whole. There were also some issues with how the difficulty estimates could be effectively and fairly converted into scores, requiring consideration about the incentives driven by the method. This analysis also only focussed on a subset of subjects taken by a relatively large number of students and excluded other non-GCSE qualifications which are usually included in performance measures. Previous studies have had issues with some subjects fitting a similar Rasch model (e.g. Coe, 2008). What to do with these subjects were a method such as this applied is a difficult question. Further evaluation would also need to be carried out around differential item functioning, essentially is the scale computed consistent across subgroups, a common problem with statistical measures of subject difficulty (Goldstein & Cresswell, 1996; Newton, 1997).

Arguably, this additional level of complexity to solve a relatively minor problem is unnecessary and creates more problems than it solves. As pointed out by Lamprianou (2009), stakeholder buy-in is highly important in national assessment schemes. Although there is currently some complaint around the current system, at least it is relatively straightforward and understandable by the non-specialist. Given the examples from other education systems (discussed by Lamprianou, 2009), applying the sort of equating created through the PCM is likely to result in other concerns about unfairness in how it is applied to different subjects. It also becomes much more obscure whether these complaints are legitimate. Despite these complexities, Coe (2010) is heavily in favour of this type of methodology, suggesting that it could be freeing in terms of the types of assessments available as long as we can calculate a 'score conversion' on to the linking construct of interest for specific purposes.

The analysis of residuals raised questions over whether GCSE scores could be considered unidimensional. Different GCSE subjects are not designed to measure a particular overarching construct, and so any interpretation as such necessarily needs to be inferred retrospectively (Bramley, 2016). It has been questioned previously whether this idea of 'general academic ability underlying all subjects can be successfully applied to groups such as arts or vocational subjects (Kelly, 1975). As discussed by Coe (2010) if we are claiming to be measuring a 'linking construct' such as general academic ability then we should present a full validity argument for how students' performance on these various assessments represent this construct. This could potentially include arguments on the representativeness of the content to the construct, how internally consistent each

assessment is and whether the measured variable correlates with other external measures as we would expect (Coe, 2010). Although this potentially represents a significant resource investment to the current system, this assumption of unidimensionality is already inherently made through simple composite scores (Tognolini & Andrich, 1996). At least by applying a model such as the Rasch analysis, the assumption of whether individual qualifications adequately represent the linking construct can be tested empirically. If individual subjects do not fit the model there is potentially an argument against them being included in any composite measures, simple or complex (Coe, 2008).

As has been stated previously, what the linking construct represents, as calculated by the Rasch model, is not entirely clear (Bramley, 2011). The trait is likely to be influenced by those subjects, which are well correlated with one another (He & Black, 2019). Although it is likely that there is a common element of skill between these assessments, as discussed previously caution needs to be taken when interpreting scores for subjects with a greater aspect of subject specific performance. Currently there is limited research of the validity of the 'general academic ability' interpretation, particularly when it relates back to these subjects (He & Black, 2019). The Rasch model also takes no account of other factors that may cause differences in grading standards beyond, for example, the amount of teaching time, student motivation, the impact on different subgroups, how effective teaching and teaching materials are, etc. All these factors could make a subject appear more or less 'difficult' (Newton, 2012). However Coe (2008) argues that this does not invalidate scores from such analyses, as if for example a subject is inherently more motivating causing students to work harder then it is valid to consider grades in that subject to be 'worth less' as the advantage gained in that subject is unlikely to transfer to other contexts.

Are Progress 8 and Attainment 8 valid measures of school performance?

To evaluate whether overall P8 and A8 are valid measures we ideally need to consider the adequacy and appropriateness of the inferences and actions based on these scores (Messick, 1989). Unfortunately, despite progress measures already being a second order use of qualification results, these measures in the UK still do not have a single clear purpose (Kelly & Downey, 2009). Current performance measures are currently used for a range of purposes, muddying the precise aims of the scores and subsequently affecting what factors should be prioritised around their calculation.

If the primary purpose of these measures is for public facing 'school league tables' and informing parental choice, then measures primarily need to be understandable without requiring expertise. In which case the effects of subject differences in grading standards on performance measures measured here are relatively minor and any method used to overcome them will introduce significant complexity, potentially undermining their meaning and value. However, if these scores were intended to be used primarily internally by government for judging school standards, there would be an argument for allowing them to be more complex, as there would be less concern about them being interpretable by the general public. In which case there would be a stronger argument for further exploring methods to correct for the effects measured here to ensure that the measures are not driving unwanted behaviours. Particularly those which result in limits to student choice and go directly against the stated aims of the performance measures of promoting a 'broad and balanced curriculum' (DfE, 2020a). Ultimately, to effectively evaluate the appropriateness of the measures a decision needs to be made to balance statistical robustness, usability and accessibility (Kelly & Downey, 2009).

Throughout this study relatively little difference was seen in the patterns showed by either the absolute performance measure (A8) or the relative progress measure (P8). However, there was one key place where they diverged. Schools which had an intake with a higher prior attainment, as measured by key stage 2 scores, had consistently higher A8 scores, but not P8 scores. Although a minor point, this highlights the value of prioritising progress based measures rather than absolute measures, as a school's score becomes less dependent on the cohort they happen to have each year giving a fairer representation of school quality (Gerwitz et al., 2019).

