Is Progress 8 a valid and reliable measure of school effectiveness?

LEDGER, MARK,RICHARD

Is Progress 8 a valid and reliable measure of school effectiveness?

Mark Ledger

Thesis submitted for the degree of
Doctor of Philosophy

School of Education

Durham University

2021

Mark Ledger: Is Progress 8 a valid and reliable measure of school effectiveness?

Abstract

Policy-makers, school leaders, parents and citizens want to know whether schools are doing their job well, and whether particular schools or types of schools are doing that job particularly well. Insofar as the job is defined in terms of pupil attainment in public examinations, value added models are currently preferred. However, both the validity and reliability of value added models have been questioned and the debates about their fairness remain unresolved. One of the major problems for value added models is that while raw-scores for each school are reasonably stable over time the value added scores based on them are more volatile. This instability does not prove that there is a problem with the measures, but it is how construct irrelevant variance would manifest. This thesis addressed these concerns by scrutinising the validity of Progress 8, the Department for Education's headline indicator of school performance in England. More specifically, it investigates whether the differences between schools' annual performance ratings and the change in schools' ratings over time can be explained by the kinds of factors that educational effectiveness is usually attributed to and perhaps more importantly, whether these factors are under the control of schools. The results show two things. First, that the Progress 8 scores are biased by external variables such as the differences in schools' intake and examination entries. This is profoundly unfair and is likely to mean that the wrong schools are identified as differentially effective. And second, that even school leaders with expert knowledge of their institutions, access to students' performance data and the previous year's attainment averages cannot make reliable predictions about schools' value-added results. This outcome invalidates the notion that parents can use Progress 8 outputs as a means of making informed decisions about the effect of their child attending one school over another.

Table of Contents:

Statement of Copyright

Acknowledgements

I would like to begin by expressing my gratitude to my PhD supervisors, Professor Stephen Gorard, Associate Professor Nadia Siddiqui and Professor Rob Coe for providing guidance and feedback throughout this project. I would also like to thank my parents and partner Abbie for their support.

1. Introduction

1.1. Chapter Introduction

This chapter provides a brief overview of the thesis, including the key research objectives, the issues surrounding them and details of how the current research project has expanded upon past research. A chapter summary presented at the end of the section then provides a more detailed breakdown of individual chapters.

1.2 Thesis Introduction

1.2.1. Thesis topic and contribution

Policy-makers, school leaders, parents and citizens want to know whether schools are doing their job well, and whether particular schools or types of schools are doing that job particularly well. Of course, a wide range of indicators can be used to judge the quality of schools, but most official accounts are based on student outcome data such as examination results. Insofar as the job is defined in terms of pupil attainment in public examinations, value added models, such as Progress 8 in England are currently preferred. These are models of pupil progress during a specified phase of schooling, and are deemed fairer than raw-score outcomes which are deemed largely a reflection of the nature and prior attainment of each school intake (Goldstein and Woodhouse, 2000). However, both the validity and reliability of value added models have been questioned and the debates about their fairness remain unresolved (Morris *et al.*, 2018).

One of the major problems for value added models is that while raw-scores for each school are reasonably stable over time the value added scores based on them are more volatile (Dumay *et al.*, 2014). This lack of stability or distinctiveness raises the question of whether value added scores are genuinely a measure of school performance. This makes it difficult for a parent to select a secondary school for their 10-year-old child as schools' current ratings may have little or no resemblance to future ratings. There are also concerns that value added scores are too dependent upon students' raw attainment level (and thus influenced by differences in school intakes, just as raw-scores are) and that the calculations contain unacceptable levels of error (Gorard *et al.*, 2013).

This leads to the key research questions for this study:

- Is the volatility of value added scores over time an indication of genuine changes in school effectiveness?

- Can annual changes in value added scores for individual schools be predicted by expert knowledge of what is going on in each school?

- And so, are value added scores a meaningful indicator of school success?

This thesis will address these concerns by scrutinising the validity of Progress 8, the Department for Education's headline indicator of school performance in England. More specifically, it investigates whether the differences between schools' annual performance ratings and the change in schools' ratings over time can be explained by the kinds of factors that educational effectiveness is usually attributed to and perhaps more importantly, whether these factors are actually under the control of schools.

This is a worthwhile undertaking in itself because Progress 8 ratings have a significant impact upon the provision of English state education and all individuals that are involved with it. Not only do the results contribute to schools' inspection ratings (Ofsted, 2019), institutions have been offered funding or threatened with closure on the strength of their scores (Leckie and Goldstein, 2017; 2019), and parents are encouraged to use the ratings when making educational choices (Wilson, 2009). Heads and teachers therefore spend a significant amount of their time analysing departmental figures and adjusting their teaching accordingly (Daniel *et al.*, 2003). In principle this sounds like a good idea. However, if Progress 8 scores were shown to be misleading then these processes would waste valuable time and resources. They could even damage the very system that they are intended to improve.

The intention, though, is that the research findings will have far wider implications than this. Since all value added models rely upon the same basic principles, the issues that are raised during this investigation will apply to other value-added models and thus to the assessment protocols of other educational systems around the world including teacher effectiveness models and even equivalent schemes in other areas of policy such as public health. It is important to acknowledge however that the specific calculations that take place in these wider models will differ. It therefore falls to the reader to assess these differences, and so judge the extent to which the lessons from this new research apply more widely.

1.3 Area of Study

1.3.1. Focus and scope

All value-added models have the same basic objective, to identify the impact that educational factors have upon students' learning once extraneous influences have been removed. To evaluate the models, however, one must be more specific (Messick 1989a; 1989b; 1995; 1996a; 1996b). This section therefore specifies the ideas that will be considered in this thesis.

The first and most fundamental requirement is to define the terms 'learning' and 'effectiveness'. Both terms are discussed at length in the literature review (see Chapter 2). For now it should be sufficient to say that learning here refers to cognitive development of students, as measured by the academic progress that a student makes between their national Key Stage 2 and Key Stage 4 examinations. This definition therefore refers to students' cross-curricular learning (as opposed to subject specific learning) and excludes alternative conceptions such as affective learning, moral and/or cultural development which are not covered by the Key Stage Curricula. The term effectiveness is used to specify whether an educational body has been successful in helping students to make more than the expected amount of progress. It is therefore the overall quality of schools' provisions that is emphasised rather than schools' ability to promote the learning of particular sub-groups of pupils.

A second decision concerns the level of analysis. Value-added models are highly versatile. Depending on their specification, models can be used to evaluate the performance of students, teachers, departments, schools, educational systems or all of the above simultaneously. This thesis is primarily concerned with establishing whether Progress 8 data provides a valid measure of school-level performance. This is in line with the principal applications of Progress 8 and the predominant interest of the UK effectiveness literature (Acquah, 2013; Chapman *et al.*, 2011). To acknowledge the fact that the international research community considers teachers and classrooms to be the locus of the educational process (Scheerens, 1992; Harris, 2009), considerable attention is paid to instructional variables during all stages of the analysis.

A third issue was whether the study should place greater emphasis upon the theoretical/technical properties of school-level value-added scores or the validity of specific applications. This project deliberately prioritises the latter. Of particular concern is whether the ratings are valid, reliable and would help parents to select a more effective secondary school for their 11 year old child. Whilst other forms of value-added models are noted within the report, in-depth discussions of their mathematical foundations were considered beyond the scope of this study. The scope of the project is, however, broad enough to allow it to draw upon the findings of past research and also to discuss the key theoretical debates within the literature.

Finally, whilst this project draws upon material from the international school effectiveness literature, any discussions of educational policy and/or the implementation of value-added methods purposefully prioritise contemporary research that was conducted in England. All of the primary and secondary data used in empirical sections was also sourced from the same educational system. This was necessary as the properties of value-added are known to be context dependent (Teddlie and Reynolds, 2000). The inevitable consequence of this, however, is that one must consider the likely impact of contextual factors before inferring information from any of the aforementioned sources to other contexts. For the same reason, greater emphasis was also placed upon secondary education and studies that have employed traditional cross-sectional value-added designs as opposed to growth models. The same caveat therefore applies.

All of the aforementioned decisions were intended to align the research with the current use of Progress 8.

1.3.2. The rationale of value-added models

School effects are not a readily-observable or manifest quality of schools (Gorard, 2011c). In fact, attaining a valid and reliable measure of the contribution that schools make to their students' learning is more difficult than one might expect (Dumay *et al.*, 2014). Whilst school effectiveness can be conceptualised in a multitude of different ways, the most common method is to use test scores as a measure of cognitive learning (Creemers and Kyriakides, 2008). However, it is now acknowledged that comparing the raw-scores of each school is insufficient as these figures are heavily biased by school intakes. Each school will attract students with different attributes and abilities that influence their final attainment level (Goldstein and Woodhouse, 2000). An extreme example of this is provided by English grammar schools (Perry, 2019). Since these schools are selective they deliberately recruit students that they believe will excel academically. Their intake will therefore contain a higher portion of individuals with favourable characteristics such as intrinsic motivation, academic focus and prior learning. If at the end of Key Stage 4 these students reach a higher level of attainment than students of non-selective schools, we cannot necessarily attribute this to the quality of the schools' educational provisions. It could be that the grammar schools merely selected their students well. It is therefore unfair to judge schools on their raw output alone as their pupils may have vastly different starting points (Raudenbush, 2004). The disparity that is evident in this example, through pupil selection, occurs naturally and to varying extents between all schools. The location of a school, its policy for selecting pupils, its specialism and reputation all influence the type of pupils that enrol at a school and the amount of family support behind them. The composition of school intakes also varies significantly in terms of pupils' prior learning and indicators of disadvantage (Gorard and Cheng, 2011). This of course does not mean that schools are not differentially effective but indicates that we need a measure of scho ol effectiveness that is capable of differentiating between the effects of pupil intake and genuine school effects.

Value-added models were developed to address this need. Although a variety of models exist, featuring increasingly sophisticated predictions, they all rely upon the same fundamental principles (Teddlie and Reynolds, 2000). Rather than measuring students' raw attainment these models judge success by the progress students make whilst attending a school (Lubienski and Lubienski, 2006). Data on a ll pupils in the relevant school population is used to predict how well students should perform in later assessments (Gorard, 2010a). The difference between this estimate and the student's actual result is then used to judge how much progress the individual has made in comparison to similar individuals from other schools (Fitz-Gibbon, 1997). A positive residual indicates that the pupil has made more than the anticipated amount of progress, whilst a negative score shows that the pupil has made less progress that expected. A residual of zero indicates that the child's progression is in line with that of similar students from other schools. The results of these individual assessments are then averaged at school-level to provide a representation of how effective the school is in comparison to other institutions. In theory, by comparing the performance of students with that of comparable pupils, this process removes the effect of differential pupil intake to schools making value-added assessments a fairer and more valid measure of school performance (Rutter *et al.*, 1979; Sandoval-Hernandez, 2008).

1.3.3. The nature of value-added effects

Value-added assessments of school performance therefore seem like a good idea. The validity of the method however is not universally accepted (Hoyle and Robinson, 2003). Much of the debate stems from the fact that school influence is envisaged as a latent property that is revealed by the calculation itself (Gorard *et al.*, 2013). Once all extraneous influences have ostensibly been accounted for, it is presumed that any differences between the predicted and actual attainment of students are causally attributable to schools and thus that the scores represent the true contribution that each institution has made to its students' learning (Marsh *et al.*, 2011). It is important to recognise however that whilst all value-added models rely heavily upon this assumption, it will never truly be the case (Coe and Fitz-Gibbon, 1998). Accurately modelling the impact of all of the extraneous influences upon students' learning is a practical impossibility (Meyer, 1997) and the problem cannot be negated with technical solutions (Sammons *et al.*, 1996; Visscher, 2001; Creemers *et al.*, 2010; Goldstein, 1997). All value-added estimates will therefore contain any genuine educational effect and an error component (Gorard, 2010a) and can at best be seen as approximations of schools' contribution (Fitz-Gibbon, 1997). A key question addressed in this thesis is therefore whether it is justified to treat these differences as an effect rather than inaccuracies because the main threats to validity, such as omitted variable bias and measurement error are difficult to rule out. Important influences are often neglected (Dearden, Miranda and Rabe-Hesketh, 2011) and some factors may even be unmeasurable in a practical context (Tymms, 1996). All of this uncertainty makes it difficult to be certain which mechanisms are responsible for the differences in schools' ratings and crucially whether they are under the control of schools. This thesis pays particular attention to the Progress 8 measure and its role in the English secondary school accountability system. The problem, however, is common to all value-added models.

Unfortunately, there are a limited number approaches that can distinguish between school effects and error. The best evidence would theoretically come from experimental research designs that utilise randomisation (Goldstein and Spiegelhalter, 1996; Shadish *et al.*, 2002). Such methodologies would negate the need to evaluate the initial differences between students because all known and unknown confounds, including errors, could be balanced between the intervention and control groups. However, few researchers have attempted to implement true experimental designs as the random allocation of students to schools is considered unethical and infeasible in most circumstances. The only exception being a few teacher-level studies that have drawn inconsistent conclusions, see Section 5.3.5. Moreover, opportunities for natural experiments, where similar conditions are created by extraneous circumstances rather than deliberate intervention are too rare for such studies to make a substantial contribution.

Only two options therefore remain. The first is to directly observe the influence of measurement error and/or bias. This is problematic, however, as most inaccuracies are not visible to the researcher. One can assess whether a particular sub-groups of students appears to be disadvantaged after differences in their prior-attainment, personal characteristics and/or background have ostensibly been taken into account but these correlations do not prove that causal relationships exist (see, for example, Chetty *et al.*, 2014a). And whilst some researchers have sort to establish causation by artificially introducing errors into the analysis and observing their impact, their results are influenced by their own assumptions and the properties of the errors that they introduce (see, for example, Goldstein *et al.* 2008).

The most popular approach has therefore been to evaluate the effect of error indirectly by observing value-added results in context. In other words, to observe the stability, consistency and statistical significance of school effects when value-added assessments are performed with different students, at different times and in different settings (Luyten and Sammons, 2010; Creemers *et al.*, 2010). This approach appeals to the face-validity of the results by arguing that if all of the year-to-year variation in schools' performance ratings must ultimately be attributed to changes in schools' effectiveness or measurement error, then logically, if the volatility in schools' results is too great to be explained by genuine differences in school performance, one must conclude that measurement error had a substantive influence upon the ratings (Isaacs *et al.*, 2013). Likewise, the legitimacy of value-added figures would be considered suspect if the differences that occur within schools are too great to be ascribed to differential effects. These studies can however only provide what Rutter (1983) referred to as circumstantial evidence, so it is up to individual researchers to draw the line as to what level of inconsistency is indicative of genuine fluctuations and the amount that would be sufficient to question the validity of the underlying calculations. Whilst this is a weak form of evidence, it is the predominant source of information available and therefore constitutes a substantial area of interest within Educational Effectiveness Research. This material is reviewed in Chapter 5 of this thesis.

The largely philosophical decision, as to whether any inconsistencies in the results of value-added analyses are indicative of complexity or error is further complicated by the fact that value-added residuals have also been used to examine the stability and consistency of educational effects (Teddlie and Reynolds, 2000). Whilst all agree that value-added ratings are intended to reflect the proficiency of teachers and schools, and should therefore exhibit a degree of stability and/or consistency (Bosker and Scheerens, 1989; Scheerens, 1993), effectiveness researchers have adapted their conception of educational effects to the point that they are now viewed as being highly complex and multi-faceted (Chapman *et al.*, 2015). Students' responses to instruction are expected to vary not only based on the quality of schools' policies and practices, but also based upon the pupil groups, cohorts, curriculum stage and outcomes that the measure is applied to (Thomas, 2001). If one takes this to the extreme it is possible to imagine that a school may have a different effect upon every single pupil in attendance, with any inconsistencies viewed as the result of differences in the underlying conditions as opposed to errors. From such a position it becomes impossible to falsify value-added ratings (Gorard, 2011a;

Popper, 2015). In essence, the models presume that which they are supposed to be seeking (Gorard, 2010a). Furthermore, when value-added residuals are used to interpret the effectiveness of teachers, schools and/or larger educational bodies, even if the results are valid, who decides which sources the residual is attributed to? If the teachers in some schools were more effective than others but the within-school variance in teacher quality was small, this is likely to be interpreted as though it is the school that makes the difference. This inference, however, is debatable. The value-added methodology cannot help to make this kind of distinction, it can only inform us where the differences lie (Perry, 2016b).

Since the concept of effectiveness is constantly being adapted to explain unanticipated aspects of schools' results the researcher is left with little to base their decision upon other than their preconceived notions of what a school effect should look like. Researchers have therefore viewed the same empirical evidence and come to radically different conclusions. Proponents of value-added models such as Reynolds *et al.* 2012 (pp. 12, ln. 3-4), for example, have claimed that "across the dozen or so countries where ERR has mature research communities, there is so much independent agreement on the size of school effects, their scientific properties, the factors responsible for them" whilst critics argue that after "four decades of school effects research, we simply do not have much confidence that state educational agencies can identify value added at the school-level" (Kelly and Monczunski, 1997, pp.279 ln. 60-63).

A key part of any assessment of Progress 8's usefulness as an indicator of school performance and informant of parental choice must therefore consider how such profound disagreements arise. Chapter 6 therefore reviews the philosophical and methodological assumptions that underpin the debate, using the dialogue between Gorard (2010a; 2011a; 2011b; 2011c) and prominent educational effectiveness researchers (Muijs *et al.* 2011; Reynolds *et al.*, 2012) as a case study. This debate informs our discussion of measurement errors and has direct implications for the interpretation of instability and inconsistency in value-added ratings. It therefore portrays the extent of the disagreement, the implications for measures such as Progress 8 and assumptions that distinguish the two positions. The issues that are discussed however are far from new, so in addition to dealing with the specific problems, the interactions give voice to the type of incongruities that have plagued the field for some time.

This thesis contributes to the debate by evaluating whether the differences in schools' performance ratings can be explained by expert knowledge of what has been going on in schools. There were three strands to this assessment.

In the first, school leaders were asked to predict their schools' value-added score in advance, based on their in-depth knowledge of their school. Since these individuals are the ultimate authority on their institutions, one would anticipate that if the variation in value-added ratings were indicative of genuine changes in school effectiveness, then the endeavour would be reasonably successful. Even after we take into account that value-added ratings are a relative construct and that the precise Key Stage 4 attainment level required to achieve a particular ratings will vary slightly each year, it stands to reason that if the measure has any pragmatic value, those with the most informed opinion should be able to foresee, at the very least, any dramatic changes in their school 'effect'. The foresight of school leaders was thus evaluated and the implications for the practical application for Progress 8 considered.

In the second empirical section a thought experiment was conducted to assess the implications of there being inaccuracies in students' Key Stage 2 and Key Stage 4 data. More specifically, the DfE national attainment averages from 2019 were used to evaluate how a 10% measurement error in students' KS2 fine-levels would impact upon students' progression scores. The magnitude of these inaccuracies was then compared to the error that would occur if students' Attainment 8 score were over-stated by 10%. The more distinct the former is from the latter, the more differential the two effects were assumed to be. To the best of our knowledge, the relative effect of the two types of error has not been explored before.

In final strand, analyses 3 and 4, Progress 8 scores were modelled using key effectiveness factors from educational effectiveness literature. The results were then interpreted based on the scientific principle of falsification; that is to say, whether the variables interacted with school performance in a logical manner that was consistent with the findings of other research. Most importantly however the analyses identified the factors that could account for the highest proportion of the variation in schools' effectiveness scores and whether these are under the control of schools.

To summarise then, the debates outlined above mean that the validity of Progress 8 ratings and other value-added methodologies is not yet assured. Much of the research evidence that has been collected thus far is circumstantial and does not get to the heart of what the school residuals truly represent. Policies with such wide application and real-life consequences require a better research base than that. This is especially so since the questions raised in this section have the potential to connect all of the individual criticisms listed in Section 1.2.1. This thesis therefore takes school effectiveness research forward by addressing one of the of the field's core problems from an entirely new perspective.

1.4. Chapter Summaries

Table 1.4a summarises the objectives behind each chapter and the content covered:

Table 1.4a: Chapter aims and overview

Chapter 2: The Design of Value-Added Models

The thesis begins by providing a more in-depth introduction to the value-added methodology, including discussions of:

- The need for a fair measure of educational effectiveness
- The origins of value-added models
- The various specifications of model
- What they provide in conceptual and technical terms

In doing so, the chapter provides the pre-requisite material for latter discussion.

Chapter 3: Progress 8 and the Current DfE Secondary School Accountability System

The specification of Progress 8 is then presented along with an overview of the DfE secondary school accountability system. This includes:

- The historical development of the DfE value-added measures
- An introduction to the current indicators of secondary school performance
- A detailed walk-through on the calculation of Progress 8 scores
- A brief statement about the unique status of Progress 8

Chapter 4: Assessing the Validity of the DfE Value-Added Models

This chapter identifies some of the operational decisions that impact upon the validity value-added models. Including:

- The specification and modelling of extraneous variables
- The quality and completeness of underlying datasets
- The difficulty of distinguishing school-effect from extraneous influences

Particular attention is paid to the specifications of Progress 8 and comparable models of educational effectiveness.

Chapter 5: Indirect Evidence of Validity

This chapter scrutinises the volatility in value-added results. It considers:

- The stability of school effects over time
- The consistency of school effects across sub-groups and types of output

In doing so it establishes whether school residuals are stable enough to legitimise the construct of effectiveness and the DfE's use of value-added data.

Chapter 6: Methodological Assumptions and the Interpretation of Value-Added Evidence

Having reviewed the evidence-base for determining whether value-added models provide a valid measure of school effectiveness, this chapter considers how and why different conclusions have been drawn. The topics covered include:

- The problem with conceptualising effectiveness as a latent variable
- The properties of measurement errors
- The level of uncertainty in value-added results and how this should be expressed

Chapter 7: Educational Effectiveness Research and the Modelling of School Performance

The aim of the section is to identify the major factors that impact upon schools' performance.

- The chapter begins by reviewing the historical development of the educational effectiveness research and the types of variable that are believed to impact upon school performance.
- Creemers and Kyriakides (2008) Dynamic Model of Educational Effectiveness is then presented in detail. This model organises the most important correlates into an integrated framework that was utilised within the empirical analyses. It therefore provides the theoretical basis for latter sections.
- The empirical support for the model is then presented.

Chapter 8: Overview of the Empirical Sections

This chapter introduces the four empirical sections of the thesis. The discussion includes:

- A statement of intent
- An overview of their respective methodologies
- The details of changes that were necessitated by the Covid-19 pandemic.

Chapter 9: Prediction Analysis

Chapter 9 contains the first set of empirical analyses. This investigation established whether school leaders' knowledge of their school allowed them to anticipate changes in their schools' progress ratings. The report can be subdivided into four segments:

- Basic statistics describing the accuracy of school leaders' predictions
- An assessment of the relationship between school leaders' estimations and schools' progress scores
- An evaluation of the unique information that leaders predictions could provide
- An appraisal of leaders' ability to predict changes

The presence of strong and logical connections was interpreted as evidence Progress 8's validity.

Chapter 10: Thought Experiment

In the second empirical section a though experiment is presented. This investigated whether errors in students' prior and final attainment data have a comparable effect upon Progress 8 ratings. Specifically, the 2019 DfE attainment averages are used to assess the inaccuracy that would result if students Key Stage 2 and Key Stage 4 records were to over-state students' true performance level by 10%.

Chapter 11: Shallow Regression Analysis

In the third empirical section the relationship between key effectiveness factors and school performance was modelled. The primary focus was the relative contribution of factors that were within and outside of schools' control. The more favourable this ratio, the more valid Progress 8 ratings were assumed to be.

The investigation was broken down into three stages:

- Individual regression analyses that modelled the relationship between each effectiveness factors and schools' performance ratings when the influences of other factors was ignored.
- A multiple-regression model that assessed the influence of the 12 most influential variables.
- A hierarchical regression model that reported upon the influence of specified categories of variable (i.e. intake differences, classroom instructional practices, schools' pedagogical policies and schools' examination entry practices).

Chapter 12: Detailed Regression Analysis

In the preceding chapter the relationship between established effectiveness correlates and school performance was modelled. The intent behind this was to determine which factors had the greatest association with schools' progress 8 ratings and whether they were under the control of schools. The aforementioned analysis was limited, however, by the number of variables that could be operationalised. In this section a more in-depth case-study was undertaken with a small sample of schools. This enabled more explanatory variables to be assessed, including alternative dimension of the factors (i.e. the intent behind actions, their timing and the level of differentiation in place). All of these have the potential to influence school outcomes. The drawback of the reduced sample size however was that the impact of variables had to be evaluated individually using simple regression models. The analysis therefore provides only a general impression of whether school policies and practices are interacting with school performance in a logical manner as extraneous variables could not be accounted for.

Chapter 13: Conclusion

The final chapter collates the evidence from the empirical sections and interprets it alongside the findings of past research. This includes discussions of any methodological limitations and the identification of topics for further research.

A final judgement is then made as to whether Progress 8 is valid and reliable enough to perform the functions that the DfE have assigned to it.

2. The Design of Value-Added Models

2.1 Chapter Introduction

This section provides a more in-depth introduction to Progress 8 and the value-added methodology. The chapter begins by discussing the need for a fair measure of educational effectiveness, and moves to the origins of value-added models and what they provide in both conceptual and technical terms. The chapter therefore provides the pre-requisite material for latter discussion.

2.2 *The Need for a Measure of School Effectiveness*

Progress 8 was designed as a school performance indicator. Its intended function is to measure the effectiveness of all state-funded schools in England and report the findings in a form that facilitates the direct comparison of institutions (DfE, 2020). Before critiquing the calculation, however, it is useful to consider why such measures are needed.

Since the late 1970s the governance of western countries has been influenced by a conservative philosophy known as neoliberalism (Olssen and Peters, 2005). The UK in particular has reportedly embraced this ideology allowing it to reshape its economy (Palley, 2004). Aspects are particularly evident in the organization of public sector services including state funded education (Mulford, 2003).

The central tenets of neoliberalism can be understood at one level as a revival of the core beliefs of classical liberalism (Olssen and Peters, 2005). These centre around the proposition that markets provide the most moral and efficient way of allocating resources (Couldry, 2010). This is due to the nature of supply and demand (Hayek, 1945). Once a group's basic physiological needs have been satisfied it becomes problematic to anticipate the products and services that each individual will desire. Each person will have preferences which are known only by that individual. Proponents of the ideology therefore claim that it is impossible to create a comprehensive list of the commodities that a society desires as the required knowledge is dispersed across the population. An individual can extend their first-hand knowledge by communicating with others but can never attain a complete picture. It is also argued that statistical data is of little assistance as this type of analysis abstracted from the contextual information that is of interest. For this reason advocates of liberalism believe that a fair and cost-efficient distribution of resources can never be achieved by human planning. The planner's decisions will always be based on overgeneralized information containing inaccuracies that will ultimately lead to injustice and waste. In the context of the current study the assertion would therefore be that the state could never hope to co-ordinate the delivery of educational services that would efficiently address the public's needs as the desires of the population cannot be accurately summarised.

Instead they look to free markets for a solution (Hayek, 1944). The theory is that if service providers have to compete for paying customers then the resulting competition will simultaneously ensure that the public's needs are met and that the most effective and cost-efficient institutions will prosper. Advocates of liberal ideology therefore argue that such a system is fairer and would ultimately increase both the quality and specificity of the available provisions (Hayek, 1945).

Whilst there are clear similarities between the neo and classical liberal discourses, the two cannot be seen as identical (Olssen and Peters, 2005). Understanding the difference between them provides an important key for understanding how neoliberal beliefs have reshaped the organisation of state education and the role value-added models play within this new arrangement. The crucial departure is in how the two ideologies envisage the role of the state. Classical liberals, such as Friedrich Hayek (1899-1992) place such faith in the market's ability to self-regulate that the role of government is restricted to the protection of optimum market functioning. Monopolies, the existence of freely available public services and market externalities all distort market functioning and lead to under- or over-production (Hayek, 1944). This is often referred to as 'market failure'. The effects of these factors can be remedied by government regulation (Hayek, 1944). Further interventions however are seen as a market externalities that interfere with market functioning and are therefore discouraged. It is this commitment to the policy of non-interference that originally confined the use of markets to the private sector. Through the application of Public Choice Theory, however, neoliberalism validates an additional form of government intervention (Olssen and Peters, 2005). Buchanan (1975) made this possible by distinguishing between the 'protective' and the 'productive' state. Whilst in its law enforcement role, the protective state is still prohibited from manipulating an individual's rights, the productive state can utilize public opinion as a means of determining the most desirable way of distributing a public good (Buchanan, 1972). This legitimises the use of quasi-markets in the provision of state education. These are assumed to operate in a similar manner to traditional markets; the main difference being that public sector providers, in this case state-funded schools, compete for students and the governmental funding that accompanies them, rather than directly for income (Mulford, 2003). The state, however, retains a greater level of control as they can then manipulate the indicators that theoretically inform parents' educational decisions to ensure that their concerns are prioritised (see the discussion of DFE priorities within Section 3.3). They also retain control of the purse strings and can therefore decide whether and how to act on the information that the system provides.

It is important to note, however, that the expansion of markets to other areas of society necessitates that these sectors are re-interpreted in economic terms. From a neoliberal perspective, education is reinterpreted as a way for self-interested individuals to maximize their earning potential. Parents are therefore assumed to choose their child's school based solely on the anticipated effects to their child's career path. Buchanan (1975) fully acknowledged that this rather crude model oversimplifies the complex nature of human decision making but maintained that it held pragmatic value by helping to explain observed behaviour.

This neoliberal style of educational management rests on three further assumptions; that schools affect children's development, that these effects differ between institutions and that parents are able to identify these variations. The first claim is beyond dispute. Children that attend school will unquestionably acquire more academic knowledge than children that receive no formal tuition.