A final thing to note is that despite the focus here on how performance measures handle grading standards between different subjects, it has been widely discussed that there are other, potentially more concerning issues with these performance measures. It has been highlighted on a number of occasions that outcome based and progress based performance measures can be biased against those that come from lower socio-economic backgrounds (Leckie & Goldstein, 2019; Reynolds et al., 2014; Teddlie & Reynolds, 2000; Neal, 2008). Due to this it has been suggested that schools in England can end up being consistently defined as 'underperforming' simply because of being situated in a relatively deprived area (Burgess & Thomson, 2013). To some extent, this was reflected in the analysis presented here. The models attempting to account for factors affecting school level P8 and A8 consistently showed that schools with higher IDACI scores (a measure of social deprivation), had lower P8 and A8 scores even once other school characteristics, including which subjects were entered and prior attainment

scores, were controlled for. Schools that had a lower mean IDACI score in 2018 compared to 2017, also on average saw an increase in their P8 score. Therefore, schools whose intake consists of these types of students can appear substantially worse, despite potentially exceeding the normal value added for the intake cohort. As the aim of all performance measures should be to focus on factors which are under the school's control (Ooghe & Schokkaert, 2009), these types of variables, of which the school has no control, potentially unfairly penalise schools, regardless of subject entry decisions discussed here.

Unfortunately, controlling for all of these potential factors statistically when calculating performance measures may be impossible (Jones, 1997) and can cause additional complications (Perry, 2016) particularly where the purposes of measures are unclear (Kelly & Downey, 2009). One pragmatic suggestion that could work to overcome this issue is not related to how the measures are calculated, but how they are reported. The NAHT accountability commission (2018) recommended that schools should only be compared to other 'similar' schools rather than nationally. The exact definition of 'similar' is potentially a point for discussion, but the laudable aim would be that schools would only be compared to other schools with similar uncontrollable factors. This would give a more accurate impression of how those schools are relatively performing, given their circumstances.

However even with these changes Leckie and Goldstein (2009, 2011) caution that school league tables are unreliable and misleading guides and are poor predictors of future school performance. They advocate greater scepticism when these scores are interpreted and conclude that although performance measures can be useful for accountability, they are a poor measure for school choice (Leckie & Goldstein, 2017). Interestingly this is echoed in the original technical design document for the P8 measure, which states that due to differences in student intake between schools the scores cannot be guaranteed to be directly comparable between schools and therefore the measures are "*less appropriate for parental choice than for school self-evaluation and inspection*" (Burgess & Thomson, 2013, p11).

Limitations

This study was inherently limited to the data that was available. The study only utilised two years of data from the National Pupil Database and so was not able to observe more long term trends in either school entry policy or changing accountability scores. A key weakness of the data used was due to some of the unmeasured variables, which would

have been valuable in pulling apart the cause of both student and school level subject entry strategies. Data on school policies for which students were entered for MFL subjects and how much choice students had in their own subject choices, would have allowed further investigation of how these policies ultimately impacted on performance measure scores. What would have been highly valuable would have been some independent measure of language ability to identify if controlling for this factor had any impact on student MFL results. Although to my knowledge no such national measure exists and would therefore require a more bespoke study, involving collecting primary data. In addition, a study collecting primary data could potentially be able to make more valid conclusions about cause and effect, whereas in this study all interpretations about causation are necessarily tentative.

One thing that came to light during the study was the particular complexities around language subjects and their relationship to achievement in more ‘core academic’ subjects. In light of this it would be valuable to repeat the analysis carried out here using an alternative ‘difficult’ subject, such as computer science. There is the possibility that because ability in computer science is likely to be more highly linked to other similar academic or science subjects it may be easier to identify any subject specific bias in scores and therefore the impact of this difference in grading standards on performance measures.

Conclusion

This study indicates through simulation that there is potentially a bias in performance measures against schools that enter a large proportion of students for MFL subjects. However, analysis indicates this bias is likely to be small and is masked in the empirical data, potentially by schools generally choosing to enter more able students, particularly those with a language aptitude, to these subjects. This does not mean the bias does not exist, but that it is being relatively effectively counter balanced by school entry practices. In general this suggests that schools are likely to be befitted by only entering students who are more able, particularly in languages, to MFL subjects, and by ensuring that if an MFL subject isn't taken it is replaced by another subject which can be used to fill the 'Ebacc bucket' in the calculation of the performance measures.

The suggestion of using a partial credit model to explicitly account for differences in subject difficulty (Coe, 2010), along the lines of a linking construct considered as 'general academic ability' was trialled. Although this methodology was successful in reducing inter-subject differences in difficulty scores across the grade range, it comes with a number of complications, which may make it difficult to apply in practice. Chief amongst these is likely to be its communicability and acceptance by stakeholders, in particular the general public. As asserted by Lamprianou (2009), assessments exist as a 'social contract' within society where general agreement needs to be held about their value and how they are used. Therefore solving the above issue caused by differential subject grading standards by using a more complex methodology may inadvertently undermine the aims of the measures. As stated by O'Neill (2013) "*Every time one performance indicator is shown to be inaccurate, or misleading, or likely to produce perverse results, some people claim that they can devise a better one that has no perverse effects. Experience suggests that they may well be as wrong as those who invented the last lot of indicators*". Instead of developing more complex measures, what is perhaps required is a more nuanced interpretation of the available data, alongside other qualitative information when making decisions about school quality and effectiveness.

References

Allen, R. (2014). *Scaling for ATAR calculations – Tasmania: 2005–2013.* Tasmanian Qualifications Authority. https://www.tasc.tas.gov.au/wp-content/uploads/2017/07/scaling.pdf

Alton, A., & Pearson, S. (1996). *Statistical approaches to inter-subject comparability.* UCLES research paper.

Backhouse, J. K. (1978). *Comparability of Grading Standards in Science Subjects at GCE A Level: A Study of Two Examining Boards in 1973 and 1974.* Evans Bros.

Baird, J. (2007). Alternative conceptions of comparability. In Newton, P., Baird, J., Goldstein, H., Patrick, H. and Tymms, P. (Eds.) *Techniques for monitoring the comparability of examination standards* (pp. 124–156). Qualifications and Curriculum Authority.

Benton, T. (2016). On the impact of aligning the difficulty of GCSE subjects on aggregated measures of pupil and school performance. *Research Matters: A Cambridge Assessment publication*, *22*, 27-30.