The second matter is less clear. The prevailing opinion is currently that it does matter which state-school a student attends (see House of Commons, 2009). Historically however this has not always been the case (see Teddlie and Reynolds, 2000). In fact, the change in perspective came about as a direct result of evidence that value-added models provided (Creemers and Kyriakides, 2008). It follows therefore that if the validity of these models is questioned, the evidence base would need to be re-examined. The final assumption however has received the most attention from policy makers. Prior to 1980, the organization of state education was not suited to providing parents with the information required to compare school performance. Like most of the public sector, education operated according to a bureaucratic system of delegated authority (Mulford, 2003). Each school had the power to decide what was best for their students and how to go about achieving it. This meant that school goals and curriculum content varied significantly. Educational provision therefore needed to be standardized to facilitate the process of comparing institutions. In 1988 the Education Reform Act established the framework for the first National Curriculum and the associated key stage testing. A school's performance could then be judged by the extent to which its students had learnt this content. The government acknowledges that these tests are not the output of education; they are a proxy measure of students' learning (House of Commons, 2008). They prioritize academic attainment as the most influential performance indicator because "pupils' life chances are to a great extent determined by their attainment in school" (House of Commons, 2009, pp. 63, ln. 23-24).

Schools' unadjusted attainment averages do not however provide a fair method of comparing institutions' effectiveness. In fact, they pre-dominantly report upon the pre-existing differences between schools' intakes (Teddlie and Reynolds, 2000). The DfE therefore introduced value-added models into the English Accountability System with the intent of providing a contextually independent measure of school performance that would inform parents' educational decisions, the allocation of support and funding decisions. The initial models, Value-Added (2002-2005), Contextualised Value-added (2006-2010) and Value-Added [Best 8] (2011-2015) were reported alongside other headline indicators. The current specification, Progress 8 (2016 onwards), however has surpassed even this standing and is now recognised as the headline indicator of educational effectiveness (DfE, 2020). Since it is these figures which shape the organization of state education, it could hardly be more concerning that the validity of the Progress 8 measure is questioned.

2.3 Origins of the Methodology

The term value-added has its origins in economics, where it refers to the difference between the sale price of an item and any financial inputs (Goldstein and Spiegelhalter, 1996). A positive value-added implies that an organisation has generated a profit and a negative score that they have incurred a loss. This essence of this definition is consistent with the educational use of the term, with one key difference. When one is dealing with the inanimate, the net profit in a transaction is calculable because it is easy to specify the financial outlay. The inputs of a school are harder to define.

As already discussed, all students possess a unique set of characteristics and skills that will facilitate or hinder their learning, regardless of their schools performance (OECD, 2008). School intakes cannot therefore be rendered homogenous by subtracting students' initial performance from the final attainment level. Were one to calculate students' absolute progress in this manner, the resulting residual would contain more components that the school effect. The students' score would be comprised of the school effects, random measurement error and extraneous sources of influence/bias. The foremost problem is that students with high prior-attainment tend to make greater academic progress than their peers (Ready, 2013). This has been a consistent finding in academic research (Dearden, Micklewright and Vignoles, 2011) and is reflected in the design of the national curriculum (TGAT, 1988). An absolute measure of value-added would therefore disadvantage students that start with lower levels of attainment. Pedagogical inputs, such as classroom instruction are also hard to differentiate in monetary terms, though early effectiveness models did try, albeit with limited success (Creemers and Kyriakidies, 2008). All of this makes an 'absolute' definition of value-added unworkable.

A valid measure of absolute progress would also require two things, a current- and prior-attainment rating recorded on the same scale and a way to separate the effect of the school from non-school factors (Cahan and Elbaz, 2000). The former is problematic in and of itself. The current Key Stage 2 and Key Stage 4 examinations, for example, do not meet this requirement. In addition to being measured on different scales, different numbers and types of subject areas are assessed using different examination boards. Practical qualifications such as BTECs are even equated with more academically orientated GCSE qualifications (see DfE, 2020). In terms of the latter, two approaches are possible. The best strategy would be to remove the effect of intake differences by design, ideally, using a randomised control trial that allocated pupils to schools (or the control group) at random. This however would be unethical in an educational context and opportunities for natural experimen ts are limited (Luyten *et al.*, 2005). The alternative is therefore to control statistically for any extraneous variables and regress students' learning gains upon their intake characteristics (Perry, 2016a). This would theoretically produce a contextually independent measure of school performance. However, the production of an absolute measure would once again necessitate that some student s did not attend school. This is because achievement gains do not only reflect the effect of intake differences and schooling, they also reflect maturation effects and other age-related factors such as informal education (Cahan and Elbaz, 2000). In order to create a valid measure of absolute school effectiveness the school effect would therefore have to be disentangled from these influences and this can only be done by studying the development of students who have not received instruction (Cahan and Elbaz, 2000). Neither approach is therefore acceptable.

To summarise then, in order to construct a traditional value-added measure of schools' 'absolute' performance the analyst would require data about students that did not receive any instruction during this period. This coupled with the demand for a continuous prior- and final-attainment metric forced an alternative approach. This adaption is discussed below.

2.4 Adapting Value-Added for use in Educational Contexts

What the stakeholders of education desire is a fair way of comparing school performance. In other words, a measure that is independent of the variables that schools cannot influence (SSCA, 1994). Recognising this is the key to understanding the specification of value-added models (Saunders, 1999). Whilst a number of factors prevent analysts from evaluating the absolute value added by schools in a like-for-like manner (see discussion above), the objective does not strictly require these problems to be resolved (Perry, 2016b). One can negate this issue by comparing the performance of statistically comparable pupils. Within the context of education the term value-added has therefore come to refer to the relative performance of schools, or how much progress students make at their school 'relative' to the progress they would be expected to make at an average or typical school (Kelly and Downey, 2011).

The nature of these comparisons varies from model to model. At their core however all value-added specifications are based upon the same premise, that pupils with similar characteristics have the same likelihood of making academic progress. At the very least these models therefore take into account differences in students' prior attainment which explains approximately 50% of the variance in students' raw results (Teddlie and Reynolds, 2000; Thomas 2001), though many consider additional factors to further distinguish between pupils. Since the effect of the main sources of bias have ostensibly been removed it then follows that if a student's final attainment level exceeds the average of their sub-group then their school must have made a greater than average contribution to their learning and vice versa. It important to recognise, therefore, that the models do not measure school effectiveness directly. What is termed the 'school effect' is, in actuality, just variance in students' performance not explained by prior attainment or by the context if that is also included (Gorard, 2010a; OECD, 2008). It is only the assumption of the researcher that attributes the effect to schools (Marsh *et al.*, 2011; Perry, 2016a). The validity of such models is thus contingent upon the adequacy of the statistical modelling (i.e. how well the model accounts for extraneous influences). As there will always be a degree of measurement error, however, the residuals can at best be considered as estimates of school performance (Visscher, 2001).

This re-invention of the term value-added for education is now widely accepted. The relative nature of the measurement and ambiguous use of terminology are however sources of confusion for some (Coe and Fitz-Gibbon, 1998; Goldstein, 1997; Luyten *et al.*, 2005). This confusion may stem from the fact that the 'school effect' is relative and centred on zero. It does not therefore refer to the level of absolute value-added or the actual progress made by pupils as the name implies. In fact, the DfE themselves have acknowledged the potential for misunderstanding and repeatedly stress the matter within the current guidelines. At one point they even went as far as adding an arbitrary score of either 100 or 1000 to schools' value-added scores, in order to discourage the notion that school with negative scores were failing (Ray, 2006). For this reason, some authors have suggested alternative descriptors such as "adjusted comparison" (Goldstein, 1997, p. 1997, ln. 23) or "adjusted academic performance" (Coe and Fitz-Gibbon, 1998, p.433, ln 24-25) as more appropriate.

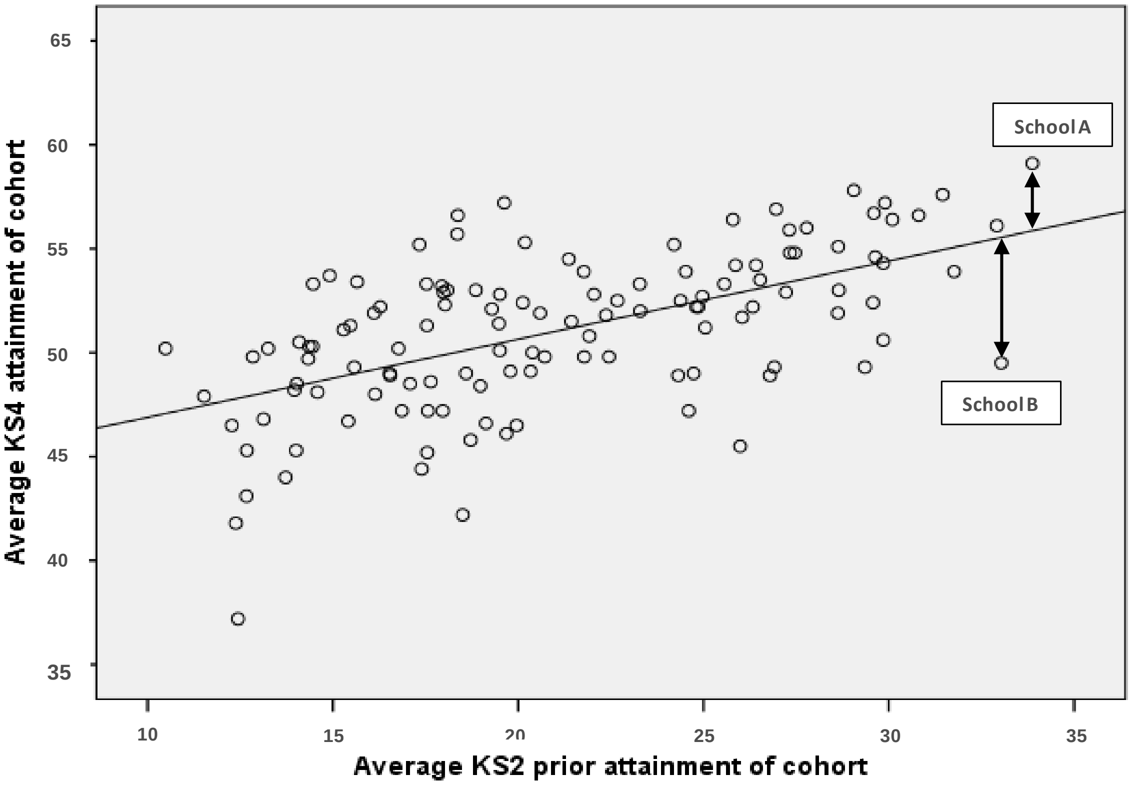
2.5 Types of Value-Added Model

Several classifications of value-added model are used within educational research and policy. Though their specifications vary, all attempt to provide contextually-independent measures of performance by controlling for extraneous influences (Teddlie and Reynolds, 2000).

2.5.1. School-level models

The first category of model uses aggregated data on non-school factors to remove bias from school-level attainment figures. To help explain this process an example is provided.

Figure 2.5.1a: Linear regression model of the relationship between KS2 and KS4 attainment



(Data artificially created for the purposes of this example)

As previously discussed differences in students' prior-attainment bias schools' raw attainment scores and prevent them from reflecting schools' true influence. Most school-level models will therefore include statistical controls which account for differences in students' starting point.

In statistical terms, the average final-attainment scores of all schools are regressed upon school-level aggregates of their students' average prior-attainment. This results in a function, such as the one depicted in Figure 2.5.1.a.

$$(1) \quad \widehat{Y} = \alpha + \beta\bar{x}$$

Where, $\hat{Y}$ is the projected final attainment average of a school, α is the intercept, $\bar{x}$ is the average prior-attainment level of the school, β is the prior-attainment co-efficient.

Each school's performance can then be described by the following equation:

$$(2) \quad \bar{Y}_j = \alpha + \beta\bar{x}_j + r_j + u_j \qquad j = 1, 2, \ldots k$$

Where $\bar{Y}_j$ is the mean final attainment in School j, α is a constant intercept, $\bar{x}_j$ is the average prior-attainment level at school j, β is the prior-attainment co-efficient, r_j is the school effect of school j and u_j is a random error term.

If, the $\alpha + \beta\bar{x}_j$ expression from Equation 1 is then substituted for $\widehat{Y}$ as they express the same thing, the difference between expected score of the school and their actual score can calculated as being:

$$(3) \quad \bar{Y}_j = \hat{Y}_j + r_j + u_j$$

or, in its reworked form

$$(4) \quad \bar{Y}_j - \hat{Y}_j = r_j + u_j$$

The performance of each school is thus judged by much how higher or lower the schools' final attainment level is than the score projected for the school by Equation 1. Within Figure 2.5.1.a, for example, the School A would receive positive rating (a score of 4) as their final attainment level is higher predicted and School B would receive a negative rating (a score of -6) as their score is lower than predicted.

All school-level value-added models rely upon this methodology. There are however several variations that may facilitate in the production more accurate estimations. The relationship between the independent (in this case average prior-attainment) and dependent variables (average final-attainment) can be modelled as a non-linear function (i.e. a curve). If this helps to account for a higher percentage of schools' results, then theoretically, it removes a higher proportion of the bias that the variable introduces. The models can also be extended so that the effect of several extraneous factors can be modelled simultaneously or to include interaction terms. By the same logic both should reduce the influence of non-school factors by providing a more accurate representation of their influence.

2.5.2. Pupil-level models

Comparable procedures can be implemented using pupil-level data.

The most common approach uses the same form of Ordinary Least Squared (OLS) regression that is utilised by school-level model. In this reiteration though students' individual final attainment scores are regressed upon their prior attainment, which results in Equation 5.

$$(5) \quad \hat{y}_{ij} = \alpha + \beta x_j \qquad i = 1, 2, \dots n. \quad j = 1, 2, \dots k$$

Where $\hat{y}_{ij}$ is the projected final attainment of student i in school j, α is a constant intercept, x_j is the student's prior attainment level x_j and β the co-efficient of prior-attainment.

The performance of the student can then be described as:

$$(6) \quad y_{ij} = \alpha + \beta x_j + r_j + u_j \qquad i = 1,2, \dots n. \quad j = 1,2, \dots k$$

Where y_{ij} is the final attainment level of pupil i in school j, which is described by a constant intercept α, the student's prior attainment level x_j, the co-efficient of prior-attainment, a student-level residual r_j (pupil-level 'value-added') and an error term u_j.

These models can likewise be extended to consider the effect of additional variables and any interaction that occurs between them. The results of all students are then collated to calculate an average value-added score for each school.

At this level however other statistical techniques can be used to predict students' scores including; lowess regression estimator methods, kernel regression, quantile regression and the new Progress 8 specification (see Section 3.3). Each variation has its own advantages and disadvantages. See Burgess and Thomson, (2013a; 2013b) and Kurtz (2018) for further details.

Despite early debates on the matter, when sufficient information is available the utilisation of student-level data is generally considered as being preferable to the school-level alternatives (Raudenbush and Willms, 1995). Whilst it is acceptable to use school-level aggregates to compare the relationships between macro-level constructs (Creemers and Kyriakides, 2008), the additional information available within pupil-level models allows the researcher to acknowledge the relationships that occur within schools (Aitkin and Longford, 1986). This broadens the range of analytical possibilities and makes for a more accurate[1] report on schools' contribution (Woodhouse and Goldstein, 1988). A school level-model could not, for example, report whether a school provided more effective instruction to a particular sub-group of pupils, whereas in pupil-level models this can be achieved by averaging and comparing the mean progress ratings of students' that share particular characteristics. Similarly, the researcher can report upon micro- (student-level) and meso-level interactions (classroom -level) interactions without fearing that the associations are not replicated in higher-level relationships (see Section 8.3 for further details). The models also tend to result in a higher percentage of variance being ascribed to the school as the aforementioned variance is acknowledged (Dettmers *et al.*, 2009) and one can more readily distinguish between meaningful and non-meaningful associations due to their higher sample sizes[1] (Woodhouse and Goldstein, 1988). From a pragmatic perspective, however, the matter has largely been rendered null by the advent of multi-level models that allow the two forms of data to

[1] Note that the term statistically significant is avoided here as the legitimacy of inferential statistics is challenged in latter sections. All other things being equal, however, higher sample sizes will still increase the representativeness of a sample.

be assessed within a single analysis (Perry, 2016b). These models, to which the discussion now turns, therefore dominate the academic literature (Teddlie and Reynolds, 2000). Though pupil-level design still feature within educational policy.

2.5.3. Multi-level models

Multi-level models were developed in the late 1980s as a way of acknowledging the hierarchical nature of educational effects (Creemers and Kyriakides, 2008). That is to say, the fact that students are clustered within schools and are therefore likely to have more in common with each other than students more generally (Snijders and Bosker, 2011).

In terms of their specification, the key distinction between single and multi-level models is that the latter partitions the residual variance into school- and pupil-level terms (Goldstein, 1997).

This can be seen by comparing Equation 6 with the simple multi-level formulation below:

$$(7)\ y_{ij} = \alpha + \beta x_{ij} + \bar{r}_{ij} + r_{ij} + u_{ij} \qquad i = 1,2, \dots n \quad j = 1,2 \dots k$$

Where y_{ij} is the final attainment level of pupil i in school j, α is a constant intercept, x_j the student's prior-attainment level, β the co-efficient of prior-attainment, $\bar{r}_{ij}$ the school-level deviation, r_{ij} the pupil-level deviation and u_j an error term.

The foremost difference between the two sets of specifications therefore concerns the $\bar{r}_{ij}$and r_{ij}terms, which describe the school-level variation (i.e. the school effect) and the pupil-level variation (how much the individual's residual differs from the average value for their school) (Goldstein, 1997).

The models are also more flexible. In addition to being amenable to the extensions mentioned in the previous sub-sections, the relationships can also be permitted to vary across schools (Snijders and Bosker, 2011). One can therefore evaluate, for instance, whether schools are more effective at instructing particular types of pupil, without the work-intensive calculations that are necessary with pupil-level models (Perry, 2016b), and can negate the of type misinterpretations can that theoretically occur in single-level analysis (Snijders and Bosker, 2011).

Proponents of the approach also argue that by acknowledging the non-independence of cases, the models produce more accurate estimations of standard errors that can more reliably distinguish between meaningful and coincidental associations (Aikin and Longford, 1986). The legitimacy of interferential statistics is however debated (see Section 6.3), which reduces the appeal of the models to some (see, Gorard, 2007).

2.5.4. Growth models

All of the value-added models discussed thus far are categorised as cross-sectional. This is because they evaluate the student outcomes at a single moment in time[2]. Growth models, however, measure student attainment across multiple years and model school effects based on the change in students' growth trajectories (Teddlie and Reynolds, 2000; Willms, 1992).

The basic idea underpinning the approach is that students will learn at different rates. Whilst most value-added models depict students' learning with a linear function (a straight line from their prior-attainment level to their final attainment level), the longitudinal nature of the assessment allows non-linear developments to be acknowledged. That is to say, that the addition of time-dependent variables allows the models to distinguish between students that initially make swift progress and then taper off over time and those which start slower and advanced more quickly during the later stages of their educational (Muthen and Khoo, 1998). This provides a more intricate depiction of students' learning that is useful in evaluating school effects and the differential performance of sub-groups (see, for example, van der Werf, *et al.*, 2008). The models can also be placed within a multi-level framework similar to that discussed above (see, Guldemond and Bosker, 2009).

It is important to recognise however that the definition of effectiveness that these models employ is notably different from that assumed under cross-sectional models. So whilst researchers such as Reynolds *et al.*, (2012) claim that this methodology is better suited to distinguish school effects, there are those that question the legitimacy of the models' explanatory power (Gorard 2011a; Perry, 2016b). The criticism being that although adding more variables and functional flexibility into a model is likely to lead to a higher percentage of variation being accounted for, this does not necessarily imply that the deviation is causal or conceptually significant. When interpreting the results of regression-based analyses it is therefore important to keep in mind that regression models are capable of 'explaining' 100% of the variance in an outcome variable, even if all of the numbers involved are made-up, random or meaningless (Gorard, 2008b).

2.6. Alternative Approaches: Regression Discontinuity Designs

Regression discontinuity designs are not classified as value-added models. They do however offer a method of evaluating the absolute benefit of one year's schooling.

In principle, these models define effectiveness as the absolute progress that a cohort of students makes during an academic year (Cahn and Elbaz, 2000). That is to say, as the difference between their aptitude in Year X and Year X-1. Since students' characteristics will remain stable during this time period (their gender, ethnicity, socio-economic status etc.) they should not, proponents argue, impact upon the achievement gains of the cohort. The benefits of this design therefore lie in it being a with-in school measure (Cahn and Elbaz, 2000).

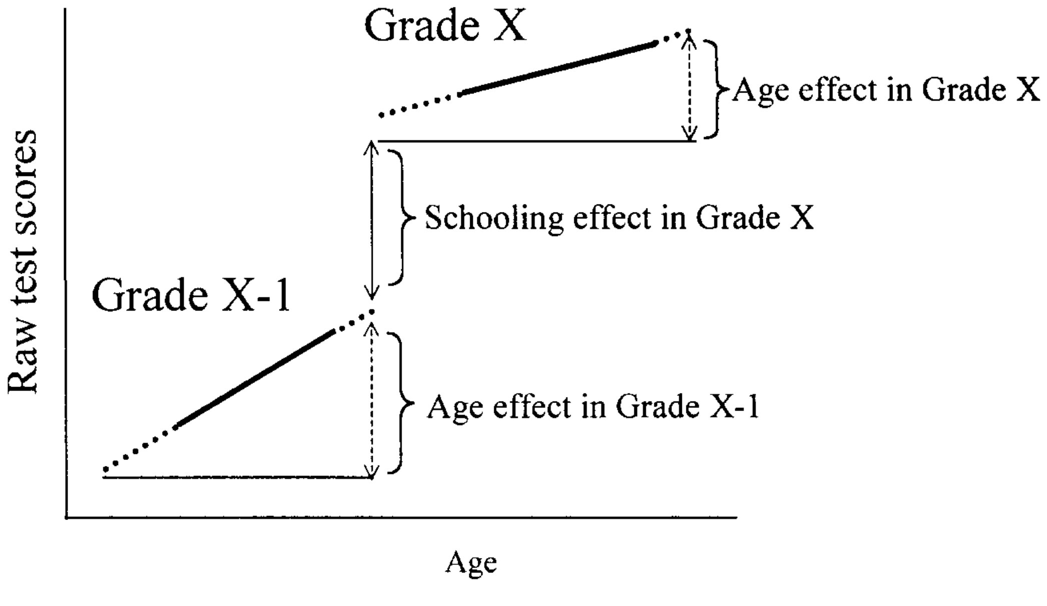
This is not, however, what is actually measured. Further assuming that that any differences in the make-up of consecutive cohorts will be negligible, allows this longitudinal design to be transformed into a cross-sectional one (Luyten, *et al.* 2009). Thus, in practice, it is the difference between the attainments of Cohort X and Cohort X-1 that is evaluated. Within the context of the English state-education, for example, one might compare the performance of Year 11 students with the

[2] Measures of prior-attainment are excluded from this statement as they act as an independent variable rather than the dependent variable.

performance of Year 10 students. It is important not to forget, though, that students' learning gains reflect more than their schools' effect. The age of the students' in the two cohorts differs, which means that maturation and other age-related factors (such as the receipt of informal education) must be accounted for. This is achieved using a between-grade quasi-experimental regression discontinuity design (Cahan and Davis, 1987). The legitimacy of this process, which involves issuing identical tests to the two groups of students, is contingent upon students' birth dates being governed by random events and upon students being allocated to cohort based solely upon their age. In other words, the models assume that students are allocated to year-groups through the application of an arbitrary cut-off and that progression through each year is then automatic, with no inter-school transfers, drop-outs or instances of students repeating or skipping a year's instruction (Cahn and Elbaz, 2000).

The difference between the mean test-scores of the two groups can then interpreted as being equal to the effect of one year's schooling plus the effect of age-related factors. Both of which are modelled using a regression discontinuity design (Cook and Campbell, 1979). Specifically, the influence of age is depicted by the slope of the within-year-group regression functions and the effect of schooling by the discontinuity between the two (see Figure 2.6a). That is to say, that the effect of age is reported by the difference in the mean predicted scores of the oldest and youngest students within each year-group, and that the effect of schooling is represented by the differences between the oldest student in lower year group and the youngest student in the higher (Cahn and Elbaz, 2000). For a more detailed description of this process see Cahan and Cohen (1989).

Figure 2.6a: The age and schooling effects in the between-grade regression discontinuity design – a hypothetical example



(Image taken from Cahan and Elbaz, 2000, p. 130)

Proponents of these models claim that this approach provides more valid and reliable indication of school performance (Luyten *et al.*, 2009). Firstly, because the intake differences that have plagued value-added designs have less impact upon inter-cohort performance and secondly because the natural experiment that the design capitalised on, i.e. the 'random allocation' of students into cohorts, will under ideal conditions isolate the effects of the remaining non-school factors (Cahn and Elbaz, 2000).

There is however one aspect of the model that should be brought to the reader's attention. Regression discontinuity designs are dependent upon their being a test that can accurately capture differences in cohorts' achievement. It is vital, though perhaps not intuitively obvious, that the assessment examines general skills that are not specific to the students' curricular. Neglecting this rule would lead to "test anchors" (Cahn and Elbaz, 2000, p. 133, ln 1). Were the test to be based upon the older year-group's curricular, for example, this would provide an unfair measure of the younger students' ability and the school effect would be overstated. Whereas basing the test upon the younger cohort's curricular would have the opposite effect. In addition to the assumptions listed above, the approach further assumes that ability and/or intelligence tests provide a suitable measure of schools' impact. Whilst there is evidence that is the case (see, Ceci, 1991; Ceci and Williams, 1997; Cahan and Cohan, 1989) and some have argued that it may even be questionable to attempt to distinguish between achievement and ability (Anastasi, 1984; Cronback, 1990), this is nevertheless a notable deviation from traditional assessment practices.

2.7 Type A and Type B Effects

Whilst the aforementioned models are intended to provide a fair and effective way of evaluating schools' impact, there are notable differences in their specifications. One of these distinctions is how they define effectiveness. In order to critique Progress 8 it is therefore necessary to be more precise about the effects that the model is intended to estimate.

Willms and Raudenbush (1989) identified two types of school effects that can be estimated by school accountability systems. These are closely related but are of interest to different types of educational stakeholders. To distinguish between them, one must first consider how school effects are brought about. Student performance is influenced by at least three factors; the students' personal characteristics, their school's practices and the context of their school. The crucial distinction between the two categories of effect concerns the last of these influences and whether or not this should be considered to be part of the school effect.

Type A effects:

Type A effects are intended to describe the difference between a student's final attainment level and the attainment that they would have achieved had they attended an average or 'typical' school. Since most of a student's personal characteristics (prior-attainment level, socio-economic status, ethnicity, gender etc.) are stable and remain the same no matter which institution they attend, the overall effect that a school has upon their learning can be defined as the combined influence of the school's practices (teaching methods, policies, and so forth) and the school context (the wider environment in which the school is located and the composition of the school's intake).

Estimates of Type A would comprise:

Type A effect = effect of school practice + effect of school context + measurement error

This information is of most interest to parents when they are selecting their child's school. These individuals are unlikely to be concerned about which characteristics of the school contribute to their development, only that it is maximised as much as possible (Raudenbush and Willms, 1995).

Type B effects:

Educational practitioners and policy makers however desire different information. Since teachers have little to no control over their school's social environment or the composition of its intakes (Coleman *et al.*, 1966; Willms, 1986; Lee and Bryk, 1989) it is deemed unfair to hold teachers to account for these effects. Teacher and schools should therefore be judged only on by the effectiveness of their practices once any differences in students' characteristics and the school context are controlled for (Raudenbush, 2004).

Thus estimates of Type B effects are comprised of:

Type B effect = effect of school practice + measurement error

The current DfE accountability system uses Progress 8 figures to report upon both types of effect. They are meant to act simultaneously as informants of parents' educational decisions and as a measure of the contribution that schools make to their students' learning. Since the model does not control for any contextual influences however it is, strictly speaking, only capable of reporting Type A effects. Thus even if the controls for student-level differences are adequate (which is far from certain), any categorisation of school practices will be biased in favour of advantaged schools. These types of observation, however, are not new and have been made previously with regards to former versions of DfE value-added models (see Kelly and Downey, 2010).

This being said, it should be noted that Raudenbush and Willms (1995) also assert that whilst it is possible to produce unbiased estimates of Type A effects using non-experimental designs. There is little prospect of producing unbiased Type B effects with the type of data available in accountability systems. As a minimum, the observation of instructional practices would be necessitated.

3. Progress 8 and the Current DfE Secondary School Accountability System

3.1 Chapter Introduction

This chapter outlines the role that value-added measures play within the English secondary school accountability system. Particular attention is payed to the calculation of Progress 8 figures and their interpretation, though details of past measures are included to provide some historical perspective.

3.2 The Evolution of the DfE Value-Added Measures

Value-added models were introduced into English state-education in 2002. The first model, national median line 'Value-Added' utilised a computationally simple algorithm that reported how much better or worse students performed in their 8 highest GCSE and equivalent qualifications, in comparison to the median score achieved by pupils with the same Key Stage 2 prior attainment. The scores of pupils were then averaged to provide a school-level value-added score that reflected the progress made by the pupils in schools' Year 11 cohorts. The initial specification, however, was criticised on two fronts (Leckie and Goldstein, 2017). Firstly, for failing to take into account other differences in school intake that continue to impact upon student achievement after their KS2 attainment had been controlled. And secondly, for failing to acknowledge that the resulting scores were estimates of school performance with a substantial margin of error.

In 2006, the initial model was replaced by Contextualised Value Added (CVA). CVA was intended to provide a fairer and more valid assessment of school performance that would not penalise schools with academically disadvantaged intakes (Kaliszewski *et al.*, 2017). It did this by utilising a multi-level specification which adjusted students' projected grades based not only upon their prior-attainment but also the average attainment level of students with similar background and personal characteristics. The additional considerations included student-level characteristics such as students' within year age, gender, ethnicity, whether students' spoke English as their first language and socio-economic status (Free School Meals status), as well as school-level influences such as the average prior-attainment level of their cohort and the affluence of the local area (Evans, 2008). Essentially, though, the model was meant to perform the same function, to provide an indication of how much progress students at each school had made in comparison to similar individuals in other schools. The resulting value-added scores were also presented alongside 95% confidence intervals that ostensibly quantified how cautious one should be when interpreting the figures.