Benton, T. and Sutch, T. (2013). *Exploring the value of GCSE prediction matrices based upon attainment at Key Stage 2.* Cambridge Assessment Research Report.

Berman, R. A. (2011). The real language crisis. *Academe, 97*(5), 30-34.

Bramley, T. (2005) Accessibility, easiness and standards. *Educational Research*, *47*(2), 251–261.

Bramley, T. (2011). Subject difficulty - the analogy with question difficulty. *Research Matters: A Cambridge Assessment Publication, Special Issue 2: comparability*, 27-33.

Bramley, T. (2016). The effect of subject choice on the apparent relative difficulty of different subjects. *Research Matters: A Cambridge Assessment publication, 22,* 23-26.

Briggs, D. C., & Weeks, J. P. (2009). The sensitivity of value-added modeling to the creation of a vertical score scale. *Education Finance and Policy, 4*(4), 384–414.

Burgess, S. & Thomson, D. (2013). *Key Stage 4 Accountability: Progress Measure and Intervention Trigger.* Centre for Understanding Behaviour Change, Short Policy Report No. 13/11.

Burgess, S., Propper, C., & Wilson, D. (2007). The impact of school choice in England: Implications from the economic evidence. *Policy Studies, 28*(2), 129-143.

Chiang, H. (2009). How accountability pressure on failing schools affects student achievement. *Journal of Public Economics, 93*(9-10), 1045-1057.

Coe, R. (2007). Common Examinee Methods. In Newton, P., Baird, J., Goldstein, H., Patrick, H. and Tymms, P. (Eds.) *Techniques for monitoring the comparability of examination standards* (pp. 331–367). Qualifications and Curriculum Authority.

Coe, R. (2008). Comparability of GCSE examinations in different subjects: an application of the Rasch model. *Oxford Review of Education*, *34*(5), 609–636.

Coe, R. (2010). Understanding comparability of examination standards. *Research Papers in Education*, *25*(3), 271–284.

Coe, R., Searle, J., Barmby, P., Jones, K., & Higgins, S. (2008). *Relative difficulty of examinations in different subjects.* CEM centre.

Cresswell, M.J. (1996). Defining, setting and maintaining standards in curriculum embedded examinations: judgemental and statistical approaches. In Goldstein, H. and Lewis, T. (Eds.) *Assessment: Problems, Developments and Statistical Issues* (pp. 57–84). John Wiley & Sons.

Cresswell, M.J. (2000). The Role of Public Examinations in Defining and Monitoring Standards. In Goldstein, H. and Heath, A. (Eds.) *Educational Standards* (pp. 69–120). Oxford University Press.

Cyprus Minister of Education. (2006). Press release 14 July. Available online at: http://www.cyprus.gov.cy/moi/pio/pio.nsf/All/F2277BE6883F716EC22571AB003B058D?Opendocument&print

Daley, D. J. (1989). *Determining relative academic achievement for fair admission to higher education.* Report to ANU, Canberra CAE and ACT Schools Authority.

Dearing, R. (1996). *Review of Qualifications for 16-19 Year Olds.* School Curriculum and Assessment Authority.

Dearing, R. and King, L. (2007). *Languages Review.* Department for Education and Skills Publications.

DfE (2014). Factsheet: Progress 8 measure. *Department for Education.*

DfE (2016a) Progress 8 measure in 2016, 2017 and 2018: Guide for maintained secondary schools, academies and free schools. *Department for Education.*

DfE (2016b). Progress 8: How Progress 8 and Attainment 8 measures are calculated. *Department for Education.*

DfE (2020a). Secondary accountability measures: Guide for maintained secondary schools, academies and free schools. *Department for Education.*

DfE (2020b). Progress scores for key stage 4: school and college performance tables. *Department for Education.*

Education Datalab. (2015). *Floors, Tables & Coasters: Shifting the Education Furniture in England's Secondary Schools.* Education Datalab. https://educationdatalab.org.uk/wp-content/uploads/2016/02/2015-Educationfurniture-04.pdf

Education Datalab. (2019). *Solutions to problems with Progress 8, part one: qualification scoring.* Education Datalab. https://ffteducationdatalab.org.uk/2019/10/solutions-to-problems-with-progress-8-part-one-qualification-scoring/

Elliott, G. (2013). *A guide to comparability terminology and methods.* Cambridge, Cambridge Assessment.

Fitz-Gibbon, C. T., & Vincent, L. (1997). Difficulties regarding subject difficulties: developing reasonable explanations for observable data. *Oxford Review of Education*, *23*(3), 291-298.

Forrest, G. M. and Smith, G. A. 1972. *Standards in Subjects at the Ordinary Level of the GCE, June 1971*. Joint Matriculation Board, OP34.

Gewirtz, S., Maguire, M., Neumann, E., & Towers, E. (2019). What's wrong with 'deliverology'? Performance measurement, accountability and quality improvement in English secondary education. *Journal of Education Policy*, 1-26.

Gill, T. (2018). *The impact of the introduction of Progress 8 on the uptake of qualifications in English schools – an update for 2016/17*. Cambridge Assessment Research Report. Cambridge Assessment.

Goldstein, H., & Cresswell, M. (1996). The comparability of different subjects in public examinations: a theoretical and practical critique. *Oxford Review of Education, 22*(4), 435-442.

Goodhart, C.A.E. (1984) *Monetary theory and practice*. Macmillan.

Gorard, S., Hordosy, R. & Siddiqui, N. (2012) How unstable are 'school effects' assessed by a value-added technique?, *International Education Studies, 6*(1), 1–9.

Griggs, J. & Walker, R. (2008). *The Costs of Child Poverty for Individuals and Society.* Joseph Rowntree Foundation.

Guarino, C. M., Stacy, B. W., & Wooldridge, J. M. (2019). Comparing and assessing the consequences of two different approaches to measuring school effectiveness. *Educational Assessment, Evaluation and Accountability,* 31(4), 437-463.

Harris, R., & Burn, K. (2011). Curriculum theory, curriculum policy and the problem of ill-disciplined thinking. *Journal of Education Policy, 26*(2), 245-261.

Harris, R., Courtney, L., Ul-Abadin, Z., & Burn, K. (2020). Student access to the curriculum in an age of performativity and accountability: an examination of policy enactment. *Research Papers in Education*, *35*(2), 228-248.

He, Q. & Black, B. (2019). Statistical evidence pertaining to the claim of grading severity in GCSE French, German and Spanish and the impact of statistical alignment of standards on outcomes*. Office of Qualifications and Examinations Regulation*.

He, Q., Stockford, I., & Meadows, M. (2018). Inter-subject comparability of examination standards in GCSE and GCE in England. *Oxford Review of Education, 44*(4), 494-513.

Hoffmann, B., & Barzun, J. (2003). *The tyranny of testing.* Courier Corporation.

Johnson, V. E. (1997). An alternative to traditional GPA for evaluating student performance. *Statistical Science*, 251-269.

Jones, B.E. (1997). Comparing Examination Standards: is a purely statistical approach adequate? *Assessment in Education: Principles, Policy & Practice, 4*(2), 249–263.

Kelly, A. (1975). The relative standards of subject examinations. *Research Intelligence, 1*(2), 34-38.

Kelly, A. (1976). A study of the comparability of external examinations in different subjects. *Research in Education, 16*(1), 37-63.

Kelly, A., & Downey, C. (2011). Professional attitudes to the use of pupil performance data in English secondary schools. *School Effectiveness and School Improvement, 22*(4), 415-437.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices.* Springer Science & Business Media.

Korobko, O. B., Glas, C. A., Bosker, R. J., & Luyten, J. W. (2008). Comparing the difficulty of examination subjects with item response theory. *Journal of Educational Measurement, 45*(2), 139-157.

Lamprianou, I. (2007). Comparability methods and public distrust: An international perspective. In Newton, P., Baird, J., Goldstein, H., Patrick, H. & Tymms, P. (Eds.) *Techniques for monitoring the comparability of examination standards*, (pp. 368–371). Qualifications and Curriculum Authority.

Lamprianou, I. (2009). Comparability of examination standards between subjects: an international perspective. *Oxford Review of Education, 35*(2), 205-226.

Lanvers, U. (2017a). Elitism in language learning in the UK. In Rivers, D. & Zotzmann, K. (Eds.) *Isms in Language Education: Oppression, Intersectionality and Emancipation*, (pp. 50–73). DeGruyter.

Lanvers, U. (2017b). Language learning motivation, Global English and study modes: a comparative study. *The Language Learning Journal, 45*(2), 220-244.

Lanvers, U., & Coleman, J. A. (2017). The UK language learning crisis in the public media: A critical analysis. *The Language Learning Journal, 45*(1), 3-25.

Laufer, B. (1992). Reading in a foreign language: How does L2 lexical knowledge interact with the reader's general academic ability'. *Journal of Research in Reading, 15*(2), 95-103.

Leckie, G., & Goldstein, H. (2009). The limitations of using school league tables to inform school choice. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *172*(4), 835-851.

Leckie, G., & Goldstein, H. (2011). A note on 'The limitations of school league tables to inform school choice'. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *174*(3), 833-836.

Leckie, G., & Goldstein, H. (2017). The evolution of school league tables in England 1992–2016: 'Contextual value-added', 'expected progress' and 'Progress 8'. *British Educational Research Journal*, 43(2), 193-212.

Leckie, G., & Goldstein, H. (2019). The importance of adjusting for pupil background in school value-added models: A study of Progress 8 and school accountability in England. *British Educational Research Journal, 45*(3), 518-537.

Leng, H. L., & Koretz, D. (2015). Sensitivity of school-performance ratings to scaling decisions. *Applied Measurement in Education, 28*(4), 330-349.

Linacre, J. (2002). What do Infit and Outfit, Mean-square and Standardized mean? *Rasch Measurement Transactions 16*, 878.

Menzies, L. & Jerrim, J. (2020). *Improving headline school performance measures.* The Centre for Education & Youth.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational researcher*, *18*(2), 5-11.

Myers, H. (2006). The 'severe grading' of MFL grades at GCSE and A level. *Association for Language Learning.*

NAHT (2018). *Improving school accountability.* NAHT Accountability Commission.

Neal, D. 2008. Designing incentive systems for schools . In Springer M. (Ed.), *Performance incentives: their growing impact on American K-12 education: forthcoming*. Brookings.

Newbould, C.A. and Hall, J. (1981). *Comparative standards in Advanced level mathematics.* University of Cambridge Local Examinations Syndicate

Newbould, C. A. (1982). Subject preferences, sex differences and comparability of standards. *British Educational Research Journal, 8*(2), 141-146.

Newton, P. E. (1997). Measuring comparability of standards between subjects: why our statistical techniques do not make the grade. *British Educational Research Journal, 23* (4), 433–451.

Newton, P. E. (2005). Examination standards and the limits of linking. *Assessment in Education: Principles, Policy & Practice, 12*(2), 105-123.

Newton, P. E. (2010a). Contrasting conceptions of comparability. *Research Papers in Education, 25*(3), 285–292.

Newton, P. E. (2010b). Thinking about Linking. *Measurement: Interdisciplinary Research and Perspectives, 8*(1), 38–56.

Newton, P. E. (2012). Making sense of decades of debate on inter-subject comparability in England. *Assessment in Education: Principles, Policy & Practice, 19*(2), 251–273.

NFER (2018). *What impact does accountability have on curriculum, standards and engagement in education? A literature review.* National Foundation for Educational Research.