Whilst many considered CVA to have been a vast improvement upon the previous methodology (Leckie and Goldstein, 2017), it was scrapped in 2011. At which point the DfE returned to using a simple model of value-added which did not take into account any compositional differences in school intakes beyond that of students' prior-attainment. This was also known as 'value-added' though we shall refer to it henceforth as 'Best 8 VA' to distinguish it from the first measure. As justification for this the following explanation was offered:

> "We will put an end to the current 'contextual value-added' (CVA) measure. This measure attempts to quantify how well a school does with its pupil population compared to pupils with similar characteristics nationally. However, the measure is difficult to

> understand, and recent research shows it to be a less strong predictor of success that raw attainment measures. It also has the effect of expecting different levels of progress from different groups of pupils on the basis of their ethnic background, or family circumstances, which we think is wrong in principle."
>
> (DfE, 2010c, p. 68, ln. 12-19)

The first of these criticisms certainly had merit. The procedure for calculating CVA figures was complex and almost certainly beyond the comprehension of parents and/or students who are unfamiliar with statistical regression, multi-level models and the assumptions underlying confidence intervals. In fact, it would not be too much of a stretch to assert that it was not fully understood by any teacher that did not have the specialist mathematical knowledge or experience of data analysis. As with all value-added models, the scores from one year were not directly comparable with those from the last and it was not even intuitively obvious upon what scale the scores were presented on (Leckie and Goldstein, 2017). That a 1000 points were arbitrarily added to each score (DfE, 2010a) to avoid the implication that some students had made no progress and that a 6 point increase in CVA equated to a student achieving one grade higher in each of the 8 qualifications was only presented within technical documentation (DfE, 2010b). There are, however, those that argue that a fine-level of understanding was not needed to grasp the conceptual intent behind the calculation and that the groundwork had therefore been laid to ensure that all individuals could interpret and act upon the results in an appropriate manner (Leckie and Goldstein, 2017). These individuals perceive the main problem to have been the way the information was summarised and presented to the public. The counter argument is that if models are so complex that stakeholders do not understand them to the point of being able to challenge them, then the benefit providing the information to these individuals is limited as they will be less likely to adapt their actions in an appropriate manner (Kelly and Downey, 2010).

The second justification, that CVA was an ineffective predictor of success, is unfortunately unclear as the Government did not cite the research to which it referred. If it was intended to express the fact that students' KS4 raw attainment are more accurately predicted by students' KS2 scores than their CVA ratings, then this would not discredit the measure (Leckie and Goldstein, 2017). In fact it would reflect the relatively small influence that schools have upon students' progress in comparison to the multitude of extraneous factors that impact upon their attainment (Rashbash *et al.*, 2010).

The final statement however was perhaps the most controversial (Bradbury, 2011). On the one hand, there is logic to the harsh argument that poverty is no excuse for failure. The proclamation is also well intended in its intent to change, rather than accept the inequalities that exist within society. That being said, the political palatability of asserting all individuals should be assumed to be capable of excelling does not negate the fact that schools often cater for vastly different populations of students. Many educational effectiveness researchers have therefore argued that returning to a point where only difference in prior-attainment are taken into account, reintroduced sources of intake bias that were previously removed and punishes schools that are responsible for the most academically disadvantaged students (Leckie and Goldstein, 2017).

3.3 The Current DfE Secondary School Performance Measures

The current system of secondary school accountability was introduced in 2016. It evaluates school performance using a variety of the measures which are subsequently published in the school performance tables (see DfE, 2019 for more information). The most influential however are the 6 headline measures of school performance.

These are outlined below:

1. Students' progress across 8 specified qualifications (Progress 8)
2. Students' attainment across the same 8 qualifications (Attainment 8)
3. Students' EBacc Average Point Score (APS)
4. The percentage of students entered for the English Baccalaureate
5. The percentage of pupils achieving a Strong Pass in English and Maths (grade 9-5)
6. The percentage of pupils who remained in education or found employment after completing Key Stage 4 (pupil destinations).

   All are aggregate measures calculated from students' individual-level data.

These indicators are used to evaluate the performance of all state secondary schools in England; that is to say, all state-funded schools (secondary, middle deemed secondary, all-through and 14-16 further education providers), academies and free schools, including special schools, pupil referral units and providers of alternative provision. In other words, any school that is directly or indirectly controlled by the state, that has pupils of the requisite age (Year 11 / 15-16 years old). Some independent schools also elect to use official state assessments, however, their results do not count towards national performance averages and are generally excluded from school performance tables.

DfE priorities:

As discussed, these 6 indicators are intended to act as a form of government control. It is therefore important to recognise that this selection of measures was deliberately chosen to encourage schools to provide "a broad and balanced curriculum with a focus on an academic core" (DfE, 2020, pp. 6, ln 19). The neoliberal view of education as preparation for employment is also reflected, both in the sixth measure and the focus on the English Baccalaureate - which was added to the list of headline indicators in 2010 as way of encouraging pupils from low socio-economic background to study for qualifications that would, it was argued, facilitate their progression into further education and employment.

From a methodological perspective it is also significant that the first three measures are evaluated on continuous scales. This demonstrates the Department for Education's (DfE) intention to recognise the development of all pupils, not just those at particular thresholds.

By way of contrast, measures 4-6 only recognise the progression of students that pass a particular boundary (e.g. those achieving a Grade 5 or above in Maths and English). These measures provide an effective method of reporting whether students meet a particular criterion, often a basic or floor standard that should ideally be achieved by the majority of pupils. By dichotomising students, however, these indicators can liken very disparate students. Measure 5, for example, would not distinguish between a pupil that had achieved a Level 1 in English and Maths and student who was operating at Level 4, though obviously the proficiency of the former student is far more concerning. Similarly, the measures will often abstract away from information that may be of value. For example, if a student had achieved a Grade 4 in English and a Grade 9 in Maths, the latter detail would not influence their schools' score. This makes them a less accurate indicator of schools' overall performance. Placing too much emphasis on threshold measures can also encourage schools to 'game the system' by paying greater attention to students that are close to key boundaries (NAO, 2003; West, 2010; West and Pennell, 2000; Wilson *et al.*, 2006).

Each of the 6 indicators is briefly discussed in turn below, with a Progress 8 described last and in more detail

Attainment 8

Attainment 8 is used in the calculation of students' Progress 8 scores but is also a headline line indicator of performance in its own right. The measure is used to summarise students' KS4 attainment. Only certain qualifications can count towards Attainment 8 however and these are weighted to encourage schools to adhere to the DfE's preferred curriculum. The four elements include:

1. The student's EBacc Maths qualification
2. The highest scoring EBacc English qualification (double weighted if both English Language and English literature are entered).
3. The three highest point scores from any remaining EBacc qualifications (EBacc maths and English qualification cannot count in these slots)
4. The three highest point scores from any remaining Ofqual approved qualifications (EBacc maths cannot count towards these slots).

These four subject-area groups are commonly referred to as the Attainment 8 'buckets'. If a pupil has not taken the maximum number of qualifications required to fill each bucket, any empty slots will receive a score of zero.

More detailed information on the inclusion criteria is available online (DfE, 2020). It is worth noting, however, that in 2015 the DfE began to reform GCSE qualifications from one using an A*- G grading system to a 1-9 system. This was done in stages over the course of 4 years. The old style of GCSEs were only eligible for these slots until the new qualifications are available. Additionally, early entry AS-level qualifications count towards the respective GCSE qualifications slots. This allows the achievements of advanced students to be recognised because high grades in these qualifications exceed the point scores available for GCSE qualifications.

This is an aggregate measure. The scores of all Year 11 pupils are therefore averaged to find the mean Attainment 8 score for the cohort in each school. This is the figure represented in performance tables.

EBacc Average Point Score

The EBacc Average Point Score provides a second composite measure of students' KS4 attainment. Each student's individual-level score is comprised of the point scores from the following subject areas:

- The highest point score achieved in either English language or English literature (both are a compulsory part of the EBacc).
- The student's point score for maths
- The two highest point scores in science (students must study three single science qualifications or the combined science award)
- The highest point score from geography or history
- The highest point score in a modern foreign language

The point scores from these qualifications are added, and then divided by 6 to produce the student's EBacc average point score. A score of zero is used whenever a student does not fill one of the elements.

The sum of pupils' Average EBacc point scores is then calculated and divided by the number of pupils in the school to produce the school's EBacc Average Point Score.

The percentage of students entering the English Baccalaureate

This indicator assesses the curriculum that students study. It represents the percentage of students entered for GCSE qualifications in maths, English Language and literature, two science qualifications, the humanities and a modern foreign language.

The percentage of pupils achieving grade 5 or above in English and Maths

This measure was also introduced into the Secondary School Performance Tables in 2017 in response to the GCSE reform. It reports the percentage of students at the school that achieved a grade 5 or above in both English and maths. The previous versions of the measure reported the number of student achieving a grade 4 or above, and before that the percentage achieving an A*-C.

Pupil destination measure

The pupil destination measure utilizes data from the National Pupil Database, Her Majesty's Revenue and Customs, the Department of Work and Pensions and local authorities, to report the percentage of students from each school that go on to sustained further education, employment or training. To be counted as sustained the graduate must participate in the activity for at least two school terms after leaving Key Stage 4. Additional breakdowns are available that specify the percentage of pupils moving onto further study, employment and training separately, and also the number students for whom data could not be found.

Progress 8

Progress 8 is the Department for Education's (DfE) headline measure of effectiveness and the primary basis upon which schools' performance is judged.

The measure is intended to report whether students made more or less progress between the end of Key Stage 2 (age 10-11) and the end of Key Stage 4 (age 15-16) than they would have if they had in theory attended another institution. More specifically it calculates the difference between each student's Key Stage 4 attainment (as assessed by attainment 8 – see description above) and that of all other pupils nationally with the same Key Stage 2 Average Point Score, and then averages these individual scores to get an aggregate rating for the school.

Since education in the UK is ordinarily split into five parts; early years, primary, secondary, further education and higher education. The assessed period will generally encompass the entirety of students' secondary-level education and is therefore intended to evaluate the amount of progress that students make whilst attending their secondary school. There are a number of reasons why this might not be the case however. These are discussed in due course.

The next sub-section will walk the reader through the calculation and the intended interpretation of schools' results. Before starting, however, there are two things to make clear. First, whilst schools' Progress 8 results are derived from individual-level progress data, no accountability is attached to the ratings at this level. Students' scores are calculated only as a means of evaluating their school's performance. And secondly, the process of comparing students' Key Stage 4 (KS4) results against other students' scores makes Progress 8 a relative performance measure. In other words, the raw-attainment level that is required to achieve a particular rating will change each year as the performance of other schools will deviate. This has obvious drawbacks. As explained in the introduction, however (see Section 1.3.2), the process is also instrumental in negating the bias introduced by school intakes. It is a key feature of the model's design which theoretically makes for a fairer and more valid assessment of schools' contributions.

The calculation of Progress 8 scores

A student's Progress 8 score is defined as their Attainment 8 score minus the average Attainment 8 score of all students nationally with the same Key Stage 2 (KS2) Average Point Score[3]. The higher this value, the greater progress the student has made in comparison to similar pupils from all schools.

Once the individual-level progress of all eligible pupils have been calculated the scores are then aggregated to give a Progress 8 rating for the whole school.

These are generally interpreted in the following manner.

- A positive score indicates that pupils at the school made more progress, on average, than pupils across England with comparable KS2 prior-attainment scores.
- A score of zero signifies that pupils made the same progress, on average, as pupils with comparable KS2 prior-attainment scores.
- A negative score suggests that pupils made less progress, on average, than pupils with comparable KS2 prior-attainment scores.

[3] Students' KS2 Average Point Scores are calculated by averaging their maths and reading fine-levels.

In fact, because of the way the score is calculated the score tells us how far above or below the expected performance level students' attainment tended to be. A rating of 1.0 for example signifies that on average students at the school achieved a full grade-point higher per subject than comparable pupils from other schools.

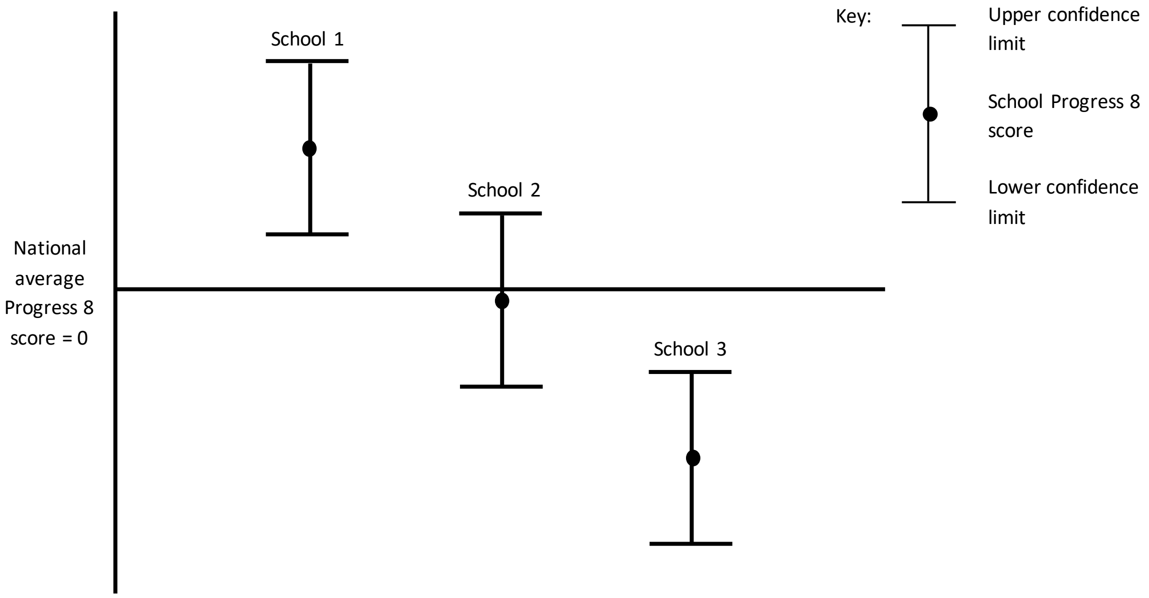
Confidence intervals

The true implications of schools' ratings are not, however, that clear. To understand why one must recall that each school's Progress 8 score is based upon the performance of a finite group of students (specifically the school's Year 11 cohort). There is therefore a chance that these students are not typical cases and that their Key Stage 4 results would have been higher or lower than other students irrespective of the school's influence. To account for this possibility and the fact that this risk is elevated when schools have a small Year 11 cohort (see Gorard *et al.*, 2013), the DfE calculates 95% confidence intervals for each school that act as a proxy for the plausible range of values within which the school's true effectiveness rating can be assumed to lie.

The upper and lower limits of this confidence limit are defined as the school's official Progress 8 rating plus or minus their C.I. value. Where C.I. = 1.96*(standard deviation of the Progress 8 scores for all eligible students nationally, divided by the square root of the number of eligible pupils at the school).

Schools are then viewed as being distinguishable from the national average only if the entirety of this confidence interval is above or below zero, as depicted in Figure 3.3a.

Figure 3.3a: The proposed interpretation of schools' Progress 8 confidence intervals



In this example both the upper and lower confidence limits of school 1 are above the national average. This school is therefore considered to be differentially effective. Similarly both of the upper and lower confidence limits of school 3 score are below zero so the school is considered to be less effective than the norm. School 2, however, has a confidence limit either side of zero, so whilst one cannot assume that the school's effectiveness score is exactly in-line with the national average, its contribution cannot be clearly distinguished from that score.

It is important to recognise, however, that one can define the population of 'other students' as the total population of students that currently attend the school, the students that could theoretically have attended the school, or as students that may hypothetically attend the school in the future. The latter two definitions refer to hypothetical super-populations (Muijs *et al.*, 2011). Super-populations do not actually exist. The objective is therefore to model the characteristics of the underlying relationships so that the inference of information beyond its original context can be justified. In this instance to imply how effective the school is likely to be in instructing students that have not actually attended the institution. This unusual practice essentially treats the differences in school intakes as random sampling error. Or to paraphrase, it views the schools' actual cohort performance data as one of many (hypothetical) random samples. Creemers et al. (2010) and Plewis and Fielding (2003) have all argued in favour of this approach, whilst others have challenged the practice (Gorard 2010b). The debate as to the legitimacy of treating the differences in schools' intakes as random sampling error is discussed in later chapters (see Section 6.3). For now, however, it is sufficient to understand that these confidence intervals are intended to inform educational stakeholders whether they can be confident that a school's performance was actually above average.

Recent changes to the Progress 8 measure

During this study there was a minor change in the way schools' Progress 8 scores were calculated. In 2018 the process was amended so that the scores of extremely low-performing pupils would be capped in order to prevent them from having a disproportionate effect upon schools' ratings. The threshold for making this determination was however set so that it would affect only 1% of students each year. The change did not therefore impact upon the interpretation of the analyses in this thesis[4].

[4] This introduction refers to the headline measures of secondary school performance at the time the empirical analyses were completed. These are generally consistent with the measures that were in place during the data collection phases of the study, with two notable exceptions. First, in 2018 the calculation of Progress 8 was amended so that the scores of extremely low performing pupils would be capped to prevent them from having a disproportionate effect upon schools' ratings. This means that the 2017 and 2018 ratings that are cited throughout this thesis are not exact equivalents. Second, prior to 2018 the English Baccalaureate APS and pupil destinations measures were not recognised as headline indicators. The percentage of pupils achieving the English Baccalaureate was reported in its place. Furthermore, in 2016-17 and 2017-18, when the analysis was conducted, the calculation was based upon students 'KS2 Average Point Scores' which were computed using students' KS2 'fine-levels'. Since 2019 however the metric that has been used has been students' KS2 Average Scaled Scores. These are based on students KS2 'fine-grades'. The two sets of terms refer to slightly different methods of converting students' KS2 performance in maths and reading into a common metric. This had no impact upon our analysis as the switch occurred after the completion of our analysis.

3.4 The Unique Status of Progress 8

The above discussion should make clear that Progress 8 has a unique status within the English secondary school accountability system. It is regarded as the headline indicator of school performance and is intended to be the primary measure by which state schools are judged. It is also the only indicator that claims to provide a fair measure of school effectiveness. That is to say, it is meant to be a metric that is not influenced by the composition of a school's intake or contextual factors such as its wider educational environment.

However, whilst the specification of Progress 8 is comparable to that of 'Best 8' and 'Value-Added' it does not attempt to control for many of the influences that Contextualised Value-Added accounted for. It is also more specific about the qualifications that can and cannot count towards students' scores and in the strictest sense, provides more valid source of information on the overall effect of attending one school over another than on the quality of schools' provision (see section 2.7). The implications of these statements are as yet unknown and will be considered in the remaining sections of this thesis.

4. Assessing the Validity of the DfE Value-Added Measures

4.1. Chapter Introduction

This chapter reviews the educational effectiveness literature, focusing specifically on the limitations of Progress 8 and some reasons to doubt value-added scores. The discussion can be broken down into two parts:

The first segment re-iterates that Progress 8 is a relative measure of performance. Whilst this does not threaten the model's validity, the matter is discussed briefly as this characteristic dictates the type information that can be attained from assessments and may mask other problems with the measure (Gorard, 2010a).

The second segment considers some the operational decisions that impact upon the validity of value-added ratings and the implications for Progress 8 assessments. These issues have the capacity to undermine the claim that value-added approaches provide a fair and accurate measure of school effectiveness. It is difficult to assess the overall effect of these biases however as it is impossible to model all of the extraneous factors that impact upon students' learning.

4.2. The Drawbacks of Relative Measures

One of the fundamental limitations of value-added models is that they provide a relative measure of effectiveness. This means that they report whether each school performed better or worse than other institutions once differences in their intake have been taken into account. While there are instances when it is useful to identify differentially effective schools, the approach has obvious drawbacks. The most notable being that the results tell us nothing about the absolute performance of schools. Within any value-added assessment, roughly half of schools will receive a positive rating and half will receive a negative rating. Likewise, the raw-attainment level required to achieve a particular rating will change each year as the performance of other institutions will vary. It is therefore possible, for example, for a school to improve and receive a less favourable rating. In fact, even if all schools improved, half would still receive negative ratings. This shortfall limits the usefulness of Progress 8 ratings and from a research perspective makes it more difficult validate the results. It is also significant that there is nothing to calibrate value-added measures against. The two issues are inter-connected. The latter however is the root cause of much educational debate (see Chapter 6).

Researchers have also argued that the competition that this form of monitoring system creates may have a negative effect upon long-term educational agendas, such as the need to develop the national pool of high quality teachers and leaders (Greany, 2017). Especially, if it is accompanied by reduction in inter-school collaboration and/or support from educational authorities.

4.3. Direct Threats to the Validity of Value-Added Models

Value-added models were created so that the true effect of schools could be evaluated. A key assumption that underpins their methodology is therefore the assertion that all extraneous influences upon students' performance can be controlled (Coe and Fitz-Gibbon, 1998; Marsh *et al.*, 2011). In fact, the validity of value-added estimates is dependent upon this being the case (Burges and Thompson, 2013b). Accordingly, any shortfalls in the modelling of non-school factors lead to bias or unexplained variance that is wrongly attributed to schools. What is more, since students with advantageous characteristics tend to be clustered in particular schools, the use of inadequate controls gives undue recognition to schools with advantaged intakes and punishes those that cater for the most vulnerable students (Burgess and Thompson, 2013a).

Controlling for non-school factors however is a complex and imperfect process (Saunders, 1999), and the topic has been the focus of a great deal of methodological research (e.g. Aitkin and Longford, 1986; Bosker and Scheerens, 1994; Goldstein, 1997; Hill and Rowe, 1996; Raudenbush and Willms, 1995; Snijders and Bosker, 2011; Timmermans *et al.*, 2011). As stated previously, it is important to recognise that the process of controlling for bias cannot be reduced to an entirely technical endeavour (Creemers *et al.*, 2010; Goldstein, 1997; Sammons, 1996; Visscher, 2001). To help understand the problem three types of issue are often cited – the technical problem of modelling the most important influences upon performance, the quality of the underlying datasets, and the theoretical problem of distinguishing between school- and non-school factors (Perry, 2016a).

4.3.1. Technical problems of model specification

In order for a value-added model to adequately control for non-school factors two criteria must be met; the model must include all student-level characteristics or contextual variables that have a substantial impact upon school performance, and the relationship between the non-school variables and performance must be adequately specified (Ladd and Walsh, 2002).

i. Omitted variables

The problem of measurement bias can be imagined as a continuum (Meyer, 1997). At one end is students' raw attainment and at the other a perfect measure of school effects (i.e. isolated from the effect of all non-school factors). If one ignores potential complications such as the propagation of measurement error (see later sections or Gorard 2010a), then it follows that correctly identifying and modelling the impact of extraneous variables will increase the validity of the model by moving it toward the latter end of this continuum.

Attempting to eliminate all extraneous sources of bias however requires a lot of data (Goldstein, 1997; Coe and Fitz-Gibbon, 1998; Gorard, 2010a). Contextualised value-added (implemented in England 2005-2010), for example, included measures of deprivation, ethnicity, English language status, gender, SEN status, in care status, age within year group, pupil mobility and school average prior attainment (Evans, 2008). Yet even this list is did not come close to operationalising all of the external factors that impact upon students' attainment.

Progress 8 (and its predecessors 'Best 8' (implemented 2011-2015) and Value-added (2002-2005)), however, deliberately utilises a simpler approach that only takes into account differences in students' KS2 Average Point Scores. In fact, this was explicitly stated as a requirement for its design (Burgess and Thompson, 2013a). The logic being that a measure of prior-attainment can not only encapsulate the direct benefits of having a higher level of pre-requisite knowledge but also, indirectly, the influence of that any background factors have had up until that point (Burgess and Thompson, 2013a). The hope was therefore that the model would still provide fair and valid assessment but be easier for educational stakeholders to understand and engage with (Kelly and Downey, 2010). There were also political reasons for not wishing to imply that it was acceptable for some sections of society to achieve less than others (see Section 3.2). Whilst the impact of the discontinued CVA indicators is small in comparison to prior attainment, past research has shown they still account for variance that cannot be explained by students' Key Stage 2 performance (see Gorard, 2006b; Teddlie and Reynolds, 2000). Within disadvantaged schools, their impact can therefore be substantial as students' with similar characteristics tend to be clustered within particular types of school (Perry, 2016a). Progress 8 assessments are therefore knowingly biased in a way that is detrimental not only to their accuracy but also the fairness of the measure (Burgess and Thompson, 2013a). What is more, these biases are not random and statistical techniques cannot estimate or remove the effects (Gorard, 2010b).

A powerful example of this shortfall is provided by state-funded special schools, which cater exclusively for students with moderate to severe learning disabilities. These institutions are required to take part in Progress 8 assessments. Despite this, however, the attainment scores of their students' are excluded from the calculation of the annual Key Stage 4 attainment averages (see DfE, 2020). Presumably this is because the DfE acknowledge that these individuals have additional barriers to their learning, which means that as a group they are likely to make less academic progress than their peers, even after differences in prior-attainment have been taken into account. Whilst this provides a better basis for comparing the performance of students' without special educational needs, it should not be forgotten that if prior-attainment acted as an effective proxy for all other non-school influences, one would expect to find that as many SEN students received positive progression ratings as negative. It is therefore shocking that in 2018, none of the 742 state-funded special schools received a positive Progress 8 score (see EduBase, 2018). The overwhelming bias introduced by school intake therefore makes the measure a profoundly unfair basis for judging the performance of these institutions, both in relation to mainstream scores and one another.

One is forced to conclude therefore that Progress 8 may disadvantage particular types of school, particularly those with disadvantaged intakes. The extent of this bias is investigated further in the empirical sections of this thesis.

ii. Specification of the relationship between prior-attainment and performance

The next question is how well Progress 8 models the relationship between prior attainment and performance. To do this effectively the model must utilise an appropriate functional form (Ladd and Walsh, 2002). This may sound complex but merely means that the relationship between prior-attainment and performance needs to be well represented by the mathematical trend that the model estimates. An example of an inappropriate functional form, for instance, would be the fitting of a linear regression line to a curvilinear relationship. This would lead to students' Key Stage 4 performance being under- and over-estimated at different points on the prior-attainment scale.

Fortunately, Burgess and Thomson (2013b) evaluated the fit that could be achieved by various statistical and non-statistical techniques while the Progress 8 model was being developed. Their study considered cubic-piecewise and percentile-based versions of Ordinary Least Squares regression, simple and extended versions of piecewise regression, multi-level regression, the Kernel approach, Lowness and the Quantile method. Each was shown to have its own strengths and weakness, and a vulnerability to different types of bias. Whilst the best of these was able to explain 58.3% of the variation in the final attainment score at mainstream schools, the pair concluded that simple piecewise regression[5] was

[5] This estimation model produces an irregular function by plotting the average KS4 attainment level for each KS2 fine-grade. See Section 3.3 or DfE (2020) for further details.

preferable as it accounted for a similar portion of the variation in results, 57.9%, yet was significantly easier for educational stakeholders to understand. The models also had comparable root mean square residuals and residuals that were unbiased by prior attainment. Progress 8 appears to adequately model the fixed effect of prior attainment.

4.3.2. The quality of underlying datasets

Inaccurate model specification is not however the only source of error and bias. Value-added designs are also heavily dependent upon the quality of the collected data. Both the validity of the outcome measure and any measures of extraneous variables will impact upon schools' apparent ratings (Goldstein, 1997; Gorard, 2010a). In the case of Progress 8 this means considering the adequacy of students' Key Stage 2 and 4 assessments. If either provides an imperfect summary of the factors that they are intended to summarise (students' prior and current knowledge base respectively) then this will add additional construct irrelevant variance into the analysis. Thus, it is not just the omission of important variables that will impact upon schools' ratings, the operationalisation of recognised variables also matters.

The problem can be broken down into two segments which are synonymous with those from the previous section; data availability and coverage of the underlying constructs. These two matters are discussed below.

A separate section, however, is dedicated to measurement errors as the characteristics of these inaccuracies are unique. The topic also plays a central part in latter discussion.

i. Data availability

The English Key Stage Assessment System collects standardised attainment data from all (or at least most) students as they enter and exit secondary education. In the vast majority of cases the tests will therefore act as well-timed measures of any pre- and post- instruction differences in performance.

An exception exists, though, for middle-deemed secondary schools that only educate student for part of this period. In such instances the Key Stage 2 assessments provide an ineffective measure of students' prior attainment as the timing of the tests and their transition between schools will not coincide. In fact, in most instances student would not join such a school until 2-3 years after the assessment. The progress scores for middle-deemed secondary schools are therefore less valid as they are partially dependent on the quality of their feeder schools. The DfE recognise this and recommend that stakeholders use alternative measures of effectiveness to judge the quality of these institutions (DfE, 2020).

It is important to recognise, however, that even when the required data is theoretically available there will be substantial portions of missing data. See Section 4.3.2.iii for details.

ii. Coverage of the underlying constructs

A second issue is whether the operationalised variables provide adequate representation of the construct that they are intended to measure. This is as much a conceptual problem as a practical issue and it is therefore a mistake to view the issue as merely a matter of data collection and variable specification (Fitz-Gibbon, 1996; Tymms, 1996; Willms, 2003). Once again, the simplicity of Progress 8 shortens the discussion considerably as only the adequacy of the prior- and final-attainment variables are of concern.

The aptitude of students

The pre-existing cognitive ability of students can be modelled in two ways, using one or both standardised tests of ability (tests of general intelligence) or students' prior-attainment data (subject specific knowledge, though the data can be aggregated or averaged across several subjects). Together these two factors are said to determine student aptitude (Creemers and Kyriakides, 2008), or the speed with which they will assimilate new information (Carrol, 1963). Strictly speaking therefore, the DfE's measure of prior attainment only covers one of these two aspects. This fact was demonstrated by Strand (2006) who showed that Cognitive Ability Tests (CAT) were more effective predictors of secondary school performance than Key Stage 2 test scores, yet a combination of the two metrics provided a more effective predictor than either of the individual measures. This result implies that both transferable learning skills (as assessed by the CAT tests) and prior-learning in specific curriculum areas (as assessed by the KS2 examinations) are required during the secondary phase of education. One can surmise therefore that neither test alone captures the entirety of student pre-existing aptitude.

Likewise, the coverage of the prior- and current-attainment measures is not the same. At present the English measures of Key Stage 2 prior attainment only considers the average finely-grade point scores of students in maths and English. At secondary level students' attainment is summarised using a weighted average of student's performance across 8 subject areas. It stands to reason therefore that the two measures are not exact equivalents of one another. This disparity was however introduced deliberately as the initial performance of students in other subject areas is less predictive of their Key Stage 4 attainment (Fitz-Gibbon, 1997).

Having said all of this, decades of research have demonstrated that Key Stage 2 prior-attainment figures tend to predict around 50% of the variance in students' Key Stage 4 attainment (Kelly and Downey, 2010). Despite the specification problems, one can therefore be assured that controlling for any pre-existing differences in test scores does make for a fairer assessment of school performance.

The final attainment-level of students

Competency is an exceptionally difficult concept to assess. The issue becomes even more complex, however, if the model in question is intended to assess cross-curricular learning (Coe, 2010; Bell *et al.*, 2007; Ray, 2006). In Progress 8 calculations, for example, Attainment 8 (the indicator of Key Stage 4 attainment) is a composite measure that summarises students' performance across multiple subject areas. The comparability of students' final attainment scores therefore relies upon all GCSE and equivalent qualifications being precisely aligned so that students' examination outputs can be converted into a common metric. This is not an easy task as each form of accreditation must receive appropriate recognition. The DfE and Ofqual (the exam regulator) must consider the quantity and difficulty of the content that students will study, the type of assessment that is used and the comparability of standards across different exam boards. Whilst these matters are assessed on a continual basis and have long been considered to be satisfactorily homogenous for the purpose of value-added assessments (Fitz-Gibbon, 1997) there are no obvious or permanent solutions and any misalignment will introduce non-random bias into the results. In England, drastic misalignments are rare but it is still easy to identify problematic topics. See for example the recent controversy concerning early entry GCSEs (Harrison, September 29th 2013), the European Computer Driving Licence (ECDL) qualification (Data Educator, April 20th 2018; Ing, 4th September 2018), the apparent abuse of vocational qualifications (Spielman, 10th March 2017) and concerns about the difficulty of the new reformed GCSEs. The comparability raw-attainment scores from different qualifications is therefore viewed as an ongoing and unsolvable issue that needed to be addressed. During the empirical sections of this thesis, schools' curricular are examined to consider the extent to which differences in schools' examination entries influence their progress scores.

Based on these discussions it is possible to tentatively conclude that students' initial aptitude and final attainment level have been appropriately operationalised. When considering the operationalisation of value-added models, however, it is important to keep in mind that no examination perfectly quantifies students' knowledge. One must not therefore confuse the operationalised version of a construct with the real thing. At best value-added models therefore provide or encompass an estimate of school effects (Coe and Fitz-Gibbon, 1998). The pivotal question however is how much the inevitable imperfections contaminate or even constitute the school effect (Gorard, 2010a).

iii. Measurement error

This section discusses measurement error and the threat that it poses for the validity of value-added measures.

Strictly speaking measurement error is merely a specific form of construct irrelevant variance. That is to say, that one could view it as being a non-school factor and consider the arguments from previous sub-section, especially in the case of Progress 8 which only controls for student-level differences in prior attainment. That being said, we discuss the matter separately because the nature of these errors is unique. This has led researchers to draw radically different conclusions about the implications for value-added models and the field of school effectiveness as a whole.

To facilitate a clear discussion it is necessary to define certain methodological terms. Specifically our use of the phrases *construct irrelevant variance*, *error*, *bias* and *measurement error*. The key distinction between these lies in the type of variation that they refer to. In line with the precedent set by Amrein-Beardsley (2014) these expressions refer to the following. Construct irrelevant variance is used as an overarching term that refers to any random and non-random sources of inaccuracy. Error here implies that inaccuracy occurs randomly, whilst bias implies that an observed (measured factors that are included in the model) or unobserved (unknown and/or unmeasured factors) and non-random mechanism is influencing the results. The nature of measurement errors is disputed (see discussions below). This final term may therefore refer to random and/or non-random error. All theorists agree however that in real-world situations measurement error will contain both random and non-random elements. It is therefore the ratio of these elements that is debated.

The validity of attainment measures underpins the calculation of value-added models (Meyer, 1997). Without accurate and reliable measures of student performance the whole process is a non-starter. An in-depth consideration of the accuracy, reliability and comparability of specific KS2 and KS4 tests is however beyond the purview of this thesis. This review must therefore trust in previous assessments of their appropriateness (e.g. Fitz-Gibbon, 1997a; Stand, 2006). That being said, one must acknowledge that whenever students' learning is quantified and compared, a certain degree of unreliability will occur

(Kortez, 2008). Whether this is due to the characteristics of the test, the process of marking, grading or extraneous events such as the student being tired (Newton, 2013). It is also inevitable that any measurement error in the initial measurement will be converted into construct irrelevant variance in students' value-added calculations (Gorard, 2010a). This section extends previous discussions by introducing four additional issues that are particularly relevant or specific to value-added measures.

Comparability of annual measures

One factor that can impact upon longitudinal analyses is the tendency for the specification of governmental value-added models to fluctuate, both in terms of new version of the measures being developed (VA, CVA, Best 8, Progress 8) and more subtle revisions such as the list of Ofqual approved subjects. What is more, the curricula for each subject area will evolve over time (Williams, 2001). All of the aforementioned variation leads to construct irrelevant variance when students' value-added score are compared over time.

Ceiling and floor effects

The scoring system of the examinations is also a concern. During the calculation of Progress 8 the highest Key Stage 2 grade that can be achieved is a Level 5 (technically grouped fine-grade level 5.8) (DfE, 2020). This means that any students that are operating above this level will receive a scale score that does not recognise their full ability. This is known as a ceiling effect (Perry, 2016a). The problem with this scenario is that at the end of Key Stage 4 these students may appear to have made more progress than they actually have. The existence of ceiling effects will therefore introduce a non-random source of bias that will advantage schools with high-achieving intakes. In fact, if Kelly and Downey's (2010) estimate that roughly one third of students achieve a level 5 is still accurate, the effect may be significant.

Similarly, there is a lowest attainable point score in these examinations. This is occurs when student are working below fine-grade level 1.5 (DfE, 2020). Theoretically the attainment of these student will also be reported inaccurately (slightly under or over the reported level – because any scores under 2.0 are rounded to 1.5), and further bias will be added to the model. It is presumed however that a far smaller percentage of pupils will fall into this category.

It is possible to control for these floor and ceiling effects statistically as the post-2005 CVA model did (Kelly and Downey, 2010), however, Progress 8 contains no such adjustments.

The same types of distortion can occur at the higher and lower end of Key Stage 4 measurements. However, it is less common for students to receive the top or bottom grade at this level. .

Missing data

In theory, school-level value-added scores are created by aggregating the individual ratings of all eligible students within a school. Each of these is in turn based upon data, which will include at least the students' prior and current attainment. In practice some of this information is likely to be missing. Schools' figures are therefore based upon the records that are available. Which is problematic because each omission introduces inaccuracies into the analysis that may impact the school's rating (Gorard, 2012a).

To help identify the source and importance of these errors, the remainder of this section discusses the National Pupil Database (NPD). This resource was used in the calculation of Contextualised Value Added, the previous method of evaluating secondary school effectiveness in England and is now used to produce Progress 8 (it is also a key part of the data used for this thesis). The NPD is a Department for Education maintained resource that collects individual-level data on every school student that receives a state-education in England. This includes details of the student's examination entries and attainment, as well as some information about their background. The system is updated in real time each year as further in information becomes available and all data goes through a rigorous process of confirmation and amendment. It is therefore considered to be one of the most comprehensive sources of educational information available to researchers. It is an invaluable tool that is regularly used by academics and practitioners alike.

Even in this exemplar the records are not complete. First, there are cases that are omitted by design. As stated above, this resource contains information about pupils that attend state-funded schools in England. It does not, however, generally provide information on individuals that attend private education or those that are home schooled. Roughly 7% of the individuals that are educated in England are excluded from the database (Siddiqui, *et al.*, 2018). It follows therefore that Progress 8 ratings can offer no insight about the relative effectiveness of their education. Furthermore, even though the resource can theoretically provide key information on all students that have attended state-education, this is not normally the case. Progress 8 scores rely upon two pieces of information; students' prior and final attainment scores. However, in some years nearly 10% of each national cohort does not have both of these figures recorded in the National Pupil Database (Gorard, 2010a). A classic example of this is when students move from private to state education. As independent schools are not required to enter their students for the same examinations as state-funded schools, many of the students that transfer into state education will not have Key Stage 2 attainment data. Similar effects can also occur when students move into the English education system from elsewhere or are absent on key examination dates. The omission of this data will often prevent the individual from being included in the value-added model and thus creates an additional source of inaccuracy.

More recent evidence, however, suggests the prevalence of missing attainment data is sometimes very limited. Perry (2016b), for example, reported that there were no students within the 2013 KS4 pupil-level attainment data that did not have the requisite prior-attainment scores. Though, several parts of this statement should be clarified. First, this figure would not include students that were omitted by design. Moreover, as this figure was derived directly from the KS4 dataset it would also omit any student that attended a state-funded school for a period but egressed before their KS4 examinations. In such instances schools' would therefore receive no recognition for the time and resources they have invested. Finally, the figure will have under-reported the extent of missing data because when the NPD matches students' prior- (KS2) and current-attainment (KS4) figures, for the purposes of the national value-added assessments, they are flexible with regards to the tests that can make up students' Key Stage 2 Average Point Scores. If student have not taken the requisite KS2 maths and reading examinations then their score in one of these qualifications is used. If they have not taken either then the scores from their teacher assessments are accepted (DfE, 2020). Evidence on the reliability of teacher-assessments, however, is lacking (Johnson, 2013) and there is reason to suspect that they can be less dependable in some circumstances (Harlen, 2005). It is also likely that teacher-assessed national curriculum levels will not correspond precisely to the grades students' would have achieved in written examinations, though moderation procedures can help improve their alignment (DfE, 2011). If one assumes that the standard assessment protocols are the most accurate, then it follows that each of these mitigations will have introduced an additional error component into the results.

Data collection and coding errors

Once the prerequisite data has been collected it will need to be coded, transcribed and stored in a database. Each stage of this process has the potential to introduce further error. Coding refers to the process of assigning a numerical or categorical value to each data-item to facilitate its analysis. Practitioners often use validation protocols to help reduce the occurrence of these mistakes. The forms that the DfE use to collect data on students' performance and personal characteristics, for example, are programmed to draw the respondents' attention to potential mistakes (for example figures that don't tally correctly). The department also release several versions of schools' value-added results each year allowing schools the opportunity to challenge potential inaccuracies. When one considers the complexity of the information that institutions must provide, however, it is unrealistic to expect that these processes remove all inaccuracies. The recoded values then need to be entered into a database. Here a degree of human error can occur, usually because the individual matches data items or cases incorrectly.

Like all value-added models Progress 8 is more susceptible to these errors than assessment of raw-attainment scores because of the amount of data that it relies upon. To perform the calculation, prior and current attainment data are required for each pupil. These attainment figures will be susceptible to the measurement, coding, transcription and storage errors discussed above. Any students without these scores will then have to be excluded from the analysis, which introduces further error. It must therefore be assumed that all Progress 8 scores will be influenced by missing and erroneous data. This will be the case even if the data for a specified student is complete, as the calculation is based upon the average performance of similar pupils nationally.

4.3.3. The difficulty of distinguishing between school effects and extraneous influences

The previous sub-section highlighted that a fair measure of school performance must control for any external influences upon student attainment, particularly differences in school intakes which have the potential to introduce considerable bias into the analysis (Gorard and Smith, 2004). The process is not, however, a purely technical matter (Creemers *et al.*, 2010; Goldstein, 1997; Sammons, 1996). The discussions thus far have spoken of school and non-school factors as if it were easy to differentiate between the two. One could therefore be forgiven for thinking that the creator of an effectiveness model has merely to select a strong pedagogical theory and use it as a road map to identify the most important sources of bias. These could then be operationalised as well as any practical constraints permitted. The matter is not that simple as value-added evidence does not provide an effective means of making this distinction (Coe and Fitz-Gibbon, 1998; Visscher, 2001). To demonstrate this, the chapter looks at preceding versions of the English secondary school value-added measures and their success or otherwise in creating a measure that was independent of prior-attainment. This provides a contextually relevant example first because prior-attainment is the most influential background factor (Teddlie and Reynolds, 2000), and secondly, because it is the only extraneous factor that Progress 8 attempts to control for. The same conceptual problem, however, applies to the specification of any extraneous influences.

i. Creating measures that are independent of raw-attainment

One of the first studies to critique the 2004 DfE (then DCES) value-added measure was Gorard (2006b). Within his sample of 124 schools from 4 Yorkshire LEAs, Gorard found a near perfect correlation between schools' total Key Stage 4 attainment scores and their value-added performance ratings. More specifically, he observed that the correlation between schools' VA scores and total KS4 attainment was 0.96, and the association between VA and the threshold measure 'proportion of pupils with five or more GCSE passes at grade A*-C' was 0.84. Whilst some association would be anticipated between these variables because, other things being equal, the schools that help students to make the most progress would also tend to have high KS4 raw-attainment. In this instance the pattern of results was too perfect with no schools bucking the trend and performing substantially better or worse than expected. Gorard therefore reached the conclusion that the two measures were actually measuring the same thing and that the model provided little information that the raw-scores did not. Despite the ostensibly logical meaning of value-added scores, the transformation of results into relative learning gains served only to obscure what was actually being reported. What is more concerning however is that, as decade of Educational Effectiveness Research can attest, raw attainment scores provide an invalid and vastly unfair method of assessing of schools' impact as the attainment data is heavily influenced by the differences in school intakes. The strength of the aforementioned association therefore implies that the value-added ratings were heavily biased by students' prior-attainment level, the very thing that they were designed to control for. In terms of the aforementioned continuum, the original DfE value-added model was only successful in moving a short way, if at all, toward a true measure of school effectiveness[6]. What is interesting for the purpose of our discussion, is that subsequent analyses by Perry (2019) concluded that whilst there was indeed a substantial school-level correlation between value-added scores from 2004 and the Key Stage 2 prior-attainment averages (r = 0.50), the association was not visible at the pupil-level (r = 0). The implications of this will be discussed shortly (see section on the 'regression attenuation' effect).

Kelly and Downey (2010) conducted a similar analysis with the 2005 CVA pilot data to test whether the addition of student background and contextual factors in the Contextualised Value Added measure had any impact upon the relationship. Their headline finding was that, within their sample of 370 schools, only 14% of the variation in CVA could be explained by the raw-attainment threshold measure (% 5 A*-C grades). Whilst this in itself implied that new controls were successful in removing additional bias from the measure, a crucial aspect of the design was that they also assessed association between the official value-added indicator (as used within Gorard, 2006) and the percentage of student achieving 5 A*-C. As this relationship accounted for 59% of the variance in VA ratings (a lower but comparable figure to that found in Gorard's study), the pair were therefore more justified in concluding that the new measures were responsible for the increased disparity between the contextualised value-added and raw-attainment data.

What is more, when Contextualised Value-Added was discontinued in 2010 and replaced with Best 8 value-added (that did not control for contextual factors), the school-level correlations between both value-added scores and KS4 attainment, and the association between the value-added and KS2 prior-attainment score increased (Perry, 2019). In fact, they returned to almost to their original level.

[6] Two further points are worthy of notation. Firstly, in later publications Gorard confirmed that a similar association existed within primary education (see Gorard, 2008a). Politicians, however, initially tried to downplay this association by claiming that it was a freak event specific to certain geographical areas (HC Deb, 2005). This lends support to the argument that neither the proponents nor critics of value-added truly know what makes up schools' residual scores.

Although further data on the association between Value-Added and raw-attainment during this period would be necessary to draw firm conclusions, the associations observed within these three studies, combined with existing evidence that high un-contextualised value-added ratings are more common in school with high-achieving intakes (see, Perry, 2016a) and that similar magnitudes of association can be explained by differences in the school-level aggregates of students' prior attainment (see discussion of Perry 2019 below), suggests that school-level value-added scores have an association with the composition of school intakes that cannot be explained by the individual-level relationship between prior-attainment and KS4 performance. This is known as a compositional effect, the nature of which is discussed below.

ii. Compositional effects

A compositional effect occurs when the aggregate of an individual-level characteristic has an independent effect upon student outcomes. For example, if being educated alongside motivated and high-achieving students has a demonstrable impact upon a student's performance that cannot be accounted for by their own characteristics, this would constitute a compositional effect. The unique influence of these peer effects are most commonly evaluated using multi-level regression models (Gorard, 2006a). In these, researchers examine the variation in student outputs that can be explained by the schools' composition after any individual-level differences have been accounted for.

Ostensibly the choice of whether to control for compositional variables within a model is straight forward (Raudenbush and Willms, 1995). If the measure will be used to inform parental choice, then any peer effects should be included within the school's effect. This is because parents will not care what aspect of the school helps their child make more progress - only that it does. Whereas, if one intends to evaluate schools' performance it would be unfair to reward/punish schools for factors that are outside of their control. Such applications therefore require any compositional effects to be removed.

The problem however is that whilst attention has been paid to these effects within academic research (see for example, Marks 2015; Timmermans and Thomas, 2014; Boonen *et al.*, 2014) and the multitude of reasons for suspecting that such effects may exist (see, Gorard, 2006a; Harker and Tymms, 2004; Willms, 1992) the overall evidence so far has been inconclusive. For every study that finds evidence of such effects, it seems there is another that does not (Nash, 2003) and the magnitude of the quoted effects has varied widely (Teddlie and Reynolds, 2000). Wilms and Raudenbush (1989), for example, found that within their sample of Scottish secondary schools every 1 unit increase in there standard deviation of the school-level SES mean was associated with an increases equivalent to 29% of a standard deviation in attainment. On the other hand, Boonen *et al.* (2014), Lavy *et al.*, (2012) and Marks (2015) did not observe any appreciable effects[7].

Compositional effects therefore remain a controversial topic (Reynolds *et al.*, 2014 pp.209). Gorard (2006a, pp 87, ln 19-20) summarised this situation by saying that compositional effects are "hard to pin down precisely because [they] are small relative to the amount of noise in the system". Due to the nature of their influence any compositional effects will always be marginal in relation to the impact of

[7] In addition to those listed in the main text, examples of studies that detected compositional effects include Brookover *et al.* (1978; 1979), Henderson *et al.*, (1978), McDill *et al.*, (1969), Rutter *el al.*, (1979), Shavit and Williams, (1985), Summers and Wolfe, (1977), Willms, (1985, 1986). While the following studies failed to find compositional effects; Alexander and Eckland (1975), Alwin and Otto (1977), Bondi (1991), Hauser (1971), Hauser *et al.*, (1976), Mortimore *et al.* (1988).

individual-level factors (Perry, 2019). What is more, the relationships observed in multiple-regression analyses only provide correlational evidence. They do not demonstrate that the independent variable is causally responsible for the differences in student outputs. Given that all measures contain measurement errors and/or missing data, it therefore becomes exceptionally difficult to make verifiable statements about the effect of school composition as one cannot distinguish causal influence from error. Researchers may then end up interpreting the data in a manner that is consistent with their personal expectations (Gorard, 2006a). This is of course the same process that is used to detect school effects and for that matter the same criticism that Gorard levels at value-added ratings in general. The problem is magnified in this instance, however, because the characteristics of a school's intake are just one aspect of the school effect and by definition less substantial. This has led some authors to conclude that "group composition matters little" (Gibbons and Telhaj, 2012, p. 26, ln. 16) or that compositional variables have a "trivial" effect that does not justify the adaptation of policy (Marks, 2015a, p. 139, ln. 16). Others, such as Harker and Tymms (2004) assert that the assessment of compositional effects merely requires a more subtle approach than researchers initially supposed. Willms (1985) takes a similar stance to this. He argues that there are circumstances that make the detection of genuine effects more likely, which include having a well specified mathematical model and output measures that are specific to the taught curriculum. Shavit and Williams (1985) cite the more generic criterion of establishing sufficient variation within ones sample.

Perry's (2019) finding that there was evidence of systematic biases in former Best 8 value-added ratings (2011-2015 system) refers to a branch of the compositional literature termed 'phantom compositional effects'. Whilst the field has yet to reach a unanimous conclusion on the importance of compositional effects, what has become evident is that some of the claimed peer influences are in fact artificial artefacts of the statistical procedures that were used to uncover them (Nash, 2003). As discussed above, compositional effects are most commonly found in multi-level models of educational effectiveness. Since compositional effects are defined as the influence that the collective characteristics of school intakes have upon pupil performance over and above the impact of students' personal characteristics, their effect is operationalised as the proportion of the variance in student outcomes that can be attributed to the aggregated measure of the characteristic(s) once individual differences between pupils have been accounted for. The problem with this arrangement, however, is that whilst the presence of measurement error in the student-level data will lower the explanatory power of the individual-level variables, their impact upon school-level aggregates is far less substantial, owing to the fact that any randomly distributed inaccuracies will tend to balance each other out when there are sufficient observations (this is not to say that all measurement error will be random, merely that a portion of the inaccuracies will be). The school-level compositional variables are therefore in an ideal position to mop up any of the variance that should have been accounted for at the student-level (Harker and Tymms, 2004). Thus, the purported effects are likely to be inflated whenever the modelling of any student-level background factors is sub-optimal. Hutchinson (2007) presents mathematical proof that these deficits can lead to the substantial school-level deviations even when there are no omitted variables and when sample sizes approach infinity. This latter part is important because it prevents researchers from quantifying the inaccuracy using statistical methods such as confidence intervals, or corrections such as Bayesian shrinkage (Perry, 2019).

iii. Regression attenuation and the grammar school effect

Measures of prior-attainment play a key role in value-added calculations. They are used to discriminate between pupils and make predictions about the future performance. Errors in student attainment record can therefore have a substantial effect upon their schools' apparent progress ratings.

Whenever the prior-attainment of a student is reported one of three situations arises. The assessment will have been accurate, in which case the measurement reflects the student's ability at the time (no error), the student may have been an 'under-achiever', meaning that their prior-attainment score is lower than their actual ability level (negative error component), or the student may have been an 'over-achiever' and performed better than their level of mastery would warrant (positive error component). If these errors occur at random, as the defenders of value-added models propose, it is tempting to assume that there would be an equal number of over-and under-performing pupils at any given point on the prior-attainment scale. This, however, is unlikely to be the case (Perry, 2019). As the distribution of students' underlying ability is likely to resemble a normal or 'Gaussian' distribution curve, with most students having an initial ability level close to the modal rating, there will be more pupils with high prior-attainment that over-achieve than under-achieve, and likewise, more pupils with low-prior attainment that under-achieve than over-achieve.

During value-added calculations the presence of random and systematic measurement errors therefore dilutes the differences between the measured prior-attainment groups because the true competency of students may be closer to mean than indicated. This ultimately suggests that average final attainment of each prior-attainment group will be closer to the overall average attainment level than they should have been. This is known within school effectiveness research as the attenuation bias (Frost and Thompson, 2000; van Ewijk and Sleegers, 2010). The process is problematic because it inflates the progress rating associated with any final attainment scores that deviate from the national average. In other words, if students' raw-attainment levels are above normal they will be disproportionally rewarded and if they are below average they will be unjustifiably punished in the analysis. What is more, the greater the measurement error the more extensive the effect. To put this another way, in models such as Progress 8, where the influence of school composition is not taken into account, the scores will not be as independent of student raw scores as proponents claim. Even if they are random, the presence of measurement error introduces systematic bias into the analysis that will advantage particular types of pupils and schools (Perry, 2019). Another way of stating this is that whenever one models a normally-distributed student-level background factor such as prior attainment, imperfect operationalisation of the underlying concept will result in a measure that only partly adjusts for the factor, rather than unbiased but 'noisy' expectations. Furthermore, the effect will impact all students and schools' ratings regardless of whether their prior-attainment scores contain errors. This is because the errors in other students' ratings will have influenced the performance of the prior-attainment groups to which they are compared.

The magnitude of the problem will depend on the specific measures and the context in which they are used (Pokropek 2015). Therefore whilst several studies have reported the impact of attenuation bias (e.g. Dieterle *et al.*, 2015; McCaffrey *et al.*, 2015; Televantou, 2015), Perry's (2019) study of the 'Best 8' value-added measures is the most applicable the research in this thesis as its specification is comparable to Progress 8. This paper found that small, medium and large errors in students' KS2 prior-attainment results[8] translated into substantial errors in schools' KS2-KS4 value-added ratings. Within the small-error dataset, for example, the average standard deviation of scores from the 'true' performance of pupils was 2.4 Best 8 points. The medium-error data resulted in a standard deviation of 7.1 Best 8

[8] The quantitative values attached to these ordinal labels were selected so as to mirror the levels of test-retest reliability observed in various sources of prior-attainment attainment data. The scale range from the standard Key Stage 2 achievement tests (Opposs and He, 2011) which had reliability ratings ranging from 0.81 to 0.85, to the Primate Indicators in Primary Schools data (PIPS) which were used by Harker and Tymms (2004). The latter tests were developed specifically for the prediction of later performance and had a more substantial test-retest reliability of 0.81 to 0.96. Thus within Perry's research the large error dataset was defined as having a correlation of 0.79 with students 'true' KS2 ability, whilst the medium error dataset had a correlation of 0.89 and the small error dataset a correlation of 0.97.

points and the large-error data resulted in a standard deviation of 11.2 Best 8 points. To put these results into context, Best 8 value-added scores typically range from around -90.7 to 68.2 (Perry, 2019). Or to put it another way, since 6 Best 8 value-added points equated to 1 GCSE grade (within a single subject), these results imply that on average schools' ratings will be biased by roughly 0.4 GCSE grades, 1.2 GCSE grades or 1.9 GCSE grades per pupil, if low, medium or large errors are present in the prior-attainment data[9]. An error of 2.4 points within a school's rating is therefore relatively small, whilst deviations of 7.1 and 11.2 points represent far more substantial quantities of bias that would have pragmatic implications for many institutions.

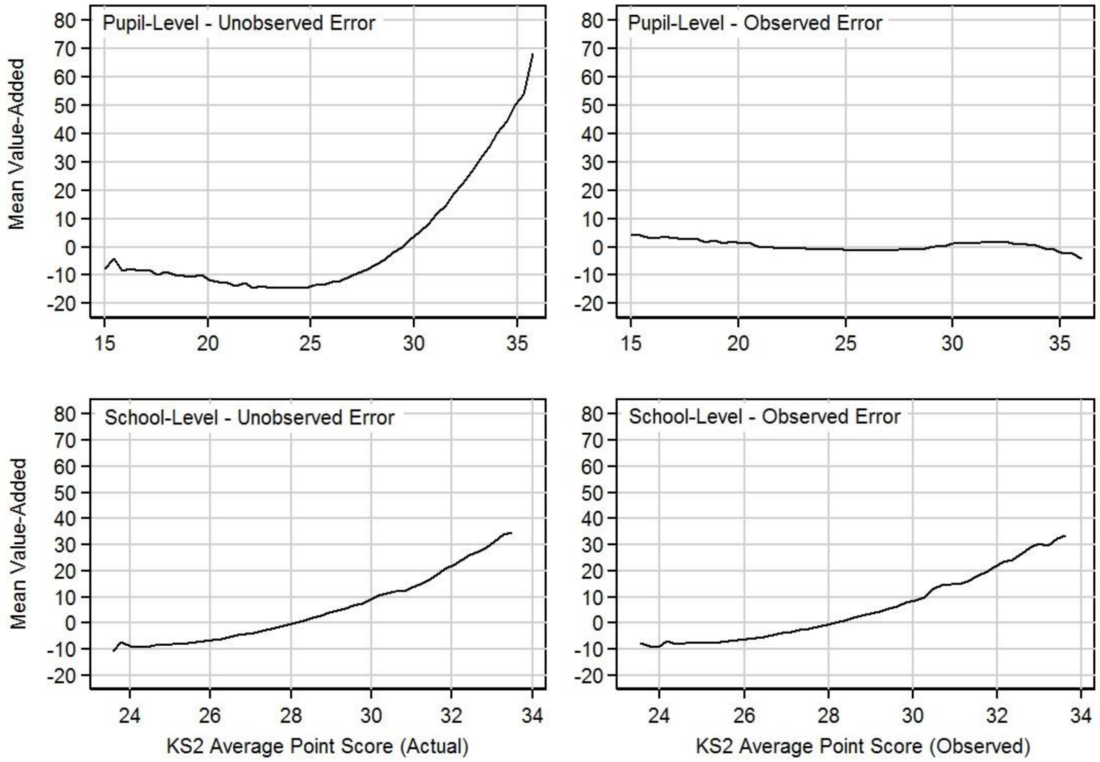
Given that, in England, value-added models are used to hold individual schools to account, it would perhaps be more appropriate to report that deviations between -5.3 and 15.0 points were observed in the small-error data, deviations between -13.9 and 36.3 points were observed in the medium-error data, and deviations of -20.8 to 56.7 points occurred in the large-error data. Since the current KS2 attainment measures are expected to contain reasonable portion of error (see Section 4.3.3.), this implies that in some school's KS4 value-added scores could be as far as 9.5 GCSE grade levels away from their true performance on account of attenuation bias alone (though it may be easier to imagine this as 1.2 GCSE grades per subject area) and even if highly reliable assessments were used individual institutions would still be unfairly rewarded or punished.

Perry's (2019) second finding was that the distribution of error was skewed to the left, meaning that there was a right tail of schools that were disproportionally advantaged by the attenuation bias. Shockingly, by excluding schools with an average prior-attainment scores in excess of 31 (national curriculum level 5C at KS2) Perry was able to demonstrate that this group was comprised almost entirely of grammar schools. This is a meaningful finding given current debates about the merits of grammar schools (e.g. Morris and Perry, 2016; Sutton Trust *et al.*, 2008), specifically, the claim that grammar schools' Best 8 ratings were 25 point higher on average than other schools (Morris and Perry, 2016). These results however suggest that small, medium and large measurement error in students' prior attainment data would have inflated schools' ratings by an average of 7.4 point, 22.8 points and 35.1 point respectively. This suggests that the compositional effects and grammar school effects reported in English 2004-2016 value-added data were largely or entirely spurious.