Noden, P. (2001) *Excellence in Cities: Teacher characteristics, expectations and attitudes*. NFER.

Nuttall, D.L., Backhouse, J.K. and Willmott, A.S. (1974). Comparability of Standards Between Subjects (Examinations bulletins / Schools Council). *Methuen Educational and Evans Bros.*

O'Neill, O. (2013). Intelligent accountability in education. *Oxford Review of Education*, *39*(1), 4-16.

Ofqual. (2015). Inter-Subject Comparability of Exam Standards in GCSE and A Level: ISC Working Paper 3. *Office of Qualifications and Examinations Regulation*.

Ofqual. (2019). Grading standards in GCSE French, German and Spanish. *Office of Qualifications and Examinations Regulation*.

Ooghe, E., & Schokkaert, E. (2009). School Accountability: (How) Can We Reward Schools and Avoid Cream-Skimming? *SSRN Electronic Journal.*

Parrish, A., & Lanvers, U. (2019). Student motivation, school policy choices and modern language study in England. *The Language Learning Journal*, *47*(3), 281-298.

Perkins, K., Brutten, S. R., & Pohlmann, J. T. (1989). First and second language reading comprehension. *RELC Journal*, *20*(2), 1-9.

Perry, T. (2016). English value-added measures: Examining the limitations of school performance measurement. *British Educational Research Journal*, *42*(6), 1056-1080.

R Core Team (2019). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing.

Reynolds, D., Sammons, P., De Fraine, B., Van Damme, J., Townsend, T., Teddlie, C., & Stringfield, S. (2014). Educational effectiveness research (EER): A state-of-the-art review. *School Effectiveness and School Improvement*, *25*(2), 197-230.

Royal Society (2008) Science and mathematics education, 14 –19. A 'state of the nation' report on the participation and attainment of 14–19 year olds in science and mathematics in the UK, 1996–2007. *London, the Royal Society.*

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581-592.

Schools Council Forum on Comparability. 1979. *Standards in public examinations: Problems and possibilities. Comparability in Examinations Occasional Paper 1*. Schools Council.

Spearman, C. (1904). "General Intelligence," Objectively Determined and Measured. *The American Journal of Psychology, 15*(2), 201-292.

Teddlie, C. & Reynolds, D. (2000) *The international handbook of school effectiveness research*. Psychology Press.

Tinsley, T., & Board, K. (2017a). Language Trends 2016/17. Language Teaching in Primary and Secondary Schools in England. *British Council.*

Tinsley, T. and K. Board. (2017b). Languages for the Future*. British Council.*

Tinsley, T. (2019). Language Trends 2019*. British Council.*

Titcombe, R. (2008). How Academies Threaten the Comprehensive Curriculum. *In FORUM: for promoting 3-19 comprehensive education (Vol. 50, No. 1, pp. 49-59). Symposium Journals.*

Tognolini, J., & Andrich, D. (1996). Analysis of profiles of students applying for entrance to universities. *Applied Measurement in Education*, *9*(4), 323-353.

Watkins, M. W., Lei, P. W., & Canivez, G. L. (2007). Psychometric intelligence and achievement: A cross-lagged panel analysis. *Intelligence*, *35*(1), 59-68.

West, A. (2010). High stakes testing, accountability, incentives and consequences in English schools. *Policy & politics*, *38*(1), 23-39.

Wilkinson, D., Bryson, A., & Stokes, L. (2018). Assessing the variance in pupil attainment: how important is the school attended?. *National Institute Economic Review*, *243*(1), R4-R16.

Wiliam, D. (1996). Standards in examinations: a matter of trust? *The Curriculum Journal, 7*(3), 293–306.

Wilson, D., & Piebalga, A. (2008). Performance measures, ranking and parental choice: an analysis of the English school league tables. *International Public Management Journal*, *11*(3), 344-366.

Wong, H., McGrath, C. & King, N. (2011). Rasch validation of the early childhood oral health impact scale. *Community Dent Oral Epidemiology, 39*, 449–457.

Young, J.W. (1990). Adjusting the cumulative GPA using item response theory. *Journal of Educational Measurement,* 27, 175-186.

Young, J.W. (1991). Gender bias in predicting college academic performance: A new approach using item response theory. *Journal of Educational Measurement,* 28, 31~M.

Appendix A – Results from individual imputed score combinations

Table A. Results from imputation of student grades and the impact on Attainment 8 and Progress 8 scores, by subject combination.

MFL Subject	Other Subject	N Students with MFL combo	N Students with other combo	MFL Prediciton Accuracy	Other Prediction Accuracy	MFL Difficulty Estimate	A8 Decrease with MFL	P8 Decrease with MFL
French Language	Art & Design	1012	1603	46%	37%	1.16	-0.07	-0.01
French Language	Geography	1012	975	46%	58%	0.44	-0.13	-0.01
French Language	Business Studies	1012	770	46%	53%	0.28	0.53	0.05
French Language	Sports Studies	1012	693	46%	54%	0.60	0.27	0.03
French Language	Food Technology	1012	610	46%	57%	0.94	0.12	0.01
French Language	Geography	862	507	47%	71%	0.65	-0.37	-0.04
Spanish	Art & Design	799	1603	46%	37%	1.06	0.18	0.02
Spanish	Geography	799	975	46%	58%	0.22	0.26	0.03
Spanish	Business Studies	799	770	46%	53%	0.12	0.84	0.08
Spanish	Sports Studies	799	693	46%	54%	0.52	0.61	0.06
Spanish	Food Technology	799	610	46%	57%	0.93	0.36	0.04
French Language	Religious Studies	787	1603	52%	45%	1.02	0.12	0.01
French Language	Geography	787	625	52%	67%	0.37	-0.10	-0.01
French Language	Art & Design	781	1341	51%	37%	1.24	-0.23	-0.02
French Language	History	781	975	51%	50%	0.45	-0.37	-0.04
French Language	Sports Studies	781	696	51%	54%	0.82	0.14	0.01
French Language	Business Studies	781	658	51%	52%	0.32	0.37	0.04
French Language	Food Technology	781	586	51%	58%	1.11	-0.08	-0.01
Spanish	Geography	631	507	48%	71%	0.37	0.07	0.01
Spanish	Religious Studies	601	1603	52%	45%	0.95	0.34	0.03
Spanish	Geography	601	625	52%	67%	0.31	0.01	0.00
Spanish	Art & Design	599	1341	53%	37%	1.07	0.20	0.02
Spanish	History	599	975	53%	50%	0.41	-0.11	-0.01
Spanish	Sports Studies	599	696	53%	54%	0.82	0.39	0.04
Spanish	Business Studies	599	658	53%	52%	0.32	0.63	0.06
Spanish	Food Technology	599	586	53%	58%	0.96	0.34	0.03

Appendix B – Rasch equated subject scores

Table B. Results of partial credit model to equate subject grades, including Rasch equated point scores.

Subject	Grade	No. Students	Infit	Difficulty Estimate	SE	Original Points	Rasch Equated Points
Art & Design	1	2415	1.34	-5.37	0.06	1	0.5
Art & Design	2	7664	1.45	-3.81	0.02	2	0.9
Art & Design	3	17134	1.43	-2.67	0.01	3	2.2
Art & Design	4	19428	1.42	-1.20	0.01	4	3.8
Art & Design	5	21755	1.48	-0.43	0.01	5	4.6
Art & Design	6	18991	1.46	0.51	0.01	6	5.6
Art & Design	7	11974	1.35	1.49	0.01	7	6.7
Art & Design	8	7633	1.29	2.12	0.01	8	7.4
Art & Design	9	4408	1.14	2.93	0.02	9	8.3
Art & Design	U	354				0	0
Biology	1	855	0.79	-2.67	0.04	1	1.5
Biology	2	2164	0.79	-2.81	0.03	2	2
Biology	3	8397	0.78	-2.45	0.02	3	2.4
Biology	4	15575	0.77	-0.94	0.01	4	4.1
Biology	5	25080	0.80	-0.11	0.01	5	5
Biology	6	28611	0.79	0.83	0.01	6	6
Biology	7	23318	0.78	1.71	0.01	7	6.9
Biology	8	18750	1.06	2.28	0.01	8	7.6
Biology	9	16132	1.52	2.87	0.01	9	8.2
Biology	U	803				0	0
Chemistry (General)	1	898	0.80	-2.37	0.04	1	1.7
Chemistry (General)	2	2459	0.81	-2.64	0.03	2	2.2
Chemistry (General)	3	9904	0.82	-2.29	0.02	3	2.6
Chemistry (General)	4	17395	0.82	-0.73	0.01	4	4.3
Chemistry (General)	5	24391	0.84	0.14	0.01	5	5.2
Chemistry (General)	6	25022	0.81	1.00	0.01	6	6.2
Chemistry (General)	7	22504	0.80	1.65	0.01	7	6.9
Chemistry (General)	8	19247	1.05	2.23	0.01	8	7.5
Chemistry (General)	9	16137	1.51	2.90	0.01	9	8.2
Chemistry (General)	U	867				0	0
Community Development	1	677	0.99	-3.96	0.08	1	0.7
Community Development	2	1287	1.03	-2.75	0.04	2	2.1
Community Development	3	2057	1.01	-1.90	0.03	3	3
Community Development	4	2024	0.97	-0.75	0.03	4	4.3
Community Development	5	2299	1.00	-0.24	0.03	5	4.8
Community Development	6	2065	0.98	0.61	0.03	6	5.7
Community Development	7	1124	0.99	1.73	0.03	7	7
Community Development	8	655	1.02	2.28	0.04	8	7.6
Community Development	9	313	0.98	3.18	0.07	9	8.6
Community Development	U	220				0	0
Computer Science	1	5471	0.95	-2.93	0.03	1	1.9
Computer Science	2	7668	0.97	-1.49	0.02	2	3.5
Computer Science	3	9590	0.97	-0.70	0.01	3	4.3
Computer Science	4	8841	0.98	0.19	0.01	4	5.3
Computer Science	5	8878	1.00	0.64	0.01	5	5.8
Computer Science	6	8157	0.98	1.24	0.01	6	6.4
Computer Science	7	6045	0.94	1.97	0.01	7	7.2
Computer Science	8	4451	0.90	2.54	0.02	8	7.9