The important thing to realise about these errors is that they are only visible at the school-level and would most likely be misinterpreted as compositional effects. Figure 4.3.3a (below) depicts the errors that were present at pupil- and school-level within Perry's analysis. Only the graphs on the right, however, would be observable.

[9] This estimate ignores the fact that students KS4 scores in maths and English were double weighted.

Figure 4.3.3a: Pupil- and school-level errors in the medium-error dataset



(* graphs sourced from Perry, 2019, p 258)

Here it can be see that the mean KS2 prior-attainment score during the study was approximately 28, and that the further from this point the pupil/school mean score is the greater their illegitimate advantage or disadvantage. What is particularly interesting from the perspective of compositional effects (and the grammar school effect) is that at school-level the individual errors cancel each other out to reveal the underlying trend within the data. The bias is however masked at pupil-level.

Fortunately, Perry states that the vast majority of school-level bias can be nullified by controlling for differences in prior attainment at the school-level (80%, 89%, 90% of the error in his small-, medium- and large-error datasets were corrected by such measures). Thus, this simple step can lessen but not remove the problem. School-level attainment controls have however been absent from English schools value added measures since 2010, which in his opinion means that Progress 8 will be as susceptible to attenuation errors as the Best 8 model used in Perry's analysis..

It is important to be aware however that in practice controlling for attenuation bias and excluding compositional effects amount to the same thing. Both describe the school-level variance in prior-attainment left over after the Key Stage 2 scores have been accounted for at the individual-level.

Returning to the problem of model specification, one can therefore see that whilst it is easy to state that the decision to include or exclude compositional factors from the school effect is determined by what one wishes to do with the information, in practice there is no way of separating genuine and spurious effects. One therefore has to make a judgement on which is more likely to be the case.

The problem however is worse than this. If one decides to treat, for example, the grammar school effect as real, what mechanism should this be attribute to? Several relationships could explain such an effect; the 'peer effect', the level of support available from parents, the commonality of disciplinary problems, the school learning environment and schools' ability to attract experienced teachers (Gorard, 2006a). It could even be that grammar schools are not more effective overall but differentially effective with particular types of pupil (Foley and Goldstein, 2012). And what if more than one of these influences is responsible? This would necessitate either the use of multiple controls, each of which will overlap with genuine school effects, or making further assumptions about which factors has the greatest impact (Coe and Fitz-Gibbon, 1998; Visscher, 2001).

In summary, this section has demonstrated that controlling for extraneous sources of bias is not merely a technical matter (Creemers *et al.*, 2010). The use of academic learning theories is therefore essential in making operational decisions, though it is not something that can resolve the matter as the value-added methodology does not offer a clear way of distinguishing between school-related and non-school factors (Creemers *et al.*, 2010). One cannot, as demonstrated, even be assured that the official DfE value-added measure will be independent of students' current and prior attainment, despite this being central to the calculation's legitimacy.

5. Indirect Threats to Validity

5.1. Chapter Introduction

In the context of value-added models, school effects are used to identify differentially effective schools, schools which enable their students to make more or less progress than they would have made had they attended another institution. For the construct to be meaningful however it must have certain properties. It must have both duration and scope (Scheerens. 1993). That is to say, that the school must have a comparable influence upon most students and its effect should be relatively stable over time. If this is not the case then there is little value in providing policy makers, practitioners or parents with school-level value-added data as there would be no meaningful way of acting upon the information. Progress 8 scores would not, for example, provide a reliable method of selecting the best secondary school for ones child if school effectiveness was highly volatile. Under these circumstances, the amount of progress the last Year 11 cohort made might have little association with the performance of students that will not sit their GCSE examinations for 6 years (Leckie and Goldstein, 2009). Similarly, if schools have a strong differential effect upon students with different characteristics and/or backgrounds, school-level summaries would not help in selecting the best school for a specific child (Allen and Burgess, 2013). A multitude of studies have therefore investigated whether value-added results exhibit these characteristics (see, for example Sammons 1996; Goldstein, 1997; Coe and Fitz-Gibbon, 1998; Teddlie and Reynolds, 2000; Visscher, 2001; Marsh *et al.*, 2011). This section reviews this material.

Before commencing the discussion it should be acknowledged that reporting upon the characteristics of school effects without knowing the validity of one's research instrument is problematic. Especially when one considers that value-added models can only provide estimates of school effectiveness (Fitz-Gibbon, 1997) and that all scores therefore contain any genuine school effect as well as random and systematic error components (Gorard, 2010a). There has therefore been much debate as to whether the volatility in schools' ratings reflects the properties of the school effects, problems with the specification of specific models or problems with underlying methodology. In fact, even associations of equal magnitude have been characterised in disparate ways (Perry, 2016b). The problem of interpretation, however, is discussed in the next chapter. For now, this thesis will focus upon reporting the stability and consistency of value-added ratings, and the implications for DfE practices. Evidence regarding the source of any volatility is therefore noted but not discussed in detail.

It is also worth noting that measures do not have universal validity. The legitimacy of evaluations therefore needs to be considered in relation to specific tasks (Messick, 1995). Since Progress 8 is used in the context of a high-stakes accountability system it is insufficient for the distribution of results to be vaguely in-line with expectations or for the results to be somewhat stable. The value-added results of individual schools must be dependable and provide a reasonable representation of school's impact upon all sub-groups of students. Furthermore, since the DfE have actively encouraged parents to consider these ratings when selecting their child's school the ratings must remain stable for a prolonged period. The fulfilment of these requirements is discussed after each section of the review is concluded.

5.2 The Stability of Value-Added Ratings

Over time schools' value-added scores will vary. The extent of this instability, however, has been characterised in different ways. Researchers have described the year-to-year stability of value-added ratings as "impressive given that the great majority of EER research usually takes place within unstable communities and rapidly changing school environments" (Reynolds, *et al.*, 2012, pp. 11-12, ln 44-1), as showing "considerable stability" in consecutive year's ratings but "much more variable" over longer periods (Thomas *et al.* 2007 p. 277. ln. 7 and 8), as "not particularly reliable or stable" (Marsh *et al.*, 2011, p. 286, ln. 6), as demonstrating "a complete lack of stability" that ensures the estimates "tell you essentially nothing" (Linn and Haug, 2002 p.33, ln. 71 and 74), or as being so inconsistent that they "simply do not have much confidence that educational agencies can identify value-added at the school-level" (Kelly and Monczunski, 2007, p. 279, ln. 60-62).

Before attempting to quantify this variation or assess its implications it is important to acknowledge that the context of effectiveness studies and the quality of the underlying datasets have varied (Dumay, *et al.*, 2013). These factors must be considered when one interprets the literature. Most researchers agree, for example, that primary school ratings are less stable than secondary school ratings because the calculations are supported by smaller samples of pupils (see, for example, Strand, 2016). Likewise, composite measures tend to be more consistent than subject specific measures (Teddlie and Reynolds, 2000) and missing or inaccurately reported data will impact upon schools' results (Gorard, 2010a; Jesson and Gray, 1991).

Different methodological approaches have also been used to examine school effects. Whilst modern research stresses the need for longitudinal assessments that control for the differences in school intakes by design, many designs are cross-sectional (Teddlie and Reynolds, 2000). That is to say, that their appraisal of schools' performance is based upon one cohort of students and one years' performance data. A cross-sectional evaluation of KS2-KS4 value-added, for example, would be based upon the learning gains exhibited by the most recent Year 11 cohort. The fundamental problem with this approach is that when one compares schools' ratings over time, one is comparing the performance of different groups of students. Any differences between the cohorts that cannot be removed via statistical controls will therefore bias and potentially destabilise the measure (Fitz-Gibbon, 1997; Gorard, *et al.* 2013; Hill and Rowe, 1996; Stringfield, 1994a; Thomas *et al.*, 1997b). Cross-section assessments therefore tend to produce more volatile estimates than longitudinal designs[10].

For a time the most popular solution to this problem was for researchers to consider several years of performance data simultaneously. Gray *et al.* (1993), for example, recommended that ratings should be based upon at least three years of performance data. This practice allowed the stable component of school effects (i.e. the mean influence of schools across the years of the study) to be distinguished from the unstable component (i.e. the year-to-year variation from this figure) (see Willms and Raudenbush, 1989 for an explanation of how this is achieved). In theory, this not only produces more accurate measures of school effectiveness but means that the stability of the schools' ratings is interpreted in a different way. That is, based on the magnitude of schools' stable components in relation to the unstable components. From this perspective the volatility of schools' performance ratings does not appear as extreme as it once did. This approach, however, will disadvantage improving schools, as poor ratings will count against a school for a prolonged period. Which may explain why, despite early recommendations (see Fitz-Gibbon, 1997), the practice has never been embraced in DfE policies.

[10] Though it should be noted that if the differences between school intakes endure over time, inadequate controls can sometimes have the opposite effect (Perry, 2019). See Section 5.2.1.

The use of more sophisticated designs is therefore becoming increasingly common (Kelly and Monczunski, 2007). Growth models, for example, model school achievement gains among the same group of students (Ballou *et al.*, 2004; Tekwe *et al.* 2004). This approach therefore has many advantages over cross-sectional assessments including the fact the differences between students are controlled by design rather than statistical intervention (Raudenbush, 2004; Rubin *et al.*, 2004). Theoretically, this should address the problem, however, new threats to validity are often evident including the design's sensitivity to student mobility and drop-out (Rubin *et al.*, 2004), and the potential confounding of achievement gains made during the school year and school holidays (Downney *et al.*, 2004; Entwisle *et al.* 1997). The models are also highly complex, difficult to understand and increase the demand for testing to such an extent that the current standardised testing system would be incapable of providing the requisite information, yet still show notability volatility within the annual results (Guldemond and Bosker, 2009). As a result some researches still recommend that performance data is averaged across a several years (Raudenbush, 2004).

Progress 8 uses a traditional cross-sectional design that only considers the performance of a single cohort of students. The intuitive expectation is therefore that its stability will to be towards the more unstable end of the spectrum. That being said, the effects may be depressed slightly by other characteristics of the model. These characteristics include the fact that Progress 8 is a composite measure that evaluates secondary school performance, the high quality the NDP datasets and its simplistic approach to controlling student-level differences. Each of these features is discussed in due course.

5.2.1. The stability of the preceding DfE models

Since there is evidence that the volatility of value-added scores is highly dependent upon model applications and the quality of any underlying datasets, the remainder of this discussion will focus on studies that have evaluated the stability of value-added measures in the context of secondary-level state education in England. The findings either relate directly to Progress 8 or to comparable measures that have been used in the same environment.

Contextualised value-added

Contextualised Value Added (CVA) was a DfE measure that was used to evaluate secondary school performance between 2005 and 2010. Its specification was comparable to Progress 8 in that it evaluated students' learning between the end of KS2 and the end of KS4, after differences in prior-attainment had been taken into account. Several characteristics, however, distinguish it from the more recent DfE measures (see Chapter 3 for further details). Whilst most of these are inconsequential to the current discussion, the fact that it controlled for a multitude of contextual variables including students' background and personal characteristics, is highly relevant.

Two studies which evaluated the stability of CVA ratings were Gorard et al., (2013) and Leckie and Goldstein (2009):

Gorard *et al.* (2013) evaluated the stability of schools' Contextualised Value Added scores between 2006 and 2010. Correlations of 0.58-0.79, 0.48-0.67, 0.56 and 0.46 were reported for results 1, 2, 3 and 4 years apart respectively. Based on this he concluded that the results of CVA were potentially "meaningless" (pp. 1, ln. 14), as even results one year apart had a modest association with one another, and after 4 years just 21% of the variance was common between the results. Whatever value-added is measuring, Gorard explains, it is not a stable characteristic of schools. This makes CVA results useless as an informant of parental choice, as schools' current rating may bear little to no resemblance to their future performance. In fact, his analysis found that no school with high-quality data records managed to receive a good rating for 5 consecutive years. Gorard therefore asserts that much of the apparent volatility may in fact have been due to unobserved errors within the data, though he is explicit in stating that his research methodology cannot prove that this was the case. It may therefore be that school effectiveness is so dynamic that schools' ratings should change dramatically year-to-year. Though, even if this were to be the case, the observed levels of instability would have been sufficient to limit the defensible applications of the measure.

Leckie and Goldstein (2009) reported similar levels of association. Specifically they estimated that value-added ratings 1, 2, 3, 4, and 5 years apart had correlations of 0.80, 0.73, 0.57, 0.46 and 0.40 respectively. The results and their ultimate conclusion, that the ratings are so volatile and uncertain that they have very little to offer as an informant of school choice, are therefore consistent with Gorard's summation. All schools in the study, however, were from the same local authority.

Allen and Burgess (2013) also assessed this issue but from a slightly unorthodox perspective. Specifically, the pair developed a framework to test whether the consultation of CVA data would help parents to select a better secondary school for their child. The results were surprising. Whilst their overall findings suggest that basing school choice upon the kinds of data that is featured in school performance tables is likely to enhance a students' GCSE performance, Contextual Value Added ratings were only more predictive of future GCSE scores when the assessments were carried out 1 year apart. Over 6 years, the low level of stability in the value-added measure meant that raw-attainment measures were the more efficient predictors. This suggests that if students transition to secondary school at the traditional point, the parents would be better of selecting their child's school using unadjusted figures. In fact, even the average KS2 attainment of the school cohorts, which initially had a low predictive capacity, help to make more favourable decisions in the long term because its prophetic ability was very stable. The conclusion then is clear. Whether this variation is indicative of genuine changes in effectiveness, or a problem with the calculation, the volatility of CVA was sufficient to limit is value. The short-term association between schools' CVA ratings and students' predicted raw-attainment levels, however, suggests that the measure did provide a degree of insight.

Best 8 value added

Best 8 value-added was used in England between 2011 and 2015. In many ways its specifications can be seen as a halfway point between the CVA model which it replaced and Progress 8 which it gave way too. Like the latter, it did not take into account any contextual variables that might impact upon school performance, for example, differences in students' background, personal characteristics or the composition school cohorts. Since this was the predominant distinction between the CVA and Progress 8, the Best 8 model should therefore be more representative of the level of stability that we should expect to see in current assessments. It does however model the effect of prior attainment using the Ordinary Least Squares method, as opposed to computing students' expected performance based upon the mean KS4 attainment of each prior-attainment fine-grade group, and applied a shrinkage factor to the school-level results. Both factors were synonymous with the specification of CVA.

In order to evaluate the stability of Best 8 value-added assessments, Perry (2016a) replicated Gorard *et al*'s (2013) study using schools' raw-attainment scores from 2011-2014, the updated value-added specifications and population data for all state-schools in England. Correlations of 0.70-0.96, 0.62-0.93 and 0.6 were recorded between the unadjusted average capped GCSE and equivalent point scores per school, 1, 2 and 3 years apart, respectively, whilst associations of 0.56-0.79, 0.49-0.68 and 0.44 occurred between the schools' value-added ratings.

Two main conclusions can be drawn from this. Firstly, schools' raw-attainment figures were considerably more stable over time than their value-added ratings. This difference is visible in the figures above but much more evident when one compares the stability of schools' value-added ratings with the average capped point scores for GCSE qualifications when non-GCSE qualifications have been excluded. The association between schools' raw-scores in GCSE qualifications ranged from 0.94-0.96, 0.92-0.94 and 0.90, respectively, for results separated by 1, 2 and 3 years (Perry, 2016b). Secondly, the association between the schools' value-added ratings 1 and 2 years apart was fractionally higher than those observed in Gorard *et al*'s (2013) study after the same lag times, suggesting that Best 8 value-added ratings were more stable than the preceding CVA measure. The only exception to this pattern was the association between the first (2011) and final (2014) ratings. Both the correlation between school's unadjusted average capped GCSE and equivalent point scores and value-added ratings dropped significantly at this point, presumably because of the GCSE reform that took place that year. These reforms reduced the number of qualifications which could be counted as GCSE-equivalent qualifications. It is therefore understandable that these changes would influence schools' value-added scores and their average GCSE and equivalent point scores, but have no discernible impact upon schools' average point scores when GCSE equivalent qualifications are excluded.

Rather that praising the increase in stability, however, Perry (2016a) cautions readers that the effect may be spurious. Since the major distinction between the Best 8 and CVA value-added models was the removal of most of the controls for differences in school intakes, he asserts that the lower level of volatility is most likely due to the increase in stable sources of bias.

To provide greater insight into the cause of this volatility, these correlations were presented alongside several subordinate analyses (see Perry 2016b, Analysis 3). These reported upon the distribution of deviations and the performance of particular sub-groups of schools:

Within a single year the performance of some schools deviated by more than 36 Progress 8 points. Since 6 'Best 8' points equated to approximately 1 GCSE grade, this represents a change of approximately 6 GCSE grades across the 'Best 8' subject areas[11]. Most school ratings, however, deviated by 0-24 point or 0-4 GCSE grades between consecutive assessments. Over two years the scale of the largest deviations did not increase substantially, though moderate and large changes were more frequent. These changes seem rather large, nevertheless it should be acknowledged that the deviations were normally distributed and that the largest changes were only evident in a small proportion of schools.

The second aside evaluated whether the stability of schools' ratings was influenced by their previous year's performance. As inadequate and outstanding schools find themselves in very different positions, it was theorised that institutions which performed poorly in their last assessment were more likely to make substantial changes to their practices. It would not therefore be surprising if the effectiveness of these schools was more volatile and changed significantly year-to-year. In contrast, having already achieved an impressive rating, there is considerable incentive for the best performing schools to be

[11] This calculation ignores the double weighting maths and English scores.

cautious and stick with proven approaches. Their performance was therefore assumed to remain more stable. Schools were thus ranked-ordered based on their previous value-added rating and split into quintiles (5 percentile groups of equal size). Pairwise correlations were then calculated between schools' 2011-2014 value-added ratings. The results show that the stability of schools' performance ratings was roughly even across the distribution of prior performance ratings with a slight tendency for higher-levels of instability in the middle range of scores. Whilst the presence of such a relationship would not have definitively proven that differences in schools' performance ratings were genuine, the near uniform stability levels that were reported suggest that the instability arises from the value-added calculation itself.

Finally, the minimum, maximum and mean change in school's annual ratings was reported, as well as the mean change across all school. That is to say, that the most, least and typical level of stability that existed in each schools' results during consecutive value-added evaluations. The results indicated that the mean minimum change was 4.7 points or roughly ¾ of a GCSE grade per pupil, the mean maximum change was 21.1 point or 3½ GCSE grades per pupil and the overall mean of each school's mean change was 12.1 points or 2 GCSE grades. Perry's interpretation of this volatility was in line with our own. Specifically he states that the magnitude of this grand mean "seems large as a typical change, but would not be surprising in a single case" (Perry, 2016b, p. 210, ln. 12-13).

Despite providing circumstantial evidence that the volatility of schools' year-to-year ratings was a little higher than many had imagined, the outcome of Perry's investigations was inconclusive. To use his own words, "the scores are not sufficiently volatile so as to rule out a meaningful school effect...[therefore] these secondary [school] figures do not obviously support either the critics or proponents of value-added" (Perry, 2016b, p. 210, ln. 12-13). The stability of schools' ratings was most likely inflated, however, by DfE's decision to remove the controls for contextual variables (see discussions of omitted variable bias).

These analyses make it clear that cross-sectional value-added models of secondary school performance, with comparable specifications to Progress 8, demonstrate a degree of stability but less than many researchers would desire. Certainly, the models are too unstable to predict school performance 6 years in advance. This makes them unhelpful as an informant of parents' educational decisions, at least if their child changes school at the traditional transition point. From other sections of our review, however, we understand that the reason for this instability it not understood. Further assessment is therefore necessary to investigate the factors which can best explain the change in schools' scores over time. This matter is addressed in the empirical sections of this thesis.

5.2.2. The stability of cohort's value-added ratings over time

One characteristic that connects the aforementioned DfE models is that they all judge schools' performance based on the progress that a single cohort of students makes from the end of KS2 to the end of KS4. When one compares the results over time, one is therefore comparing the development rates of different groups of students. A common criticism of this form of cross-sectional value-added design is therefore that an unknown proportion of the year-to-year variance will ultimately stem from unacknowledged differences between the cohorts, rather than changes in the schools' effectiveness (Guldemond and Bosker, 2009). Studies which evaluated the stability of cohort's scores over time would therefore provide a useful basis for evaluating the level of bias that said shortfall introduces. The theory being, that if the ratings of individual cohorts were to be highly stable over time, then we could be relatively assured that it is the differences between cohorts that destabilise the official DfE measure

(Perry, 2016b). Unfortunately, we have been able to locate only one study that performed such an assessment.

Perry (2016a; 2016b) evaluated the stability of CVA ratings for specified cohorts of students using as simplified replica of the model and data from the Making Good Progress study[12]. His results indicated that cohort ratings which were separated by 1 year had correlations of 0.43-0.69. Thus, even when the CVA specification was applied to the same cohort of students, the stability of ratings over a short period of time was only "moderate" (Perry 2016b, p. 193, ln. 1). Furthermore, a correlation of 0.62 was recorded between cohort results that were separated by 2 years. We hasten to add however that only one group of secondary cohorts was assessed for three consecutive years. This rating therefore represents a reduction, from the 0.69 biannual association score, to which it is attached. Thus, the correlation between cohorts ratings decreased over time as one would expect.

It is interesting to interpret this instability alongside the findings from the previous sections. If omitted variables such as intake differences were primarily responsible for the variation in schools' annual performance ratings, then a large portion of this variation would disappear when the same group of students is evaluated. This does not appear to be the case. This result therefore suggests a more general problem of stability (Perry, 2016b).

5.3. The Consistency of Value-Added Ratings

The previous section discussed the stability of school effects and how examining the volatility of schools' ratings over time can provide indirect evidence of Progress 8's validity. When looking across time however it is difficult to distinguish whether the changes in pupils' progression are due to genuine deviations in school effectiveness, variations in examination systems or the limitations of value-added measures as some degree of variation is expected. For this reason an additional branch of research has investigated the consistency of school performance ratings at a single point in time. That is to say, that these studies assessed whether the same substantive judgements are made about schools' effectiveness when their impact is evaluated using different indicators. This literature is reviewed below.

5.3.1. The consistency of effects across different types of output

One of the key roles of schools is to help improve students' cognitive development. Educational policy and research therefore place considerable emphasis upon examinations which assess students' mastery of specified material (Scheerens, 2013). Schools' results can be assessed in two distinct ways. Progress 8 is primarily concerned with educational quality. That is to say, that it strives to identify the institutions that help students to make the most academic progress once differences in prior attainment have been taken into account. Many of the early educational effectiveness studies however had a slightly different focus, to evaluate whether schools were able to reduce the attainment gap between specified groups of students. The idea being that effective schools would contribute to social justice by helping to address inequality within society, most commonly socio-economic disadvantage (see for example Edmonds, 1979). Whilst modern research has more modest expectations of schools' ability to address fundamental civic problems such measures still inform political thinking, policy and research, and have

[12] Making Good Progress was a large scale DfE study that used teacher-assessments to evaluate the progress of 148,135 KS2 and KS3 pupils, from 342 schools in 10 local authorities. It contains three consecutive years of data for students in the specified age range, including, crucially, the study periods between the national curriculum tests (see DfE, 2011 for further details).

helped to construct a rounded picture of schools' impact. The extent to which schools can purposefully impact upon the performance of particular subgroups is discussed below. At this stage, though, it is sufficient to highlight that the two goals do not always go hand in hand, and sometimes conflict with one another. Thus, whilst schools that are effective for one group of students tend to be effective for all, there can sometimes be inconsistencies when effectiveness is considered from both perspectives simultaneously (see, for example, Thomas *et al.*, 1997a or Dearden, Micklewright and Vignoles, 2011).

Most people would agree however that there is more to students' education than the acquisition of academic knowledge. Schools can also be evaluated based upon outcomes such as their contributions to students' attitude, social development, moral values, personal competencies (e.g. reflection and initiative) and affective state (e.g. psychological health and well-being) (Eisner, 1993; Oser, 1994; Raven, 1991, Cheng, 1996; Lewis and Tsuchida, 1997). Within many countries schools are therefore expected to pursue objectives that do not directly relate to the student attainment. Often with the expectation that schools will help to compensate for aspects of children's upbringing which are believed to be lacking. This is one of the reasons lessons such as civic education are taught in many educational systems, as a means instilling within pupils everyday knowledge, values and social skills that are not necessarily provided at home (Delors, 1996). Since Progress 8 is not intended to evaluate these types of outcomes, an in-depth review of this literature it unwarranted. It is nevertheless interesting to note that effectiveness research has found weak or non-existent associations between schools' cognitive outputs and alterative indicators of effectiveness (see, for example, Gray, 2004; Knuver and Brandsma, 1993; Smith *et al.*, 1989; Thomas *et al.*, 2000). Now, it may legitimately be the case that the proficiency of a school in teaching traditional curricula has little correlation with its ability to prepare students for adult life. Especially since the available evidence suggests that schools' influence upon these kinds of outcomes is minimal (Knuver and Brandsma, 1993; Opdenakker and Van Damme, 2000). One could even argue that due to the finite time available for instruction and the ever increasing demands that are placed upon teachers and schools, the pursuit of additional goals has the potential to divert attention from the task of raising attainment. However, whilst there is some evidence that schools can teach this material effectively by integrating it into their core curricular (Creemers and Kyriakides, 2008), and that reciprocal relationships sometimes exist between the achievement of cognitive outcomes and affective outcomes (Knuver and Brandsma, 1993), it would appear that for the most part schools' cognitive value-added ratings do not reflect the schools' success in other areas. There is therefore little evidence that value-added ratings based on different types of outputs agree with one another.

At this stage it should be re-iterated that there are two potential explanations for these differences and the other inconsistencies noted in this section. The variation in effectiveness ratings may stem from schools genuinely being effective in one area and not in another, or they may stem from random-error and/or bias in the underlying data. In this instance it is tempting to view the former as being more likely, however, it is important not to rule out the latter.

5.3.2. The consistency of effects across departments/subject-areas

Progress 8 is a composite measure that evaluates students' learning across multiple subject areas. Whilst it is convenient for educational stakeholders to have generic measure of schools' influence, a key question that any critique must address is whether it is appropriate to evaluating learning in an aggregate manner. In other words, do schools that provide effective instruction in one subject area tend to be more effective overall or does the effectiveness of an institution depend on which output is evaluated? To help make this assessment this thesis draws upon the findings of educational effectiveness research, where it has become common practice for studies to simultaneously operationalise several measures of schools' output.

Overall this evidence base suggests that there is a low-moderate level of consistency in schools' departmental effects (Fitz-Gibbon, 1997; Reynolds *et al.*, 2014). There is, however, also a consensus that school effectiveness is multi-faceted (Thomas, 2001). That is to say that the effectiveness of a school varies depending on the specific outputs and pupils that are evaluated. The precise level of consistency that is observed in schools' departmental ratings is therefore context dependent (Mortimore *et al.*, 1988; Bosker and Scheerens, 1989; Luyten, 1994; Sammons *et al.*, 1996). Students' performance tends to be most comparable in subjects that emphasise key skills (i.e. Maths, reading and writing) (Teddlie and Reynolds, 2000). This is attributed to the overlap in their content. Departmental effects in primary school are also more consistent than those in secondary schools, which makes sense because primary school teachers are usually tasked with the delivery of students' entire curriculum, whereas, in secondary education teachers tend to specialise in specific areas (Teddlie and Reynolds, 2000). Secondary education is therefore delivered by a several groups of individuals which increases the potential for disparity between students' instructional experiences. Other explanations of this phenomenon have however been posited, including the fact that smaller number of subjects are evaluated at primary level (Fitz-Gibbon, 1997). If, for example, the correlations between students English, maths and averaged English and maths scores were computed, it should not be considered surprising that both individual subjects share a close association with the aggregated score as each contributes 50% of the rating. And equally, unless all value-added estimates are 100% accurate, one would expect departmental effects to be more comparable when the same students study each subject (Reynolds *et al.*, 2012). This occurs most often in primary education schools where all of the students in a given cohort study the same curriculum, and in compulsory subject-areas during secondary education.

Given that the results of these analyses deviate, the research of Thomas *et al.* (1997b) and Telhai *et al.* (2009) is presented as contextually relevant exemplar:

Thomas et al. (1997b)

In this UK based study Thomas *et al.* used multi-level modelling to examine the characteristics of school and departmental effects. In particular they wanted to establish the magnitude of departmental effects, whether some schools were consistently effective or ineffective across subject areas and whether differences in departmental effectiveness persistent over time.

Their results indicate that whilst there were important differences in schools' overall performance (6.2% of the variance in total GCSE scores was explained by schools), schools' performance deviated substantially across subject areas (between 4.1% and 15.4% of the variance in each subject areas was explained by schools). In fact, with the exception of GCSE English, these variations were either comparable to, or larger than, the overall differences in schools' scores.

It was also shown that departmental effects were higher in non-compulsory subjects such as English literature, French and History, and lower in compulsory subjects. The aforementioned example of GCSE English being a particularly extreme cases, presumably because students will develop their language skills outside as well as inside of school (Thomas *et al.*, 1997b).

This outcome suggests that evaluations of school performance should consider departmental-level variations in effectiveness. It is important to acknowledge, however, that the aforementioned effect sizes were calculated using three years of GCSE results. That is to say, that the author utilised a longitudinal research design that siphoned of any year-to-year variance. The percentage of variance attributable to the school and departmental effect during any given year was therefore slightly higher than a traditional design would have calculated.

To address the second research question, whether some schools were consistently effective or ineffective across subject areas, the group examined the correlations between schools' overall performance ratings (based upon students' total GCSE scores over a 3 year period) and their departmental effects (GCSE scores in maths, English, English literature, French, history and science). Correlations of 0.38 to 0.52 were observed between total GCSE performance and specific departmental scores, and relationships of 0.20 to 0.72 were found between the individual subject scores. Since all associations were positive, there was at least a minimal level of agreement between the measures. The relationships, however, were far from perfect. This signifies that schools which were differentially effective in one area were not necessarily as effective in others. Similarly, whilst some of the strongest associations seem logical, e.g. the 0.72 association between scores for English and English literature, other subjects with overlapping content were oddly disparate, e.g. the 0.35 correlation between maths and science. This reinforces the observation that the concept of effective and ineffective schools may be too simplistic to capture the essence of schools' performance.