Computer Science	9	2261	0.89	3.60	0.03	9	9
Computer Science	U	2067				0	0
Dance: General	1	238	1.26	-3.99	0.14	1	0.7
Dance: General	2	727	1.50	-3.07	0.07	2	1.7
Dance: General	3	1222	1.53	-1.75	0.04	3	3.2
Dance: General	4	1176	1.45	-0.53	0.04	4	4.5
Dance: General	5	1234	1.40	-0.03	0.03	5	5
Dance: General	6	1202	1.28	0.57	0.03	6	5.7
Dance: General	7	808	1.16	1.46	0.04	7	6.7
Dance: General	8	574	1.19	1.96	0.05	8	7.2
Dance: General	9	312	1.25	2.87	0.07	9	8.2
Dance: General	U	66				0	0
English Language	1	10144	1.13	-4.55	0.02	1	0.5
English Language	2	30649	1.15	-3.73	0.01	2	1
English Language	3	88037	1.09	-2.82	0.01	3	2
English Language	4	81252	1.02	-0.82	0.00	4	4.2
English Language	5	93961	0.98	-0.25	0.00	5	4.8
English Language	6	76929	0.90	0.82	0.00	6	6
English Language	7	45732	0.85	1.83	0.01	7	7.1
English Language	8	24127	0.92	2.65	0.01	8	8
English Language	9	11891	0.99	3.50	0.01	9	8.9
English Language	U	3392				0	0
English Literature	1	17476	1.13	-4.03	0.01	1	0.7
English Literature	2	35260	1.16	-3.05	0.01	2	1.7
English Literature	3	60401	1.09	-2.15	0.01	3	2.7
English Literature	4	80520	1.02	-1.16	0.00	4	3.8
English Literature	5	91030	0.99	-0.27	0.00	5	4.8
English Literature	6	75549	0.96	0.74	0.00	6	5.9
English Literature	7	46123	0.95	1.72	0.01	7	7
English Literature	8	27962	0.94	2.40	0.01	8	7.7
English Literature	9	15243	0.96	3.26	0.01	9	8.6
English Literature	U	7033				0	0
Food Technology	1	1545	0.96	-4.86	0.06	1	0.5
Food Technology	2	5119	0.96	-3.71	0.03	2	1
Food Technology	3	8942	0.98	-2.30	0.02	3	2.6
Food Technology	4	6970	0.98	-0.74	0.01	4	4.3
Food Technology	5	6980	1.00	-0.31	0.01	5	4.7
Food Technology	6	5735	1.00	0.52	0.02	6	5.6
Food Technology	7	3213	0.97	1.51	0.02	7	6.7
Food Technology	8	2235	1.05	1.93	0.02	8	7.2
Food Technology	9	1013	1.09	3.10	0.04	9	8.5
Food Technology	U	326				0	0
French Language	1	2578	1.07	-2.76	0.03	1	2.1
French Language	2	8755	1.11	-2.44	0.02	2	2.4
French Language	3	20808	1.12	-1.35	0.01	3	3.6
French Language	4	17037	1.07	0.39	0.01	4	5.5
French Language	5	20888	1.11	0.59	0.01	5	5.7
French Language	6	11071	1.06	1.98	0.01	6	6.7
French Language	7	10667	1.03	1.91	0.01	7	7.2
French Language	8	6819	1.18	2.90	0.01	8	8.2
French Language	9	3676	1.45	3.77	0.02	9	9.2
French Language	U	1192				0	0
Geography	1	14642	0.84	-4.02	0.02	1	0.7
Geography	2	24969	0.83	-2.32	0.01	2	2.5
Geography	3	35267	0.81	-1.38	0.01	3	3.6
Geography	4	28774	0.81	-0.17	0.01	4	4.9
Geography	5	30339	0.82	0.17	0.01	5	5.3
Geography	6	28884	0.80	0.83	0.01	6	6
Geography	7	22379	0.79	1.60	0.01	7	6.8
Geography	8	15226	0.85	2.33	0.01	8	7.6
Geography	9	8366	0.92	3.29	0.01	9	8.7
Geography	U	3595				0	0
German	1	825	1.03	-2.59	0.06	1	2.3
German	2	2479	1.08	-2.18	0.04	2	2.7
German	3	6492	1.06	-1.33	0.02	3	3.6
German	4	6937	1.02	0.24	0.02	4	5.3
German	5	8043	1.05	0.78	0.01	5	5.9
German	6	5330	1.02	1.92	0.02	6	7.2

German	7	3925	1.04	2.38	0.02	7	7.7
German	8	2475	1.16	3.13	0.02	8	8.5
German	9	1420	1.36	3.92	0.03	9	9.4
German	U	381				0	0
History	1	17317	0.93	-3.33	0.02	1	1.4
History	2	25437	0.92	-1.93	0.01	2	3
History	3	33394	0.91	-1.15	0.01	3	3.8
History	4	27070	0.91	-0.08	0.01	4	5
History	5	30348	0.93	0.13	0.01	5	5.2
History	6	30317	0.92	0.75	0.01	6	5.9
History	7	24745	0.93	1.48	0.01	7	6.7
History	8	17947	0.96	2.18	0.01	8	7.5
History	9	10700	1.00	3.10	0.01	9	8.5
History	U	6174				0	0
Latin	1	34	0.94	-1.66	0.33	1	3.3
Latin	2	76	0.95	-0.81	0.18	2	4.2
Latin	3	156	1.03	-0.24	0.12	3	4.8
Latin	4	173	1.08	0.79	0.09	4	5.9
Latin	5	245	1.12	0.93	0.07	5	6.1
Latin	6	341	1.14	1.34	0.06	6	6.5
Latin	7	493	1.12	1.74	0.06	7	7
Latin	8	619	1.16	2.41	0.05	8	7.7
Latin	9	669	1.03	3.25	0.06	9	8.6
Latin	U	13				0	0
Maths (General)	1	23307	0.97	-4.31	0.01	1	0.5
Maths (General)	2	40883	0.98	-2.85	0.01	2	2
Maths (General)	3	61322	0.98	-1.94	0.01	3	3
Maths (General)	4	100039	0.95	-1.26	0.00	4	3.7
Maths (General)	5	90810	0.94	0.05	0.00	5	5.1
Maths (General)	6	57717	0.94	1.07	0.00	6	6.3
Maths (General)	7	45890	0.95	1.47	0.01	7	6.7
Maths (General)	8	32680	0.96	2.20	0.01	8	7.5
Maths (General)	9	17199	1.01	3.23	0.01	9	8.6
Maths (General)	U	6848				0	0
Music Studies (General)	1	628	1.09	-3.52	0.07	1	1.2
Music Studies (General)	2	2030	1.25	-3.05	0.04	2	1.7
Music Studies (General)	3	4843	1.29	-1.90	0.03	3	3
Music Studies (General)	4	3888	1.34	-0.03	0.02	4	5
Music Studies (General)	5	4501	1.41	0.27	0.02	5	5.4
Music Studies (General)	6	4923	1.37	0.92	0.02	6	6.1
Music Studies (General)	7	3564	1.28	1.90	0.02	7	7.2
Music Studies (General)	8	2735	1.34	2.44	0.02	8	7.7
Music Studies (General)	9	1781	1.26	3.32	0.03	9	8.7
Music Studies (General)	U	266				0	0
Photography	1	414	1.35	-5.06	0.11	1	0.5
Photography	2	1339	1.46	-4.03	0.05	2	0.7
Photography	3	3493	1.44	-3.15	0.03	3	1.6
Photography	4	4146	1.42	-1.68	0.02	4	3.2
Photography	5	4671	1.45	-0.93	0.02	5	4.1
Photography	6	4164	1.42	-0.03	0.02	6	5
Photography	7	2404	1.35	1.03	0.02	7	6.2
Photography	8	1437	1.30	1.61	0.03	8	6.8
Photography	9	732	1.18	2.44	0.04	9	7.7
Photography	U	89				0	0
Physics (General)	1	579	0.84	-1.82	0.04	1	1.5
Physics (General)	2	1907	0.86	-2.83	0.03	2	2
Physics (General)	3	9484	0.86	-2.52	0.02	3	2.3
Physics (General)	4	17343	0.85	-0.79	0.01	4	4.2
Physics (General)	5	24554	0.86	0.12	0.01	5	5.2
Physics (General)	6	26133	0.84	0.96	0.01	6	6.1
Physics (General)	7	22470	0.83	1.70	0.01	7	6.9
Physics (General)	8	18718	1.16	2.27	0.01	8	7.6
Physics (General)	9	15872	1.53	2.90	0.01	9	8.2
Physics (General)	U	921				0	0
Religious Studies	1	8668	1.00	-3.73	0.02	1	1
Religious Studies	2	16513	1.04	-2.55	0.01	2	2.3
Religious Studies	3	26971	1.05	-1.69	0.01	3	3.2
Religious Studies	4	22747	1.05	-0.39	0.01	4	4.6

Religious Studies	5	28248	1.07	-0.21	0.01	5	4.9
Religious Studies	6	30526	1.05	0.48	0.01	6	5.6
Religious Studies	7	23519	1.01	1.36	0.01	7	6.6
Religious Studies	8	18087	1.03	1.95	0.01	8	7.2
Religious Studies	9	11700	1.03	2.82	0.01	9	8.2
Religious Studies	U	3036				0	0
Science Double Award	11	6746	0.96	-3.81	0.02	1	0.9
Science Double Award	21	12244	1.01	-3.58	0.01	1.5	1.2
Science Double Award	22	19058	1.04	-2.94	0.01	2	1.9
Science Double Award	32	27296	1.04	-2.38	0.01	2.5	2.5
Science Double Award	33	37641	1.03	-1.87	0.01	3	3
Science Double Award	43	43916	1.03	-1.23	0.01	3.5	3.7
Science Double Award	44	37105	1.03	-0.47	0.01	4	4.6
Science Double Award	54	36605	1.00	-0.22	0.01	4.5	4.8
Science Double Award	55	39944	1.00	0.06	0.01	5	5.1
Science Double Award	65	18479	1.01	1.27	0.01	5.5	5.7
Science Double Award	66	15386	0.98	1.01	0.01	6	6.2
Science Double Award	76	12583	0.95	1.32	0.01	6.5	6.5
Science Double Award	77	8404	0.92	1.81	0.01	7	7.1
Science Double Award	87	5874	0.95	2.05	0.01	7.5	7.3
Science Double Award	88	3841	1.00	2.41	0.02	8	7.7
Science Double Award	98	2482	1.09	2.75	0.02	8.5	8.1
Science Double Award	99	2519	1.24	2.71	0.02	9	8.6
Science Double Award	U	5112				0	0
Spanish	1	2257	1.19	-2.66	0.04	1	2.2
Spanish	2	6505	1.23	-2.25	0.02	2	2.6
Spanish	3	14958	1.24	-1.32	0.01	3	3.6
Spanish	4	11962	1.19	0.37	0.01	4	5.5
Spanish	5	14089	1.17	0.55	0.01	5	5.7
Spanish	6	7996	1.12	1.79	0.01	6	6.4
Spanish	7	8729	1.09	1.64	0.01	7	6.9
Spanish	8	5524	1.26	2.72	0.02	8	8.1
Spanish	9	3253	1.53	3.47	0.02	9	8.9
Spanish	U	1102				0	0
Speech & Drama	1	1097	1.17	-4.46	0.07	1	0.5
Speech & Drama	2	3704	1.25	-3.47	0.03	2	1.3
Speech & Drama	3	8409	1.20	-2.27	0.02	3	2.6
Speech & Drama	4	7457	1.12	-0.56	0.01	4	4.5
Speech & Drama	5	9031	1.13	-0.19	0.01	5	4.9
Speech & Drama	6	8817	1.07	0.65	0.01	6	5.8
Speech & Drama	7	4714	1.02	1.85	0.02	7	7.1
Speech & Drama	8	3483	1.06	2.13	0.02	8	7.4
Speech & Drama	9	1861	1.09	3.15	0.03	9	8.5
Speech & Drama	U	265				0	0
Sports Studies	1	1343	1.06	-5.54	0.09	1	0.5
Sports Studies	2	5760	1.08	-3.75	0.03	2	1
Sports Studies	3	15405	1.05	-2.38	0.02	3	2.5
Sports Studies	4	11651	1.03	-0.34	0.01	4	4.7
Sports Studies	5	13854	1.03	-0.15	0.01	5	4.9
Sports Studies	6	13062	0.99	0.68	0.01	6	5.8
Sports Studies	7	7326	0.94	1.77	0.01	7	7
Sports Studies	8	4643	1.01	2.24	0.02	8	7.5
Sports Studies	9	2416	1.07	3.13	0.02	9	8.5
Sports Studies	U	140				0	0