To complicate matters further, the differential effects did not persist over time. As a follow-up to the aforementioned analyses Thomas *et al.* (1997b) calculated the correlation between school and departmental value-added results 1 and 2 years apart. Whilst the schools' effect upon total GCSE examination score was reasonably consistent (0.85-0.88 and 0.82 for ratings 1 and 2 years apart respectively), departmental effects were much more volatile (0.48-0.92 and 0.38-0.71, for ratings 1 and 2 years apart). This means that in some subject areas as little as 23.0% of the variance in schools' value-added scores was common between results 1 year apart and just 14.4% was shared across two years. Both of these figures refer to the consistency of ratings for schools' French departments. Beyond this, however, there was no discernible pattern to the results. That is to say that no other subjects or groups of subjects emerged as being clearly more consistent or volatile over time than the others.

Finally, in an attempt to summarise the overall effect of this volatility, the authors used two sets of criterion to identify schools that had been differentially effective across the majority of the measures. The first set of indicators identified schools that had a statistically significant positive (or negative) outcome in their overall value-added assessment and 2 or more statistically positive (or negative) scores in their departmental ratings. The second set distinguished schools which either had a statistically significant positive (or negative) overall ratings and/or significant results on each of their departmental ratings. Additionally, in both instances a school could not be regarded as differentially effective or ineffective if a statistically significant result contradicted that judgement. With a single year, only a small minority of school met these benchmarks. Based on the former criterion, for example, just 13 schools from the 1991 sample (14%) were identified as being consistently effective, and 15 (16%) were identified as consistently ineffective. And over time the number of consistent results diminished even further. Based on the same criteria, just 3 schools (3%) were judged to be differentially effective for all three years of the study, and only 3 (3%) were judged ineffective. The majority of schools therefore did not have clear cut results.

Overall, the evidence presented in this study therefore suggests that the effectiveness of departments within a school can vary substantially. In fact, in most cases the variation within schools will exceed the variation between schools. These relative strengths and weakness would be masked in a school-level measure of school performance. Such measures therefore provide an incomplete measure of school performance. Despite this, the authors stress that schools' effect upon total GCSE performance still provides an important indicator of overall effectiveness, provided that it is presented alongside more detailed breakdowns of schools' performance. They also recommend that any ratings take into account at least three years of performance data as they did, or random year-to-year fluctuations are likely to interfere with the assessment. Though personally, we do not see the best course as being this clear. Whilst a reasonable proportion of the year-to-year variance in Thomas *et al.*'s (1997b) overall value-added calculation cancelled itself out, this was not the case for the individual departmental measures. This becomes clear when one compares the percentage of variance that is attributed to time in the total GCSE score calculation (1.1%) with the unstable component of the departmental effects (1.8% to 7.8% of the total variance). The ratio of school to temporal effect was therefore almost 6:1 in the overall calculation, yet as low as 1:1 in one subject area. If one recalls that this variation could theoretically be attributed to random measurement error and/or uncontrolled differences in schools' intakes rather than changes in school practice, then the validity of the departmental-level measure is far less assured.

Telhai et al. (2009)

These five authors examined the size and stability of departmental effects within a sample of 450 English secondary schools. Their focus was exclusively upon history and geography which were optional subjects in all institutions. Substantial differences in departmental effects were found. Specifically 44.4% of schools had departmental results that could be distinguished from one another. The relative performance of departments, however, varied significantly over time with few managing to persistently outperform the others. Within their sample, for example, if the value-added rating of a school's history department exceeded that of their geography department, there was only a 60% chance that this difference in performance would persist during the following academic years. Likewise, if the history department performed worse than the geography department, there was a 59.1% chance that it would do so again. Both percentages are close to 50%, meaning that historical data on departments' performance could not predict which would perform best in the future.

This conclusion is supported by subsequent calculations that evaluated the number of schools in which the history/geography department consistently outperformed the other for four consecutive years. These showed that only a fraction of departments were differentially effective throughout (4 of the 264 schools with moderate to large departments). What is interesting, however, is that when the consistency of performance was assessed in relation to students' raw-attainment, the stability of departmental ratings hardly changed. The variation therefore seems to stem from the volatility of test scores. There are two possible explanations for this (*Telhai et al.,* 2009), either departmental value-added measure are adversely effected by the influence of construct irrelevant variance, that is to say, the idiosyncrasies of the pupils that happened study for the stated qualification at a given point in time, or the effectiveness of individual departments was so volatile than it changed on an annual basis.

The findings of these studies challenge the validity of Progress 8 ratings. The research suggests that there may be significant differences between the effectiveness of departments within a school. From one perspective then departmental-level breakdowns of the value-added figures would provide more targeted information that their school-level equivalent. However, department-level value-added scores are also more volatile, which is itself may be indicative of invalidity and of there being unacceptable levels of construct irrelevant variance within the results. Within comparable models there has therefore a tension between the need for specificity and stability. It is also worth acknowledging that this conflict is problematic whether the departmental-level volatility reflects genuine changes in the underlying construct (i.e. departmental effectiveness) or arises because fewer observations take place (which is likely to increase the amount of construct irrelevant variance). It would not, for example, be rational to make educational decisions based on department-level ratings if the performance of department varies dramatically year-to-year. And whilst averaging students' performance across several subject-areas might help to stabilise performance ratings, one could argue that this only masks the problem.

5.3.3. The consistency of effects across pupil groups

Another concern is whether schools have a comparable impact upon different pupil-groups (Reynolds *et al.*, 2014). Theoretically speaking there are numerous within-school processes that have the potential to make some schools more effective at enhancing the performance of particular sub-groups. Factors such as the policy for designating students into classrooms and the allocation of instructional resources may intentionally or unintentionally favour some students, especially if the schools has explicitly implemented compensatory programs to enhance the progress of disadvantaged students or provided enrichment activities for gifted students (Dearden, Micklewright and Vignoles, 2011). The students themselves are also diverse and may have different reactions to the same instruction (Thomas, 2001). This type of variation presents a problem for school-level indicators such Progress 8 as the more differential school effects are the less value there is in estimating a school's influence upon the average or 'typical' student.

Four types of relationships have been theorised (Jesson and Gray, 1991). These are depicted in Figure 5.3.3a:

Figure 5.3.3a: Four conceptions of differential effects

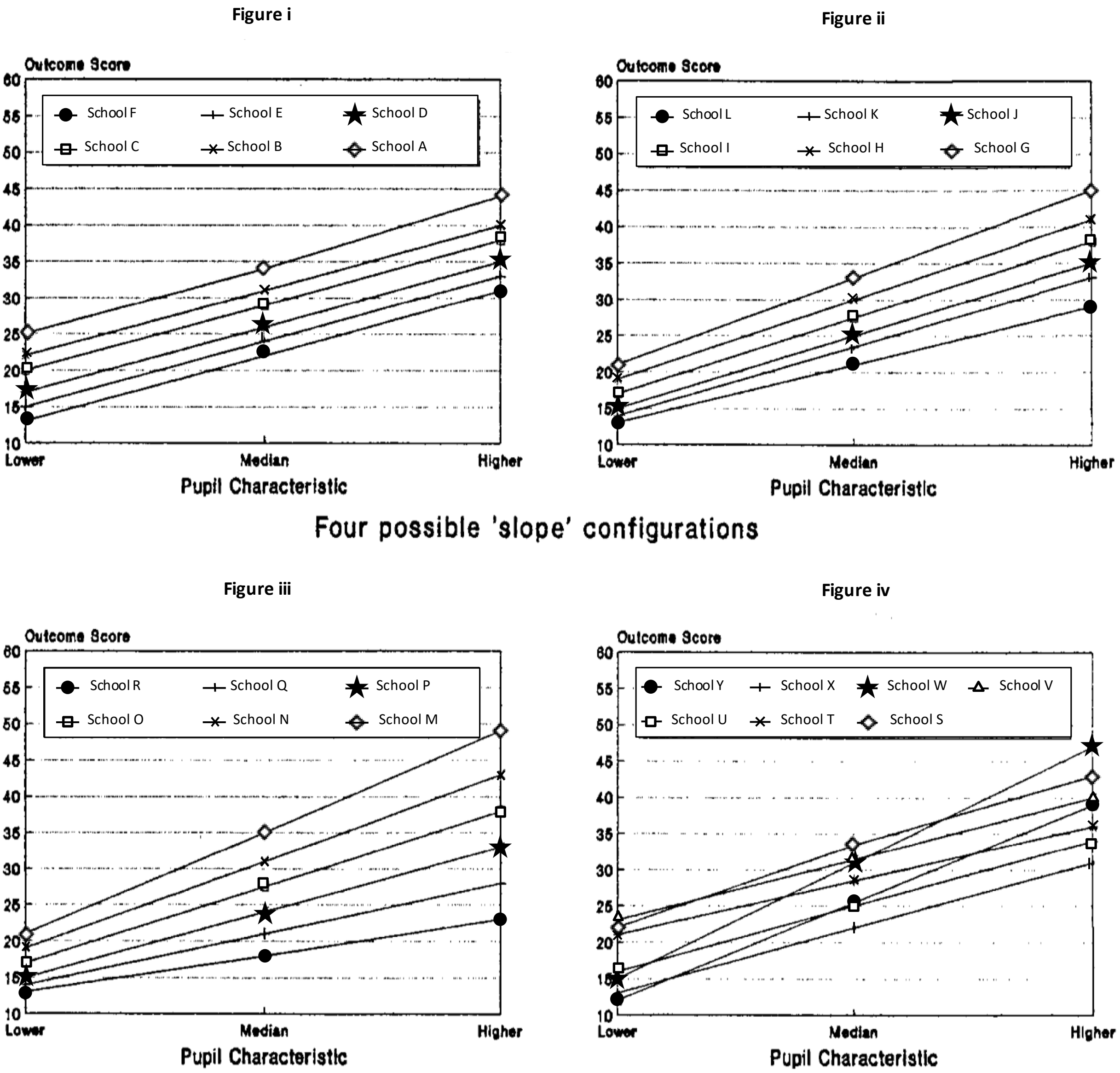


Figure i

Figure ii

Four possible 'slope' configurations

Figure iii

Figure iv

* Graphs taken from Jesson and Gray (2001).

School-level performance ratings have the greatest validity in the first scenario (Figure 5.3.3ai). In this instance the results of each school, i.e. the upward sloping lines, suggest that there is a consistent relationship between students' characteristics and performance. That is to say, that whilst attributes such as ability, prior-attainment, gender, socio-economic status and ethnicity may influence students' attainment, the effects are uniform across all schools. Thus, whether one compares the performance of students with a high, medium or low level of a characteristic the same schools are judged to be the most effective. In fact, in the purest form of this relationship even the magnitude of any deviations will remain the same (i.e. the distance between lines is constant). Under these conditions an accurate measurement of school-level effectiveness would be equally applicable to all students. It is assumed, however, that some relationships may not adhere to this pattern. The three remaining relationships deviate from this exemplar in ways that make it increasingly difficult to interpreted schools' performance as a whole.

The results depicted in Figure 5.3.3aii share many of the same attributes. A positive relationship between the characteristic and performance is seen in all schools and whether the performance of students who had a high, medium or low prevalence of the factor is compared, the same substantive conclusion is reached. In this instance however the slopes are no longer parallel. That is to say, that the variation in students' learning is slightly higher amongst the high prevalence students than amongst students with a low prevalence of the characteristic (or vice versa). In such instances, as long as the deviation is minor, school-level ratings of effectiveness retain most of their validity, though they do provide a slightly inferior representation of the learning of individual students.

If the differential effects are larger, however, the slopes of each school 'fan out'. Schools which encourage similar outputs from the low-prevalence students can then elicit distinctly different results from high-prevalence pupils (or vice versa). This effect can be seen in Figure 5.3.3aiii, where for example, School O and P are roughly 5 output points apart amongst low-prevalence students, 7.5 points apart amongst medium-prevalence pupils and 12.5 points apart among the high-prevalence pupils. A school level summary of this information would therefore omit meaningful deviations in schools' effects.

In the final diagram the profile of schools' results deviate to a far greater extent. This makes the overall effect of the factor difficult to interpret. Consider, for example, the slopes of school S, V and W. If the learning of low-prevalence students is considered School V has the best output. School S has the best results amongst medium-prevalence students and School W has the highest attaining high-prevalence pupils. A school-level summary of this information would omit all of this information.

All four of these examples are of course hypothetical cases. Real performance data is usually harder to interpret and will contain some erroneous results that don't adhere to the predominant type of association.

Academic researchers have come to different conclusions on this matter. Whilst many studies have reported that schools were particularly effective with certain sub-groups of pupil (Dearden, Micklewright and Vignoles, 2011; Goldstein *et al.*, 1993; Sammons *et al.*, 1993; Smith and Tomlinson, 1989; Strand, 2010; Nuttall *et al.*, 1989; Thomas *et al.*, 1997a; 1997b; Thomas, 2001) others have failed to find evidence of differential effects (Jesson and Gray, 1991; Rutter *et al.*, 1979; Willms, 1986). For this reasons, Teddlie and Reynolds (2000) extensive review of the subject matter categorised the evidence base as being inconclusive.

The most consistent evidence relates to prior-attainment and/or ability (Sammons *et al.*, 1993; Strand, 2010; Thomas, 2001; Thomas and Mortimore, 1996; Thomas *et al.*, 1997a). Smith and Tomlinson (1989), for instance, found greater variety in the performance of pupils with high prior-attainment. Their explanation for this was that having low prior-attainment in basic skills such as English or maths may prohibit the future learning of students. This serves to cap the effect of instruction meaning that low ability students would likely achieve similar Key Stage 4 outputs at most schools. Conversely, the outputs of pupils who excelled at Key Stage 2 are more varied as these individuals are equipped to progress and are therefore more sensitive to differences in the quality of their secondary education. This sentiment has been echoed in more recent research and has led the now widespread view that it is crucial to deal with gaps in attainment early on in the educational process before they become entrenched (Dearden, Micklewright and Vignoles, 2011; Scheerens and Bosker, 1997; Strand, 2010; Kyriakides, 2004). Dearden, Micklewright and Vignoles (2011) however point out that the appearance of differential effects can be exaggerated if a significant number of students score the top mark of their Key Stage 2 attainment tests (see Section 4.3.2 on ceiling effects and their impact upon value-added measure).

Even here, though, where the discrepancies are greatest, the differential effect of schools is small in comparison to the difference between institutions (Reynolds *et al.*, 2014; Strand, 2010; Thomas, 2001; Thomas et al., 1997a). Therefore, whist some researchers claim that more than one-quarter of the secondary schools in England are differentially effective for students of differing prior-attainment levels (Dearden, Micklewright and Vignoles, 2011), the value-added scores of low, medium and high performers all have close association with the school average (Dearden, Micklewright and Vignoles, 2011; Thomas et al., 1997a). This implies that that schools' performance slopes are more akin to the functions in Figure 5.3.3a.ii and 5.3.3a.iii than those in Figure 5.3.3a.iv, and that the extent of any inconsistencies is insufficient to challenge the validity of school-level ratings.

With regards to other categorisations of pupil, research has produced evidence that schools can be differentially effective for pupils of different ethnicities, socio-economic status and gender (Strand, 2010; Thomas, 2001; Thomas *et al.*, 1997a). Though these deviations were small, both in relation to the differences between schools and the differential effects associated with prior attainment. The presence of such effects does not therefore challenge the construct validity of school-level indicators such as Progress 8.

That being said, it is worth acknowledging a potential source of tension. Whilst this thesis is primarily concerned with the quality of school practices, schools can also be evaluated based on their ability to address inequalities within society (see Section 7.2.2). From the latter perspective, it not only appears as though schools have a limited capacity to close pre-existing attainment gaps, but that schools that are considered effective in absolute terms may be characterised as such because they are the most efficient at maximising the attainment of advantaged students (i.e. middle-class student, female students and pupils with high prior-attainment). A school could therefore be viewed as being effective from one perspective and ineffective from another (Thomas et al., 1997a). This statement, however, does not apply to student ethnicity. This is because ethnic minority students tend to have lower levels of absolute attainment but make greater academic progress. Their sensitivity to differences in educational quality should therefore help to reduce rather than exaggerate inter-racial attainment gaps (see Thomas *et al.*, 1997a; 1997b).

5.3.4. The consistency of cohort ratings

Few studies have investigated the consistency of cohort ratings. A fact that some attribute to the general shift of interest from schools to teachers within academic research (Muijs *et al.*, 2014) and the assumption that when performance deviates, this reflects differences in the effectiveness of individual teachers rather a fundamental problems with the value-added methodology. The evidence that is available, however, suggests that value-added estimates vary substantially across the different year-groups of a school.

Mandeville and Anderson (1987) were amongst the first to study this phenomenon. Their results showed that amongst their sample of 423 South Carolinian elementary schools the consistency cohort value-added ratings was "discouragingly small" (Mandeville and Anderson, 1987, p. 212, ln 10-11). In fact, the highest correlation that they observed was r = 0.17, with the majority of associations being far less substantial. The pair therefore concluded that this level of inconsistency contradicted the notion of a 'main' school effect. Mandeville (1988) later reinforced these observations by repeating the study a year later, using the same methodology, and an additional years' performance data. The reported cross-cohort correlations were equally small, with a median r correlation of 0.07 (-0.2 to 0.18) between cohorts in English and a median r correlation of 0.13 (0.00 to 0.19) in maths.

Bosker (1989 secondary cited from Bosker and Scheerens 1989) have also assessed inter-cohort consistency in Dutch secondary schools and reported modest correlations of r = 0.50 and 0.47 in arithmetic and language respectively.

These studies, however, are somewhat dated and are not contextually specific. They therefore have a limited capacity to inform us about the consistency of value-added evaluations in UK-based secondary schools. Thankfully, one author has recently addressed the gap in the literature.

In his critique of English value-added measures, Perry (2016a; 2016b) compared the progress made by six consecutive year-groups (years 3-9) using a simplified version of Contextualised Value Added and data from the DfE's Making Good Progress study (DfE, 2011). His results suggest that within secondary schools there were correlations of approximately 0.7 between cohorts that were one year apart and a correlation of 0.45 between cohorts two years apart. Similar associations were observed in primary school cohorts but the relationships tended to be weaker. Perry therefore concluded that cohorts' value-added "results cannot be safely generalised from a single cohort to the school at large" (Perry, 2016b, p 211, ln 11-12).

The nature of these associations was interesting. Were the instability in cohorts' value-added results attributable to random fluctuations in cohort characteristics or random measurement error, one would have anticipated that the correlations between group ratings would be similar irrespective of how many years they were apart (Perry, 2016a). This was not the case. In fact, the magnitude of the correlation between cohorts' value-added ratings appears to have been dictated by their proximity. Although one can only speculate about the mechanisms behind the phenomenon, this may indicate that cohorts that have spent a comparable amount of time at the school are more likely to have received the comparable educational experiences (Perry, 2016b). The more separate the groups however the less likely it is that they will have shared key educational inputs, the same teachers and curricular materials for example. This would be consistent with the view of educational effectiveness researchers, who claim that teachers and classroom-level instruction are the locus of educational effects (Creemers and Kyriakides, 2008; Muijs *et al*, 2014). Though in Perry's view, given the low level of stability for teacher-level effectiveness ratings (discussed in next section), it is more likely that during certain periods a combination of beneficial factors acted together to create a more effective learning environment that benefited students who attended the school through that period. Readers should be cautious of inferring too much from these speculations, however, as correlational evidence does prove causal relationships exists. This is a recognised weakness of cross-sectional research designs (Coe and Fitz-Gibbon, 1998; Marshal *et al.*, 2011). It is possible therefore, that the difference in student outcomes were actually due to extraneous factors, such as differences in students' background that evaded the value-added controls. Particularly since Perry's model only controlled for differences in the prior attainment, gender and FSM status of pupils. These factors would however have had to change gradually over time, in-line with the observed relationships. The skills and curricular assessed in each school-year also change, and the diminishing correlations might reflect this disparity.

Although the Perry's (2016a; 2016b) study relied heavily upon teacher-assessments, sufficient steps were taken to rule out any bias that this might have introduced. It seems fairly clear therefore that whether the variation in inter-cohort ratings is due to genuine differences in the quality of students' instruction, the make-up of each cohort or another shortfall in the value-added methodology, value-added ratings do not apply uniformly to all year groups within a school. Certainty it would seem that the chance of the current Year 11 cohort's ratings providing meaningful insight to parents who are in the process of selecting their child's secondary school is low. The systematic nature of the inconsistencies may signify however that at least a portion of this instability does reflect genuine changes in schools' effectiveness. This research evidence could therefore be considered sufficient to discourage some but not all applications of cross-sectional value-added designs.

5.3.5. The characteristics of teacher-level value-added scores

This study is primarily concerned with the reliability and validity of school-level Progress 8 ratings, as used by the DfE in England. Within the academic literature however the focus has gradually moved away from school-level assessment and towards classroom-level interactions. This is where learning actually takes place and where educational effects are highest (Luyten, 2003; Reynolds, 2008). Whilst a full critique of teacher-level value-added assessments was considered to beyond the scope of this thesis, it would therefore have been ill-advised to overlook them completely. Especially as these lower-level units provide further insight into the properties of school effects, as defined by value-added models. This brief review will therefore discuss the most robust research evidence alongside the other indicators of consistency. This evidence comes from the United States where teacher-level value-added measures have been more openly embraced in state policy (McCaffrey and Hamilton, 2007). New York City Department of Education, for example, went as far as publishing teachers' scores as a means of holding staff to account (Amrein-Bearsley, 2014). Although this caused considerable controversy, including law suits and a statement from the American Educational Research Association which warned educational stakeholders of the limitations of value-added evidence and advocated 8 technical requirements that should be met before implementing such measures (See, AERA, 2015). The three cited studies are also particularly useful as the researchers were able to implement experimental or quasi-experimental research designs that provide more effective controls for any differences between pupils.

The first paper is Nye *et al.*, (2004). Nye and her colleagues re-interpreted data from the Tennessee Class Size Experiment (also known as Project STAR), a project which followed the progress of Kindergarten students from 79 elementary schools for four consecutive years of their education. Within each school, students were randomly allocated to classes (of differing size) that they would retain throughout the experiment. Teachers were then randomly assigned groups of students on an annual basis. Theoretically, this process ensured that any differences in students' performance could be traced to one of three sources; the differences in class size, differences in teacher effectiveness or sources of invalidity in the experiments design (Shadish *et al.*, 2002). Their results indicated that each year between-classroom differences (teacher effects) accounted for 12.3-13.5% of the variance in mathematical learning gains and 6.6-7.4% of the deviation in reading. The magnitude of these effects is consistent both with the non-experimental research cited by the authors (median r-squared score = 0.11 across the two subjects), including the fact that learning gains in Maths were noticeably higher than in reading (Teddlie and Reynolds, 2000). Most researchers presume that this is because a higher percentage of parents provide reading instruction at home, though it is also possible that the teaching of mathematics is more varied. Another consistency is that the differences in teacher effectiveness explained far more variation than the differences between schools (average of 6.1% of the variation per year in mathematics and 4.7% in reading respectively), but less than within-classroom differences (average of 34.4% of the variance per year in mathematics learning and 35.6% in reading). The estimated teacher effects were normally distributed and of appreciable size, meaning that if students had a 25th percentile teacher (a less effective teacher) instead of 75th percentile teacher (effective teacher), or an 90th percentile teacher (a very effective teacher) over an 50th percentile teacher (an average teacher), the change in their educational experience would account for approximately 1/2 of a standard deviation in mathematics and 1/3 of a standard deviation in reading. This evidence therefore suggests that there were meaningful differences in teacher performance that would not be captured by school, or even departmental level summaries.

The two remaining studies, Kane and Staiger's (2008) and Chetty *et al.* (2011) were not true experiments but are nevertheless more robust than the standard correlational studies.

The initial purpose of Kane and Staiger's (2008) research project was to determine whether National Board for Professional Teaching Standards certified teachers were more effective than uncertified teachers. As part of this assessment the authors implemented a quasi-experimental design that compared the effect that the two groups had upon students' examination scores. 78 pairs of elementary teachers were chosen to take part in the study, each comprised of a certified and uncertified individual. These were selected so that both teachers were from the same school, taught the same grade of student and could be allocated classes of students at random. Students were then provided with 1 year's instruction, after which the performance of their teacher was evaluated using a variety of indicators. Attainment measures were then retaken 1 and 2 years after the experiment had been completed to assess the longevity of teachers' effect. This research design therefore helped to negate assignment bias, but was inferior to a true randomised control trial as students were not randomly allocated to classrooms.

Most of the author's findings reinforced the results of Nye *et al.* (2004), in so much as the magnitude of teacher effects were broadly comparable, the majority of the variation in students' performance was shown to exist within rather than between schools and teacher effects in mathematics tended to be greater than those in English. A more concerning finding, however, was that the influence of teachers faded out over time. That is to say, that after students were exposed to a differentially effective teacher, the initial differences in students' achievement deteriorate by approximately 50% per year. This meant that after just two years the student's attainment was no longer distinguishable from other students. This could signify that value-added effects are short-lived. If this is true this would imply that aggregated measures of school-level effectiveness such as Progress 8 not only oversimplify the multi-faceted nature of school effects but also over-estimate teachers and schools' long-term influence.

The authors were quick to point out however that one should not jump to conclusions too quickly. Whilst it would indeed be troubling if students were simply forgetting the knowledge that they had acquired, or if value-added were measuring something short-lived like the benefits of teaching to the test, alternative explanations of this relationship are possible. It might be, for example, that the content of students' curricular changes as they progresses through the education system. A portion of the knowledge that is acquired during the school year may therefore be redundant the following year. Whilst this is undoubtedly true to some extent, intuition says that core subjects such as maths and English Language are likely to have wider ranging applications, especially the basic skills that are taught in elementary school education. This suggestion therefore grates against the notion of scaffolding and structured curricular, as well as the theory that pupils with high-prior attainment are more able to capitalise upon high-quality instruction (see Section 5.3.3). Nye *et al.* (2004) also posit that the initial impact of effective teachers might have spilt over, increasing the prior-attainment scores of the class in the latter value-added calculations, thus lowering their apparent ratings. Such a mechanism is plausible, though it still conflicts with the aforementioned differential effects (see discussion above regarding schools' impact upon students with differing levels of prior-attainment). Crucially, however these or alternative explanations might account for the fading out of learning gains without implying that the long-term effect of teachers has been exaggerated. The available research evidence does not allow us to distinguish between these and other eventualities. The topic therefore merits attention in future research. Nye's observations however are consistent with the findings of other research, including Krueger and Whitmore (2001) and Konstantopoulous's (2007; 2008) reanalyses of the data from the aforementioned Tennessee Class Size Experiment, as well as non-experimental studies (McCaffrey *et al.*, 2004; Jacob *et al.*, 2010; Rothstein, 2010).

The final study, Chetty *et al.* (2011), was eventually re-published as two separate papers (Chetty *et al.* 2014a; 2014b). The first of these evaluated whether value-added measures provide an unbiased measure of teachers' impact, whilst the latter was successful in connecting the construct to long-term outputs such as higher level of college attendance and future salary. It is the foremost report that is of most interest, due to its quasi-experimental design.

In this section Chetty *et al.*, examined the impact that teacher mobility has upon class performance ratings. Theoretically speaking if a differentially effective teacher leaves a school their department ratings should decline. And conversely, if an effective teacher is recruited to a department its ratings should rise. In fact, if one uses a longitudinal effectiveness model to calculate teachers' ratings and the distribution of any unobserved determinants remains constant, then the effect should be predictable. For example, if a maths teacher with a VA estimate 0.3 points above their colleges' proficiency rating leaves a school with three classes per grade, then the average class rating should fall by 0.1 (0.3/3) (Chetty, *et al.*, 2011). In practice, of course, the change in groups' performance will deviate from this amount. This is because all value-added estimates are imperfect predictors of future performance. However, since the influence of random variance will cancel out during repeated observations, the presence of any systematic deviation is indicative of bias (Chetty, *et al.* 2014a). The three researchers therefore used these change events to create a natural experiment.

Based on observations of over 4000 staffing changes within English and mathematics departments they calculated that the entry of a highly effective teacher (top 5% of the performance distribution) raised the mean test scores of their new grade-department cells by an average of 0.035 standard deviations. The egress of an effective teacher caused results to fall by a similar amount (0.045 standard deviations), whilst the entry and exit of an ineffective teacher (rated in the bottom 5%) cause departmental scores to fall/increase by 0.021 and 0.034 standard deviations respectively. The fact that these figures were highly comparable with the changes in departmental performance that were predicted based on the teachers' value-added performance data (0.042, 0.042, 0.033 and 0.034, respectively) therefore implied that the estimates were reasonably accurate and free from substantial sources of bias. Because of this and the fact that the influence of teachers is widely dispersed, the authors go on to conclude that changes in teacher effectiveness should make a significant difference to students' test scores.

Chetty *et al.*'s (2011) conclusion however is surprising given their reported effect sizes. Especially since the initial sections of the report acknowledged that teacher effects are very unstable and have a low level of correlation that deteriorates over time. In elementary schools, for example, correlations of 0.43 and 0.3 were found between the teachers' ratings that were taken 1 year apart in maths and English respectively, which dropped to 0.25 and 0.15 over 5-7 years. The stability of middle-school teachers' ratings was more difficult to calculate but comparable to these figures (see Chetty 2014a p.2607 for further details). In the secondary context, similar problems with inconsistency have been observed across multiple studies. McCaffrey *et al.* (2009), for example, studied the stability of English value-added estimates. Specifically they reported that in consecutive years there were correlations of 0.2-0.7 between teacher's value-added ratings, meaning that less than half of the variation was common between years. This analysis took place at the classroom-level and therefore provides an inexact parallel with the school-level figures that were reported earlier. The finding is nevertheless interesting from a theoretical perspective as one would have thought that the skill and behaviours of specific teachers would have remained more stable than school-level actions, where for example, key members of staff may leave. The finding therefore hints that measurement errors make up a substantial portion of the variation. Likewise, after reviewing current research evidence Amrein-Beardsley (2014) reported that the year-to-year correlation between teacher-level value-added ratings ranges between 0 and 0.5, with most associations being in the 0.2 to 0.4 region. At most therefore teachers' value-added scores can be expected to explain 25% of the variation in the next years' appraisals (0.5 squared), though in practice they will usually account for much less than this (4-16% going off the aforementioned estimates). Taking all of this into account it is therefore conceivable that the immediate effect of a teacher moving school might not be representative of their long-term impact and that Chetty *et al.* (2011) may have been over generous in their interpretation of teachers' influence. In fact, in the line charts of their results (see Figure 3, Chetty 2014a, p. 2620) it appears as though 1-2 years after a teacher transition the initial improvement (/declines) in mean school-grade-cohort test scores either reduced dramatically (in 1/4 of the aforementioned scenarios) or reversed (in 3/4 of the aforementioned scenarios).

Collectively these three experimental studies therefore suggest that individual teachers have a small but meaningful impact upon students' performance. The stability of the reported effects however is concerning, especially when one recalls that this may be interpreted in different ways. If we are prepared to take the results at face value, one reaches the conclusion that teachers' influence is very volatile and likely to fade out over time. If this is the case, then it follows that the provision of teacher-level value-added data may be useful for formative purposes, though the scope for further application would be limited (see Amrein-Beardsley, 2014 for a discussion of why these ratings are likely to be damaging in high stakes contexts). This explanation also implies that school-level aggregations of the data, such as Progress 8 figures would mask a lot of underlying variation. However one could also attribute this instability random-errors and the smaller number of observations per calculation. It may therefore be that a reasonable portion of the volatility is spurious.

6. Methodological Assumptions and the Interpretation of Value-Added Evidence

6.1. Chapter Introduction

One of the surprising characteristics of the effectiveness literature is the range of opinions that are held regarding the worth of value-added models. The proponents of such models claim that the methodology has had a "major impact" (Reynolds *et al.*, 2012, p. 12, ln. 36) by facilitating the detection of "practically (as well as statistically) significant" school effects (Muijs, *et al.*, 2011, p. 24, ln 44-45), as well as factors that are consistently associated with educational effectiveness (Muijs, *et al.*, 2011). Whereas critics argue that the models have "fatal flaws" (Gorard 2010a, p. 746, ln 6) that make them "useless for practical purposes" (Gorard *et al.*, 2013, p. 8, ln 5). In terms of their stability, the results have been interpreted as having "substantial year-on-year-stability" (Reynolds *et al.*, 2012, p. 11, ln. 43) or as so volatile that the results are "meaningless with current datasets" (Gorard *et al.*, 2013, p. 7, ln 39). Whilst some of these differences can be attributed to the context of individual studies or the specification of individual models, the core dispute can ultimately be traced back to the 'fragility' of the value-added assumptions (Marsh, 2011) and how readily these are accepted by researchers (see also Coe and Fitz-Gibbon, 1998). This section explores the differences in the interpretation of value-added data and how this has led to such profoundly different conclusions.

To focus this discussion the recent dialog between Gorard (2010a; 2011a; 2011b; 2011c) and prominent educational effectiveness researchers (Muijs *et al.* 2011; Reynolds *et al.*, 2012) is used as modern exemplar of the longstanding underlying dispute. This case is particularly relevant to our study as it helps to tie together the three interrelated topics of measurement error, the volatility of value-added results and use of probability based statistics. All aspects of the debate, however, seek to answer the question:

> "Is the variation in school outcomes unexplained by student background just the messy study left over by the process of analysis? Or is it large enough, robust enough and invariant enough over time, to be accounted a school 'effect'?"
> (Gorard, 2010a, p.746, ln. 40-43)

6.2. School Residuals: Genuine Effect, Random Error or Bias?

As stated in the introductory chapter of this thesis, school effects are not an observable quality of schools (Gorard, 2011c). In the case of Progress 8, what practitioners call the 'school effect' is merely the variance in students' Key Stage 4 attainment that remains after taking into account the differences in students' prior attainment. That is to say, the amount left over (positive or negative) after deducting the mean performance of all students with the same Key Stage 2 Average Point Score. Phrased like this it is easier to see why some critics argue that the estimates might be wildly inaccurate. After all, any number of factors could impact upon students' progression and cause it to deviate from the typical amount. Or more crucially, impact upon certain types of pupils more than others (Gorard, 2010a). Under what conditions then can we be sure that a school's residual constitutes a genuine school effect?

This question does not have a definitive answer. Two main options have traditionally been adopted by researchers. The first is to argue that all extraneous influences have been taken into account. The best way to rule out external influences is to use an experimental design that automatically balances out both known and unknown factors (Shadish *et al.*, 2002). However, such robust sources of evidence are rare due to the practical difficulties and ethical issues involved in employing randomised interventions within an educational setting. In their absence the main way to examine the validity of Progress 8 directly is to use statistical controls to assess the effect of known but unobserved variables (unobserved in the sense that the variable is not specified within in the model). As discussed in Chapter 4 (and later in Chapter 7), educational researchers have identified a range of student-level characteristics that have verifiable associations with school performance, yet remain largely or entirely outside of schools' control (Creemers, 2007). It stands to reason therefore that if one can demonstrate that the value-added figures are biased by this type of characteristics then one will have increased the evidence for concluding that schools' residuals do not provide an accurate assessment of schools' contribution. And indeed there have been studies which have demonstrated that Progress 8 and comparable models of educational effectiveness are vulnerable to this form of bias (see Leckie and Goldstein, 2019 and Perry, 2016a respectively). The theoretical and technical problems that undermine such modelling were discussed in Chapter 4, but include the practical impossibility of collecting accurate data on all variables and the conceptual problem of distinguishing between school and non-school effects, especially when some variables are unknown or unobserved. So whilst school effectiveness researchers can make claims as to the prevalence of systematic bias, one can never be one hundred percent confident that the relationships are causal or act in the direction that is theorised (see, Coe and Fitz-Gibbon, 1998).

The second approach is to examine the stability and consistency of results across different outputs (see for example Marks, 2015b). This is known colloquially as the indirect approach or examining the results in context (Chapman *et al.*, 2015) and it forms a significant part of Muijs *et al.* (2011) and Reynolds, *et al.* (2012) defence against Gorard's (2010a) criticism of value-added analyses. The rational being, that the process of triangulation should increase our confidence that observable properties of school effects are not freak occurrences (Teddlie and Reynolds, 2000). The process, of course is also imperfect because these assessments can only act as a guard against inappropriate interpretations, they can never establish validity. Were value-added measures to be influenced by an unknown but consistent source of bias, for example, how would one identify this? Furthermore, within the context of secondary education, where the level of instability and inconsistency of value-added ratings across outputs is not sufficient to immediately identify the results as being invalid, how does one make an objective judgement? The problem of interpretation, is then made worse by broad definitions of effectiveness factors (Coe and Fitz-Gibbon, 1998) and the fact that value-added data has also been used to examine the properties of school effects (Gorard, 2010a). In fact, the concept of school effectiveness has evolved from the point that school effects were expected to have an certain amount of duration and scope (Bosker and Scheerens, 1989; Scheerens, 1993) to position where effectiveness is often viewed as a multifaceted construct that varies substantially across outcomes, departments, pupil groups and over-time (Lang *et al.*, 1992; Levine and Lezotte, 1990). In extreme cases it is even presumed that the absolute effectiveness of a school will vary even if the characteristics of the school remain exactly the same, as the external environment and the goals of education will mutate (Creemers and Kyraikides, 2008). From such a perspective, almost any form of variation could be explained post-hoc (Popper, 2005) (see the cautions of researchers about 'fishing for correlations' (Luyten *et al.*, 2005; Scheerens, 1992) or 'data dredging' (Gorard, 2015)). Where one draws the line between a reasonable and unreasonable level of inconsistency is therefore a personal decision and a clear point of distinction between researchers (including the debate between Gorard vs Muijs *et al.*).

6.3. Communicating Uncertainty

The previous section re-introduced the problem of justification and how researchers' assumptions play a decisive role in the interpretation of value-added evidence. In this type of general discussions however it is difficult to grasp exactly what it is that the researchers disagree on and how this could lead to radically different conclusions. This section therefore takes a closer look at debate between Stephen Gorard and five of the most prominent individuals in the field; Daniel Muijs, Tony Kelly, Pam Sammons, David Reynolds and Chris Chapman. The dialog begins by reviewing the author's stances on measurement error and whether they expect inaccuracies cancel out or multiply during the value-added calculation. It then compares their views on probability-based statistics and whether these kinds of approaches can be used to quantify the amount of error that is likely to occur within school-level measurements. Finally, it considers the implications of applying inferential statistics in the context of a national accountability system. The source of all disagreements within each of the aforementioned debates however concerns the nature of measurement error and whether any inaccuracies can be presumed to occur at random.

A. The nature of measurement error

As discussed in Section 4.3.2, all value-added calculations contain inaccuracies as they are based upon imperfect information. These emerge from a combination of omitted variable bias, imprecise measurements, missing data, mistakes during data collection and coding errors. This seemingly innocuous backdrop set the stage for one of the most divisive debates in modern educational effectiveness research. This dialog centred on Gorard's (2010a) article 'Serious Doubts About School Effectiveness' and his claims of propagated error.

In this paper Gorard evaluated the amount of missing and erroneous data within one of the highest quality datasets available to UK researchers, the National Pupil Database. This resource underpins the calculation of Progress 8 figures and therefore has direct implications for this thesis. Despite their renowned quality, however, Gorard concluded that the records contained enough measurement error to invalidate value-added calculations. His reasoning is outlined below.

Gorard (2010a) began by observing that all value-added models calculate the difference between students' actual and predicted attainment. To assess the level of uncertainty in a simple value-added calculation he therefore conducted a thought experiment, wherein each of these measures was assumed to contain a modest error component (a relative error of 10%). The former was presumed to occur directly through the imperfect process of assessing and reporting upon the students' performance (i.e. the error in examination results) and the latter indirectly because the predicted scores are based upon, at the very least, the prior attainment of the pupil in question, and this prior-attainment data would be likely to contain a comparable percentage of error. He then asserted that unless one knows the direction of these inaccuracies (which one would not in any real-world situation), one must allow for the possibility of the errors acting in opposite directions. That is to say, for the student's actual score to be over-estimated and their predicted score under-stated, or vice versa. Therein, the maximum error possible within a simple value-added calculation is equal to the sum of the two initial error components.

This would not be so problematic, he states, were it not for the fact that actual and predicted attainment scores are intended to be very similar. There would be no point in comparing students' performance to an expected level of attainment otherwise (Gorard, 2010a). When the latter is subtracted from the former, one is therefore left with a small residual (the school 'effect') and a sizeable error component. To borrow his example, suppose that the actual point score of a pupil was 100 and that their predicted score was 99. This would make their value-added rating 1 (100-99). However, Gorard argues that the absolute error in the calculation could be as high as 19.9 points, as 10% of 100 = 10 and 10% of 99 = 9.9, and these two errors may occur in opposite directions. Thus, in this instance, he asserts one would end up with a value-added score that could theoretically contain an error up to 1990 times as large as the individuals' residual score. He concluded that we therefore have no idea whether the individual actually improved or not and that the measure would be useless for any practical purposes.

Understandably the article received a response from five prominent educational effectiveness researchers, who sort to defend the validity of their field (see Muijs *et al.*, 2011; Reynolds *et al.*, 2012). The aforementioned papers criticised the following aspects of Gorard's argument:

The first criticism was that Gorard "treats the predicted a score (P_{ks4}), as an *attempted measurement*, which it is not. [Relative error] cannot be applied to a prediction in the way the measure suggests" (Muijs *et al.*, 2011, p. 25, ln 36-38). This is technically true but somewhat of a trivial technicality. Whilst the prediction itself does not involve any measurement and cannot therefore contain measurement error, the estimates are underpinned by measurements which do (see He *et al.*, 2013).

The second criticism concerns the terminology which is used in Gorard's argument. There are two elements to this. Reynolds *et al.* (2012) assert that in Gorard's example he claims to report the maximum relative error in the value-added scores but actually concluded with a statement of the maximum relative error range. They also take issue with the way maximum relative error was calculated. Specifically, that it was expressed as a function of the students' value-added residual (i.e. maximum absolute error / difference between student predicted and actual KS4 scores). This is because the official specification of Contextual Value Added (the model utilised in Gorard's example) used to add an arbitrary number (100 for primary schools, 1000 for secondary school) to the ratings so that the general public would not misinterpret a negative residual as a sign that students had made no progress. Muijs therefore argues that it is the CVA score of 1000 that the maximum relative error should be compared to[13]. This would make the maximum relative error range a fraction of the value that Gorard reported. Both arguments are correct, though one can see why Gorard chose to discuss the potential for error in relation to the component of the model that is actually malleable. Neither point, however, makes any substantive difference. Regardless of the terminology used or whether the calculation estimates the maximum relative error range of the CVA score or the CVA residual, Gorard's maths still dictates that the error in the students score could be as far as 19.9 points out if the errors in the actual and predicted scores do not cancel out. The semantics of which denominator is most appropriate only influences how severe this maximum absolute error is made to sound.

The most significant disagreement, however, concerns the nature of measurement error, in particular, whether Reynolds *et al.* (2012) are correct in their assertion that measurement errors tend to be randomly distributed. If this is the case, any measurement error in pupils' value-added calculations would, they argue, more-or-less cancel itself out when the figures are aggregated to school-level and it would be unlikely that their influence would be systematically different for different types of school. In support of this claim the group cite several studies (Ferrão and Goldstein, 2009; Goldstein *et al.*, 2008; Woodhouse *et al.*, 1996) that have examined the effects of random error within multi-level

[13] In their second retort Reynolds *et al.* (2012, p.8, ln 1-2) actually stated that the measurement error is "of course related to the contextual value-added score (CVA) of 100". However it is presumed they must have wrongly assumed that the example took place within the context of primary education. The mean official contextualised value-added score in Gorard's example would have been 1000 as he explicitly stated that it errors were in Key Stage 4 final attainment and Key Stage 2 prior-attainment data.

effectiveness models. These papers conclude that when their models were adjusted for a lower level of reliability, the effects of error only influenced the fixed part of the model. The school residuals were unaffected. Ferrão and Goldstein (2009) also found correlations of 0.97 or higher between schools' value-added estimates when measurement error is and is not considered. These studies, however, all presumed that measurement errors would be random and developed their research methodologies accordingly. Their results may therefore paint a favourable picture of the situation.

Gorard's (2011a) response was that measurement error cannot be presumed to be entirely or even mostly random and that is unfair to casually assume otherwise. To support this assertion he describes how the population of students that do not claim Free School Meals contains a super-deprived group (this observation was eventually published in Gorard 2012b). The deprivation of these students would therefore go unacknowledged within a model that corrected for the effect of socio-economic status using the FSM indicator. This type of non-random bias, he argues, may have a significant impact upon schools' results.

In combination, the aforementioned disagreements led Gorard and Muijs *et al.* to drastically different conclusions about the potential scale measurement errors and the worth of value-added data. Ignoring the dispute about terminology, Gorard (2010a) states that the absolute error in his example could be as high as 19.9 which means that the actual progression of the pupil would fall be between -18.9 and +20.9. This is then expressed as a maximum relative error of 1990% the size students' residual (Gorard, 2011a, p.17). Muijs *et al.* (2011 p. 25, ln 53-55) however assert that the students' progression will fall between -9 and 11, which represents a range for the Relative Error of 10%. Whilst most decisions do not result in differences of this magnitude, it is argued here that the subjective nature of the available research evidence (see previous section) makes is difficult for researchers to distance themselves from the effects of their pre-existing assumptions.

Furthermore, whilst the examples of systematic bias that Gorard (2010a) provided do not apply directly to Progress 8, it should be acknowledged that failing to account for intake differences is likely to be more damaging than utilising imperfect controls. What is more, recent research has shown that even randomly distributed inaccuracies can lead to systematic bias (Perry, 2019). It is therefore inevitable that some degree of systematic error will occur within Progress 8 results, and Gorard's warning should therefore be taken seriously. Without further evidence or a better understanding of how student-level errors translate into the school-level scores, however, it is difficult to know how much bias is present. Though it should be noted that Gorard's (2010a) calculation emphasised the worst case scenario, given the stated parameters. It was not therefore his intent to argue that the error in value-added calculations is as high as his example suggests, rather that the true progress of the individual could lie anywhere within the stated range and that researchers have no way narrowing the matter down any further without relying upon unjustified assumptions. In fact, in other research papers he identifies mechanisms that would only function if there were a random component to value-added scores (see Gorard *et al.*, 2013). Likewise, Reynolds *et al.* (2012) do not dismiss the notion that there would be some degree of systematic error in value-added measures. The true measurement error therefore lies at an unknown point between these two extreme ends of the continuum.

B. The use of confidence intervals and the legitimacy of viewing uncertainty as sampling error:

The last sub-section illustrated how fundamental differences between the assumptions of Gorard (2010a; 2011a; 2011b) and educational effectiveness researchers' (Muijs *et al.*, 2011; Reynolds *et al.*, 2012) have a decisive influence upon the magnitude of error that these individuals expect to find in schools' value-added ratings. Their debate did not, however, stop there. The authors also disagreed on how uncertainty should be expressed.

As discussed in Chapter 3, Progress 8 figures are not interpreted in isolation, they are presented alongside 95% confidence intervals that are intended to quantify how much confidence one can have in the results. The upper and lower limits of this confidence limit are defined as the school's official Progress 8 rating plus or minus their C.I. value. Where C.I. = 1.96*(standard deviation of Progress 8 scores for all eligible students nationally, divided by the square root of the number of eligible pupils at the school). Schools are then viewed as being distinguishable from the national average only if the entirety of this confidence interval is above or below zero. Likewise, one can ostensibly assess whether two schools are differentially effective by whether their confidence intervals overlap.

These intervals are deemed necessary because each school's value-added score is based upon the performance of a finite group of students (specifically the school's Year 11 cohort). There is therefore a chance that the cohort's scores are not typical cases and that some groups of students would have made more or less progress than others irrespective of their schools' influence (DfE, 2020). To account for this uncertainty and the fact that this risk is elevated when schools have a small Year 11 cohort, the DfE specifies a range of values within which the school's true effectiveness rating is assumed to lie. In theory the use confidence intervals therefore helps to prevent schools from being unfairly judged (or unfairly credited) with variations that are not due to genuine differences in their effectiveness.

This practice is endorsed by the majority of educational effectiveness researchers as being an appropriate method of communicating potential for error within value-added calculations (including, theoretically, the threats to validity discussed within this thesis) (see, for example, Goldstein & Spiegelhalter, 1996; Leckie and Goldstein, 2011; Mortimore, *et al.*, 1994; Nuttall, *et al.*, 1989; Sammons, *et al.*, 1995; Sammons, *et al.*, 1993; Wilson *et al.*, 2008). In fact, Reynolds *et al.*, (2012) explicitly state that they advocated for their use, through their academic writing and in reports to government. Another body of researchers however have voiced their concerns that this form or probability-based statistic is often abused (see, for example, Glass, 2014; Howe, 2014; Trafimow and Rice, 2009; White, 2014), and that they provide an ineffective means of summarising the threats to value-added analyses (Gorard, 2014; Perry, 2016b). It is therefore worth unpacking the issue. Particularly since the logic underpinning the approach is scarcely made explicit (Styles, 2014).

The theoretical justification for equating uncertainty and sampling error

As the preceding sections of this thesis have made clear, one cannot evaluate school effectiveness by comparing the raw attainment of students (Goldstein and Leckie, 2008; Goldstein and Thomas, 1996). Each institution has a unique intake that is more or less likely to succeed, irrespective of their school's influence (Goldstein and Woodhouse, 2000). Whilst Progress 8 attempts to control for these differences by accounting for the most potent source of bias, students' prior-attainment (Burgess and Thompson, 2013a), it is widely acknowledged that this form of statistical control is imperfect and that value-added calculations can, at best, be considered as an estimate of schools' true contribution (Fitz-

Gibbon, 1997). Some attention must therefore be given to the sources of error and/or bias that were not taken into account.

The conventional approach to estimating uncertainty thereby assumes that the remaining sources of inaccuracy stem from the allocation of pupils into school, and in doing so equates the potential for measurement error with sampling error. The argument is then made that confidence intervals are the best method for weighing uncertainty (Leckie and Goldstein, 2011; Wilson et al., 2008) because they purport to quantify sampling error (Neale, 2015).

However, several criticisms have been levelled at this form of probability statistic. Critics have argued that they are often used in inappropriate contexts, such as with the NPD population data (e.g. DfE value-added assessments), that the potential for measurement error cannot be reduced to technical issue and that their underlying logic is flawed (Gorard, 2015). These matters will be dealt with in turn.

In his article 'The widespread misuse of statistics', Gorard (2014) argued that it does not make sense to calculate confidence intervals when one is working with population data. Since no sampling has taken place, there is no need to test whether the information can be generalised.

Whilst many agreed with this assertion (for example, White, 2014), the practice is often defended by arguing that the figures apply to a hypothetical super-population, (for example the population of students that could have attended the school or the population of students that might attend the school in future) and/or that the process can be used to quantify how much variation could be expected to occur by chance if the sample had been selected at random (Glass, 2014; Styles, 2014). This defence seeks to draw a distinction between traditional design-based inference and model-based inference (Goldstein and Noden, 2004; Plewis and Fielding, 2003; Reynolds *et al.*, 2012; Snijders and Bosker, 2011). While design-based statistics are used to make inferences from a sample to a real-world population, model-based inference is concerned with broader questions (Plewis and Fielding, 2003). Specifically, the intent is to learn about the processes that produce the observed outcomes (Snijders and Bosker, 2011). The latter does not, proponents argue, require the randomisation to have actually taken place (Reynolds, *et al.*, 2012). Instead the analysis attempts model the degree of variation that would have occurred, if students were allocated to schools at random (Goldstein and Noden, 2004). The derivatives of probability theory (i.e. confidence intervals, significance tests, p-values, standard errors) are then used as a basis for comparing the amount of between-school variation with that which would have occurred naturally (Styles, 2014).

From this point opinions diverge. Some researchers see this as licence to apply the same logic outside the context of model specification (see, for example, Plewis and Fielding, 2003; Snijders and Bosker, 2011). Others such as Perry (2016b) and Glass (2014) disagree. Instead they argue that there is a sharp distinction between using inferential statistics as a yardstick against which to judge whether something could have emerged by chance and claiming that it did emerge by chance. The latter does not, they argue, permit inference to practical situations because the process forces the analysis to make assumptions that are not supported by evidence (Berk and Freedman, 2003). Say for example, that confidence intervals were used to suggest that a school's mean attainment level was sufficiently below the national average for us to be reasonably assured that the schools' results are unlikely to have arisen by chance alone. This does not, Perry (2016b) argues, help us to make a meaningful judgement about the school's effectiveness. The institution's results could be low because there are sources of systematic bias within the assessment or because it is genuinely ineffective. Inferential statistics cannot help one make this distinction. In fact, in almost all circumstances, the schools' score is likely to contain a random-error, systematic-error and genuine effect components (see, Coe and Fitz-Gibbon (1998)). The same logic also prevents model-based inferential statistics from being used to generalise observations to other context, such as pupils that would attend the school in future. The process requires an inductive scientific argument. This latter group of researchers therefore argues that confidence intervals have a narrow yet legitimate function, but that they should only be used as an aid to model specification (i.e. to help select the most appropriate variables to be represented within a model) (Perry, 2016b).

Gorard (2014) accepts these denunciations but adopts an even sterner position. In fact, he goes as far as to argue that in addition to not being able to provide the type of information that researcher's desire, the underlying logic of confidence intervals is flawed. This, he claims, makes the inferential statistics essentially meaningless. Specifically he asserts that the calculation wrongly equates the conditional properties for the population (or super-population) with those of the sample (or population). This rational of his position is reviewed below.

All probability statistics are underpinned by the same propositional logic:

> If A is true, then B is also true
>
> B is not true
>
> Therefore, A is not true either

This form of argument is known as *modus tollendo tollens* which means to deny the consequent. The premise begins with a conditional statement, 'that if event A occurs, event B will also occur'. This is then followed by a second statement, 'that the consequence of the first statement (event B) did not occur'. By interpreting the two statements together one can therefore deduce that, if B is always the consequence of A, and B has not occurred, then A must not have occurred either.

In the context of value-added assessments, the two events are therefore as follows:

> A = the population mean is a certain distance (or within a certain distance) of the sample mean
>
> B = the sample mean is the same distance (or within the same distance) of the population mean
>
> Where the sample mean refers to a school's official value-added rating and the population mean to the true value-added score that would emerge if an infinite number of students had attended the school. Though these two events could equally refer to the population data and super-population data respectively (see discussion above for a discussion of super-populations). Hence forth, however, these datasets will be referred to as the sample and the population so as not to confuse the discussion.

Here, one can see that if A is true, B must be true, and vice versa. This is a valid form of argument but one that is depends on both statements being 100% accurate (Gorard, 2014). If any uncertainty or inaccuracy is added to the statements the logic of the argument breaks down. Not least because, contrary to what one might assume, the probability of A given B will not necessarily be the same, or even close to the probability of B given A (see, Trafimow and Rice, 2009). This, Gorard states, is the problem that hinders all probability-based statistics.

According to the DfE guidelines their confidence intervals are intended to estimate "the range of values within which we are statistically confident that the true value of the Progress 8 score for the school lies" (DfE, 2020, p.5, ln. 2-3). However, since the analyst is unaware of the population parameters (i.e. the schools' true value-added score and standard deviation of pupils' true scores nationally) they instead use their best estimates of these values, the sample parameters (i.e. the school's measured Progress 8 result and the standard deviation of pupils' reported value-added scores). Gorard's (2014) assertion is that this logic is fundamentally flawed as it assumes from the outset what the calculation is intended to test, the accuracy of the sample statistics. Therefore whilst educational researchers such as Goldstein (2008) assert that confidence intervals should report what the DfE wishes to know. In Gorard's view what they report is the following:

> "If we assume that [the measured Progress 8 score] from a complete random sample is identical to the true [value-added by the school], then the CIs [confidence intervals] of many repeat random samples of the same size would contain the [true value-added score] for 95 per cent (or selected interval) of these samples"
>
> (Statement adapted from Gorard, 2014, p.7, ln.22-28)

This is why all confidence intervals centre on the sample mean (i.e. the reported score) as opposed to the population mean (the true score). What confidence intervals actually provide, Gorard (2014) argues, is therefore meaningless for any practical purpose. It cannot be used to report a range of likely values for the school's true value-added score as the logic of the argument is dependent upon the sample statistic being accurate. Moreover, if one allows for the fact the estimated Progress 8 score may not be identical to the school's true score, then it is no longer true that 95% of the projected scores will fall within the projected range.

The proposed shortfall and the calculations over-dependence upon the sampling statistics are best illustrated using an example.

Imagine that a school receives a Progress 8 rating of 0.5. This rating suggests that the school is performing above average. However we wish to calculate a confidence interval to determine whether this inference is justified or whether random error could conceivably account for the result. Now suppose that the school's actual rating was 1.0. In this scenario the schools confidence interval would project a range of conceivable scores around the observed score of 0.5 which may or may not include the true rating of 1.0. However, what happens if the schools' true value-added score was 2.0, or -2.0 for that matter? In both instances the same confidence interval would be projected because the observed rating and the standard deviation of the observed scores have remained unchanged. The formula does not (and cannot) take into account any differences in schools' actual performance, only the data that is observed. What is more, the scenario above implies that if a school with an actual performance rating of 0.5 is given a rating of 1.0, this is just as accurate as reporting their value-added as 2.0 or -2.0.

Gorard (2014) therefore asserts that, even when it is used as intended, the confidence interval formula does not work. In fact, embracing the flawed statistic, he argues, increases the risk or spurious conclusions and distracts data users from considering the type of error and/or bias that might have impacted upon the result.

Concluding statement on the use of confidence intervals:

This chapter has discussed the fragility of value-added evidence and how seemingly innocuous differences in researchers' prior assumptions can have a substantive impact upon their interpretation of value-added evidence. A particularly divisive topic is measurement error, where researcher have viewed inferential statistics both as being essential in conveying that value-added calculations are imprecise and prone to error (Leckie and Goldstein, 2011) and as meaningless and potentially damaging practice (Gorard, 2015).

It is argued here that the latter stance is correct. Not only are confidence intervals ill-suited to reporting upon the threats to validity that have been discussed within this thesis (see Chapter 4), their underlying logic is unsound (Gorard, 2014). They do not therefore provide a meaningful measure of either statistical or practical significance (Trafimow and Rice, 2009). What is more, their use encourages data-users to focus upon random error, thereby distracting from more pressing concerns (Perry, 2016b), whilst encouraging a level of uncritical acceptance and over-confidence in value-added results (Perry, 2016a).

The simplest solution would therefore be to stop reporting confidence intervals alongside schools' Progress 8 results and within other forms of value-added analyses. Informed data users would then be encouraged to consider the validity of any inferences more thoroughly. As to the claim that the general public would be more likely to accept the results at face value, we are sceptical of how much substantive difference this would make. Is it not more likely that parents and/or teachers that do not understand the intricacies of confidence intervals would more or less ignore them in any case?

7. Educational Effectiveness Research and the modelling of school performance

7.1 Chapter Introduction

This chapter underpins the empirical sections by identifying the main factors that impact upon student attainment. Particular attention is paid to the relationships that are outlined in the Dynamic Model of Educational Effectiveness (Creemers and Kyriakidies, 2008) as this informed our research methodology and instruments.

7.2 Disciplinary Perspective on Educational Effectiveness Research

Researchers that have sought to identify factors that distinguish between more and less effective schools have traditionally approached the problem from one of three disciplinary perspectives. They implemented an economic rational that focused primarily upon the interaction between purchased inputs and school outputs, they took a sociological stance that observed the association between students' background, prior-attainment and academic performance or they focused upon the learning of individual pupils and extrapolated from there. These three starting points led researchers to focus upon different types of relationships and thus three distinct categories of model emerged.

7.2.1. The economic perspective

Economic models of effectiveness were concerned with the productivity of schools (e.g. Elberts and Stone, 1988; Brown and Saks, 1986). That is to say, the efficiency with which purchased inputs such as teaching materials and staff salaries were converted into specified outputs. Their ultimate goal was to produce a mathematical function that described the association between any financial or material outlays and schools' performance once differences in their intakes had been taken into account (Monk, 1992). The more profitable the conversion, the more effective the school was deemed to be. These relationships could be represented as linear functions, consisting of main effects and interaction coefficients, or they could be non-linear (Brown and Saks, 1986). In either case, these models were characterised by their use of inanimate input variables, the examination of direct effects and the aggregation of all data to a single level of analysis.

It is easy to see the appeal of this approach as the kinds of factors they operationalised were easy to manipulate, especially from the perspective of the administrators who govern educational policy. A strict implementation of the economic principles was however problematic as both the inputs and outputs of education needed to be quantified. It was vitally important, for example, that all independent variables could be expressed in monetary terms. The same was true of the dependent variable with the added complication of having to codify students' learning. Furthermore, whilst the use of specific practices such as the choice of instructional behaviours, curriculum decisions and the schools' organisational structure could be evaluated, the assessments told us little about the mechanisms underpinning their effects (Cheng, 1993). This would not have been obstructive, except for the fact that the relationship between fiscal inputs and school outcomes is more complex than was once presumed. Increases in funding or the improvement of teacher-to-student ratios, for example, have inconsistent effects that do not necessarily lead to improvements in school performance

(Hanushek, 1986, 1989; Hedges *et al.*, 1994). The contribution that these models made to our understanding of effectiveness mechanisms was therefore limited.

Whilst the type of input variables that these models utilised are seldom seen within modern effectiveness research, the logistical approach and overriding concern for productivity have endured.

7.2.2. The sociological perspective

The sociological arm of effectiveness research made substantial contributions to our understanding of school effects.

Even before the field of effectiveness research was founded it was well-established that sociological differences between students such as their gender, socio-economic status and ethnicity impact upon an individual's access to and utilisation of educational opportunities (Coleman *et al.*, 1966; Jencks *et al.*, 1972). The first contribution of sociological studies was therefore to evaluate these influences.

This leads us to the second contribution which concerns the criterion used to define school or teacher effectiveness. This thesis is primarily concerned with the overall quality of schools' provisions. That is to say, with the impact that schools have upon the average or typical student and whether this is reflected in schools' performance ratings. The comments above should, however, make it clear that it is also possible to assess school effects in relation to the achievement gap that exists between particular types of pupil. Or to put it another way, whether the school helps to compensate or reinforce existing disparities. This constitutes a distinctly different perception of effectiveness that concerns the equity of educational opportunities rather than their quality. Whilst the equity of education may not be our primary focus, this branch of educational effectiveness research highlights that educational practices may not be equally beneficial to all types of students. This has implications for effectiveness measure such as Progress 8 (see Section 5.3.3). It also implies that evaluations of school effectiveness must consider the extent to which schools' policies and practices are differentiated in order to address different types of learning need.

Arguably, the greatest contribution of sociological studies however was to help connect classroom instruction with organisational and environmental factors that impact upon these interactions. This was an immense contribution as education does not take place within a vacuum. Instruction takes place in classrooms, classrooms are located in schools, and schools are located within wider educational structures and the wider contextual environment (Creemers, 1994). In modern-day effectiveness models it is therefore common to see variables that describe the school climate, the culture and/or structure of the school. The inclusion of these variables was inspired by organisational theories, such as Thompson (1967) and Mintzberg, (1979) which often adhere to the notion that there are many ways of conceptualising effectiveness (Cameron and Whetten, 1983). Evaluations can consider the productivity of the organisation, their adaptability, the involvement of its members, the continuity of the working environment, or its responsiveness to external stakeholders (Scheerens and Creemers, 1989). Though, in the contextual of educational effectiveness, all can be seen as pre-requisites for enhancing educational attainment (Scheerens, 1992).

7.2.3. The psychological perspective

Whilst the other perspectives of effectiveness research initially concerned themselves with factors that manifest at the school-level, a significant body of educational research had already investigated the influence of classroom-level interactions. Specifically the discipline of teacher effectiveness research (Teddlie and Reynolds, 2000).

Though interest in what makes some teachers differentially effective can be traced back as far as the 1920s (Domas and Tiedman, 1950), the origins of this research paradigm are primarily associated with the formation of the AERA Committee on Criteria of Teacher Effectiveness (Barr *et al.,* 1952; 1953) and the subsequent publication by Gage (1963). These events revitalised researchers' interest in teacher effectiveness and emphasised that teaching activities should be related to learning gains (Doyle, 1977). A multitude of studies followed, that were then summarised by researchers. Rosenshine (1976; 1983), for example, used this body of literature to develop the Direct Instructional Model of teaching which stressed the importance of six instruction practices or 'functions'. These were; the reviewing and checking of the previous day's work, the presentation of new content, student practice/guided practice, feedback and correctives, independent practice and periodic reviews of covered material. Brophy and Good (1986) developed a similar model known as Active Instruction. This recommends a comparable set of behaviours but places a greater emphasis of the teachers' ability to protect instructional time using a combination of forward planning, thorough explanations of learning activities and multi-tasking.

Methodological advances, however, have now allowed components from the three historical research traditions to be combined within multi-level frameworks (see, for example, Stringfield and Slavin, 1992; Creemers, 1994). These 'integrated' models use an economic style production function to depict the impact that organisational and process variables have upon students' performance after the differences in students' prior-attainment and background have been taken into account. Most however are connected by learning theories from the psychological research tradition in recognition of the fact that classroom instruction is the locus of the educational process (Scheerens, 1992).

7.3. The Dynamic Model of Educational Effectiveness

The Dynamic Model (Creemers and Kyriakides, 2008) is a multi-level model of effectiveness that attempts to outline the most influential factors from the student-, classroom-, school- and context-level. It is comparable to other integrated models in most regards, including the fact that it places most emphasis on the teaching and learning situation, and thus upon the role of teachers and students. School and contextual factors may have a direct impact upon student performance but are expected to operate primarily by influencing the conditions at lower levels. These attributes are depicted in Figure 7.3a below.

Figure 7.3a: The multi-level structure of the Dynamic Model

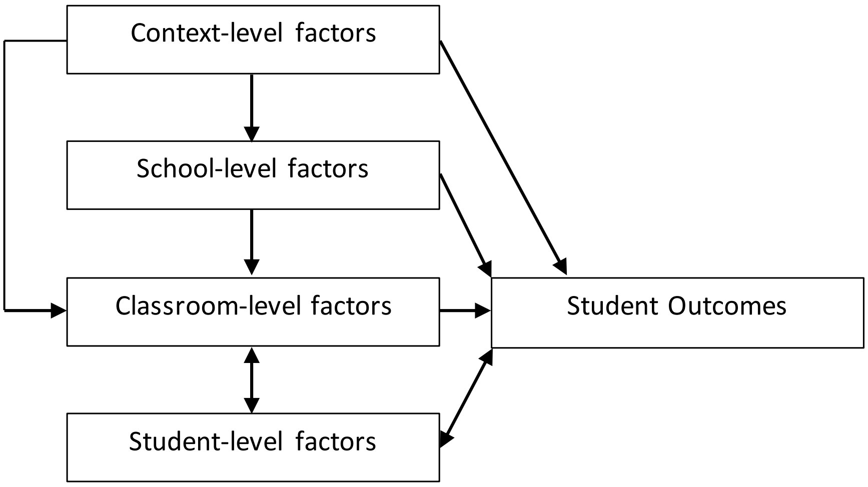


*Image replicates Figure 5.1 from Creemers and Kyriakides (2008, p.77)

The model deviates from other models in three main regards. Firstly, it recognises that effectiveness factors are not unidimensional constructs that have only one important variant. The authors therefore specify that all actions should be evaluated from five different perspectives; the frequency with which they are performed, their focus (specificity and purpose), the timing and duration of their implementation, their quality and the level of differentiation that takes place. Each of these dimensions is thought to play an important role in understanding the different effects that educational practices can have upon student performance.

This brings us to the second distinction. The Dynamic Model acknowledges that the relationship between effectiveness factors and student attainment will not necessarily be linear. It is presumed, for example, that if a teaching activity is performed too regularly it may start to have a lesser or even detrimental effect. Likewise, whilst school policies need to be specific in order impact upon teacher behaviour, too specific a policy may be restrictive. And whilst actions may be more likely to achieve a single objective, their effect may be limited if their influence is too narrow and disconnected from other educational activities. These non-linear effects, however, are limited to the frequency and focus dimensions of the model. More consistent implementation, higher quality actions and differentiation are all presumed to have a linear association with performance. Though it important that differentiation strategies are reviewed by the school's internal evaluation mechanisms as some approaches can be counter-productive (see, for example, Kyriakides 2004; Peterson *et al.*, 1984).

The final distinguishing feature gives the model its 'dynamic' character. Creemers and Kyriakides state that school- and contextual-level factors need to be evaluated in a different way to classroom-level influences. Specifically they argue that the effect of new policies is dependent upon the school's situation. That is to say, on whether the guidelines refer to one of the stronger areas of the schools' provisions or a weakness. The logic being that it is significantly harder to improve upon more proficient areas. For this reason, the pair recommends that the stage dimension of any evaluation considers whether changes to the school policies were based upon data from the school's internal evaluations. This should ensure that worrisome areas are addressed and that there is continuity in the schools' approach.

7.3.1. The theoretical basis of the Dynamic Model

The core constructs from the Dynamic Model were taken from Carrol's model (1963) of school learning.

This model states that the degree to which a student masters an activity is determined by the ratio between the time they have dedicated to the task and the time they would need to master it. Time actually spent on learning is purported to be the lowest of three values; the time allowed for learning (i.e. opportunity), the time for which the learner is prepared to engage actively with the learning activity (i.e. perseverance) and the time that the student needs to master the task under optimum conditions (which is determined by the students' aptitude). This last value may be increased by less than optimal tuition if this interferes with the learner's ability to understand instruction (i.e. the quality of instruction). In doing this Carrol essentially identified the main student- and classroom-level factors that influence students' learning and provided a theoretical explanation of how they interact with each other and student outcomes.

Carrol's definitions of these factors were, however, rather vague. In order to develop this instructional theory into a multi-level model of educational effectiveness, Creemers and Kyriakides (2008) elaborated upon these descriptions by further distinguishing between the components. They began by dissecting the construct of *opportunity to learn*. In the original model this referred to the time that was available for learning. The pair recognized, however, time alone cannot induce learning. In order to achieve educational outcomes students must also be given the chance to acquire the requisite knowledge and skills. In other words, they require access to the content they are expected to learn. Their model therefore includes factors concerned with the time (*time for learning*) and content (*opportunity to learn*) that is made available to students. Whilst this was a significant step in the development of the model, it is also a potential source of confusion. It is therefore reiterated that within the Dynamic Model of Educational Effectiveness the term *opportunity to learn* refers to content not time.

Creemers also acknowledged that there is a difference between allocated learning time and opportunities, and those utilized by teachers and students. The Dynamic Model therefore distinguishes *time on task* (the time for which students are actively engaged in learning activities) from the time that is made available at the classroom, school and contextual levels. The same distinction was made between the provision and use of opportunities to learn (i.e. content). The first two constructs, however, are not measured directly, rather they were used alongside the concept on instructional quality as criterion for defining and evaluating effectiveness factors at higher levels. That is to say, to select and categorise the teaching behaviours and policies that help to enhance the time, learning opportunities and quality of the educational experiences that are made available to teachers and students. [14]

[14] This explanation has been truncated. These constructs were originally expanded within Creemers' Comprehensive Model (1994), the forerunner to the Dynamic Model that is utilised within this thesis. The revised version, however, makes the same methodological assumptions as the original.

7.3.2. Student-level factors

All of the effectiveness factors within the Dynamic Model are presumed to have a positive association with school performance[15]. This applies to all factors from all levels of the model, though the caveat about the plausibility of non-linear effects at the classroom, school and context levels still applies. At the student-level, however, 11 influences are identified.

The first two factors, *time on task* and *opportunity to learn* were introduced above. These are defined as the time that students' are actively engages in learning and the content that students actually engage with, respectively. They are referred to as 'task-related' variables to signify that they describe the actions of the individual learners rather than their background or personal characteristics. Teachers and students can therefore exercise a degree of control over these variables, though their actions are framed by decisions that are made at higher-levels.

The remaining factors refer to differences in students' background and characteristics that predispose them to making more or less progress than other students. The most important of these is the student's aptitude, which reflects both students' intelligence and prior-attainment. These differences are the primary determinant of academic success. In fact, they often account for more than 50% of the variation is students' raw attainment (Teddlie and Reynolds, 2000; Thomas, 2001). The popular theory being that one's natural ability and prior-learning regulate the speed at which one can process new information (Carrol, 1963). More capable individuals with pre-requisite knowledge of the topic are therefore presumed to master learning activities in a far shorter period because they approach the task from a favourable starting point. Support of this has been provided by studies that have investigated the validity of the Comprehensive Model (the precursor to the Dynamic Model) and other integrated models of educational effectiveness (e.g. de Jong *et al.*, 2004; Kyriakides, 2005).

The next three variables; students' socio-economic status, ethnicity and gender were taken from the sociological branch of effectiveness research. Many studies have shown that the greater part the variation in students' educational outcomes can be explained by this type of background characteristic (Sirin, 2005). Moreover, the initial differences between students tend to expand during the educational process (Kyriakides 2004a) and the capacity of schools to address these inequality is limited (Jesson and Gray 1991; Thomas *et al.* 1997a). Two points, however, should be noted. The first is that the precise nature of these influences is unknown. One would imagine that the inequality in educational outcomes occurs because of underlying differences in the educational opportunities that are available or utilised by the respective groups, though, there are a multitude of reasons why this could be the case. Middle-class parents, for example, may place a higher value upon education than working-class parents, have better access to learning resources or be more likely to pay for additional tutoring. It may even be that these individuals find it easier to help their child with school-work by virtue of their own educational experiences or that there is less pressure to start working and contribute to the household income. Most effectiveness studies have not attempted to delineate these effects and neither does the Dynamic Model. A cynic would therefore argue that these variables act as proxies for the underlying but unconfirmed mechanisms. Researchers have, however, long abandoned the notion that the higher performing groups are naturally more intelligent and assume that all (or the vast majority) of students would achieve success under the right conditions (Bloom, 1968). The second point is that the variance that can be explained by these factors will overlap with the variance that is predicted by prior-

[15] The relationship between ethnicity and performance is more complex than this statement suggests with some groups being more or less disadvantaged than White-British pupils (Thomas *et al.*, 1997a). In the UK, however, most groups tend to outperform the White-British group. See Section 5.3.3 for further details. The relationship between gender and performance also varies from subject to subject but in within secondary education girls overall performance tends to exceed boys (Thomas et al., 1997a; Leckie and Goldstein, 2019).

attainment (Fitz-Gibbon, 1997). Essentially, this is because any pre-existing measures of students' learning will encompass the impact that any extraneous sources of bias have had up until that point. When prior attainment has been controlled, these factors therefore account for a lesser but far from insignificant proportion of the variation in student performance.

Similar attention is also given to differences in students' personality, thinking style, perseverance (stable context-irrelevant motivation), subject-motivation (subject-specific motivation) and expectations. These factors also have consistent associations with student performance that makes some students more likely to succeed in an academic context (de Jong *et al.*, 2004; Kyriakides *et al.*, 2000; Kyriakides, 2005; Kyriakides and Charalambous, 2005; Kyriakides and Tsangaridou, 2008; Valverde and Schmidt, 2000; Wentzel and Wigfield, 1998). The concepts, however, were taken from psychological research as opposed to the sociological branch of effectiveness research.

Since the collective impact of the aforementioned variables is far greater than that of teachers or schools (de Jong *et al.*, 2004; Kyriakides 2005). Creemers and Kyriakides also distinguish between the variables which can and cannot be influenced by schools. Students' expectations, subject motivation, thinking style and engagement level (time on task and opportunity to learn) are all considered to be malleable and responsive to students' educational experiences. Their aptitude, socio-economic status, gender, personality traits and perseverance however are considered to be stable, at least in the short term. Whilst the former can be treated as educational outcomes, or as a means of improving student attainment, the latter must be controlled within any value-added assessment of school effectiveness, or the bias that they describe will impact upon schools' ratings.

7.3.3. Classroom-level factors

Factors at this level provide the conditions for students' learning. Whilst each of these variables may have a direct impact upon student performance, their main influence is upon the amount of time and learning opportunities that students engage with. They may also effect students' expectations, thinking style and subject motivation.

In recognition of the important role that teachers play in the instructional process, each classroom-level factor is defined as a teaching behaviour[16]. Eight are used to describe the teacher's influence upon students' learning. These include; orientation, the structuring of lesson content, questioning, teacher-modelling, application, the teacher's role in creating an effective classroom learning environment (which is sub-divided into three components; teacher-student interaction, student-student interactions and classroom disruptions), the management of lesson time and classroom assessment. All of these behaviours were taken from teacher effectiveness research and have an established relationship with performance (see, for example, Brophy and Good, 1986; Darling-Hammond, 2000; Dunne and Wragg, 1994; Kyriakides, Campbell and Christofidou, 2002; Muijs and Reynolds, 2000; Rosenshine and Stevens, 1986; Wang, Haertel and Walberg, 1993).

The factors, however, are not based upon a particular approach of teaching. Instead Creemers and Kyriakides adopt what they refer to an "integrated approach" to defining teaching quality (2008, p. 103, ln 27). This means that the 8 classroom-level factors were designed to cover the key aspects of

[16] That is to say, that all factors refer to the quality of teachers' instructional behaviours. Available learning time is assessed as part the evaluation of the classroom learning environment and management of time factors. Opportunity to learn, however, is not referred to explicitly at this level as it was felt that in most cases the curricular that students follow and the textbooks that students use would be set at the school or departmental level.

both traditional and modern teaching theories. Whilst the structuring and questioning factors may play a central role in, for example, the direct teaching model (Rosenshine, 1983) or mastery learning (Bloom, 1976), orientation and teacher modelling are more important in approaches that attempt to improve the learning disposition of students (e.g. Choi and Hannafin, 1995; Collins, Brown and Newman, 1989; Savery and Duffy, 1995; Simons, Linden and Duffy, 2000). Creemers (2007) reviewed these approaches and demonstrated that they were sufficiently covered the proposed behaviours.

With the exception of classroom disruptions and off-task interactions which distract students' from learning, there is considerable research evidence to suggest that each of these behaviours will have a positive interaction with performance, meaning that increasing or improving upon their use will enhance the learning of students. It should be remembered however that the overuse of any one technique may have detrimental effects and that Creemers and Kyriakides intended for all of the aforementioned behaviours to be evaluated using the 5 dimensions set out in introduction to this section.

7.3.4. School-level factors

Creemers and Kyriakides' conception of the school level is based upon an assumption that schools will influence students' learning in a different manner to teachers. Whilst teachers are directly involved in the delivery of instruction, school-level factors, for the most part, are not. Instead these influences are thought to affect student performance by shaping teachers' behaviour and the conditions under which classroom instruction is delivered. They are therefore modelled as having a predominantly indirect influence upon student achievement. Some direct effects on attainment are thought to exist but it is believed that these are rarer and their influence less substantial. These suppositions are collaborated by the findings of educational effectiveness research which have repeatedly demonstrated that factors at the classroom level explain a higher proportion of the variance in student attainment than school- or context-level factors (see for example, Kyriakides, Campbell and Gagatsis, 2000; Yair, 2000; Teddlie and Reynolds, 2000). For this reason, the Dynamic Model refers to school-level factors that have a clear empirical and theoretical link with both classroom instruction and student attainment. These factors are also grouped according to their envisaged impact upon instruction. This means that the three core constructs of learning time, learning opportunities and the quality of teaching, which were emphasised at the classroom and student levels, play a central role in the definition of the school-level variables. It is also worth noting that, since the primary aim of educational effectiveness research is to identify ways for education providers to enhance student achievement (Creemers 2002), the school-level factors are defined as school policies and/or actions. This means that issues such as the students' behaviour outside of lessons are not assessed by monitoring students' interactions but by the extent to which differences in content of the school behaviour policies are associated with differences in student attainment.

Specifically, four aspects of the school policies are considered:

The first over-arching factor refers to the school teaching policies. These guidelines contain a set of rules and agreements that help to regulate classroom-level instruction by directly influencing the time that is made available for learning, the content that students are exposed to and the instructional behaviour of teachers during lessons. There is therefore a clear theoretical and empirical link between these regulations, classroom behaviours and student performance.

The second overarching factor identified by the Dynamic Model refers to the policies for creating an effective learning environment at the school. This aspect of school procedures can be broken down into several sub-factors, namely; the policies govern student behaviour outside of classrooms, teacher collaboration, the provision of learning resources, teachers' and students' attitude towards learning, and the school partnership policy. All describe school-level measures which can be taken to promote favourable forms of interaction between school stakeholders outside of lessons. Like the other school-level factors in the model, these policies are intended to improve pupil attainment by enhancing the characteristics of classroom-level instruction. That is to say that they are intended to improve teachers' use of the 8 instructional behaviours and thus the quantity of active instruction and learning opportunities that are made available to students. Whilst the link between these policies and specific classroom- or student-level factors is therefore less tangible, their effect upon student performance is no less significant.

The two remaining over-arching factors are concerned with the school policies for monitoring and evaluating its educational provisions. A substantial body of research evidence both from the early stages of school effectiveness research and more recent multi-level studies suggests that a school's evaluation procedures will have an independent influence on students' performance (e.g. de Jong *et al.*, 2004; Harris, 2001; Kyriakides *et al.* 2000; Kyriakides, 2005; Thomas, 2001; Torres and Preskill, 2001). To be most effective, however, schools must continually appraise all aspects of their internal environment, not just student attainment levels. The dynamic model therefore includes factors that refer to the school policies for evaluating the institution's teaching policies and the policy for creating an effective learning environment at the school.

The four overarching school-level factors represented in the Dynamic Model are therefore; the school policy for teaching (including any actions taken to improve classroom instruction), the evaluation of the school teaching policy, the school policy for creating a learning environment at the school (and actions to improve school learning environment) and the evaluation of the school learning environment.

More extensive guidelines on the intended operationalisation of these factors are available in Creemers and Kyriakides (2008). The important facts to recall, however, are that all four over-arching factors are expected to have a positive association with school performance, though too great an emphasis on one aspect of the schools' provisions may be counterproductive.

7.3.5. Context-level factors

Even further away from the classroom-level there are contextual factors that establish the conditions that schools must operate within. Before identifying these factors, a few words should be said about the type of variable that Creemers and Kyriakides considered. In order to select variables that had a clear and discernible influence upon students' learning the pair operationalised factors that had a theoretical connection with classroom instruction, or, more specifically, the core concepts of *time on task*, *opportunity to learn* and *quality of instruction* that were used to identify the key educational factors at other levels. The use of this selection criterion had two effects. First, it excluded some of the operational characteristics that are commonly used to distinguish between the educational systems of different countries. One of the most significant areas of educational research that this dismisses is the consideration of how the structure of national educational systems impacts upon student attainment levels. Whilst such characteristics undoubtedly influence the delivery of educational provisions their overall impact upon performance is unclear. Several international studies and meta-analyses have therefore concluded that the effectiveness of national education systems is not determined by whether they are, for example, centralised or decentralised (Kyriakides and Charalambous, 2005; Schmidt *et al.*, 1998). The omission of such considerations was therefore deliberate. A secondary consequence of this selection criterion was that it allowed the contextual factors to be grouped in a comparable way to the lower-level constructs. The reader should therefore be acquainted with many of the concepts that are expressed and the nature of their influence.

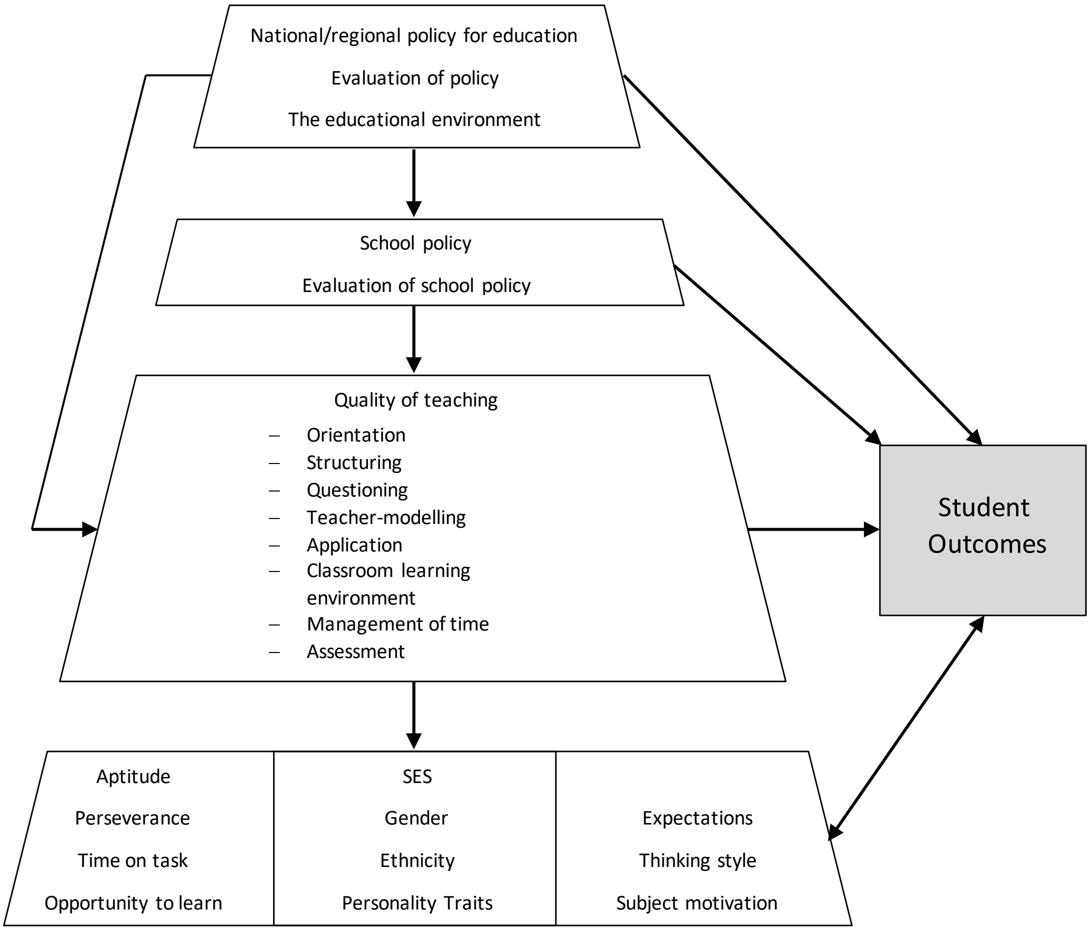
The first context-level factor refers to the national and regional policies on education. The model assumes that these policies will directly affect students' learning by influencing classroom instructional practices and stakeholders' learning outside of classrooms. There is therefore a close association between these policies and the school polices for teaching and the creation of a school learning environment, as the former provides the framework that the latter must operate within. In the U.K., however, middle-level organisations have a limited affect upon pupil attainment levels as local educational authorities are not in a position to directly influence policy (Tymms *et al.*, 2008).

The second factor is concerned with the mechanisms that are used to evaluate the aforementioned policy. As highlighted in the outline of the classroom- and school-levels, there is considerable evidence which suggests that feedback on the performance and behaviour of key educational stakeholders plays an important role in the development of effective practice. On the assumption that this principle will apply equally to the provision of national-level education, Creemers and Kyriakides chose to operationalise the concept as their second contextual-level factor.

The final context-level factor refers to the wider educational environment, specifically the support that that schools receive from local stakeholders and the expectations of these groups. This was Creemers and Kyriakides attempt to acknowledge that learning does not only take place within schools and that factors such as the national attitude towards education and the availability of learning opportunities outside of school can also enhance student attainment (Valverde and Schmidt, 2000).

The full models and the envisaged interaction between factors is outlined in Figure 7.3.5a.

Figure 7.3.5a: The Dynamic Model of Educational Effectiveness



*Replica of Figure 7.3 from Creemers and Kyriakides (2008)

7.3.6. A notable omission

Before concluding the section we should acknowledge why school leadership is not included as a factor within the Dynamic Model. This is because its effects are difficult to measure either directly (Witzier *et al.*, 2003) or indirectly (Leithwood and Jantzi, 2006). Creemers and Kyriakidies therefore focused upon the implementation of positive actions, rather than who is implementing them. This is in-line with way the factors are defined at other levels. It is nevertheless acknowledged that school leaders are instrumental in developing their school's mission, structure, policies, culture, resources and strategies for improvement (Leithwood *et al.*, 1998).

7.4. Empirical Support for the Dynamic Model

This section presents three studies that have tested the suppositions of the Dynamic Model; Kyriakides *et al.'s* (2013) meta-analysis of classroom-level effectiveness factors, Creemers and Kyriakides' (2008) meta-analysis of school-level effectiveness factors and Creemers and Kyriakides'

(2008) empirical appraisal of their measurement framework. This research provides support for the model and further explicates the relationships that exist between student attainment and the school-related variables.

7.4.1. The effect of classroom-level factors

Traditional literature reviews have often struggled to summarise the effectiveness literature because the effect of classroom- and school-level factors varies across studies. In fact, it is reasonably common for a practice to be described as beneficial in one assessment and unhelpful in another. Whilst this may appear perplexing, it is made more understandable when one recalls that the influence of school-related variables is presumed to be small and that the results of all analyses contain measurement error. Meta-analyses which integrate and summarise the results from multiple studies are therefore invaluable. Firstly, because they reveal the underlying relationships that exist between variables and secondly, because they can be used to identify moderators that impact upon the reported associations. The methodology can also be adapted to validate theoretical frameworks by determining whether the factors that are included in a model have a greater association with students' performance than other correlates.

Though several researchers have conducted meta-analyses of the teacher effectiveness literature (see, for example, Hattie (2009) and Seidel and Shavelson, (2007)), this discussion focuses upon the results of Kyriakides *et al.*, (2013). This assessment is more up-to-date, includes a reasonable number of studies (n=167) and refers directly to the factors outlined in the classroom-level of the Dynamic Model of Educational Effectiveness.

The core results of the analysis are reported in Table 7.4.1a.

Table 7.4.1a: The average effect size of classroom-level factors within Kyriakides *et al.* (2013) meta-analysis of teaching behaviours

Teaching behaviour	Average effect (z-score[17])	Number of studies
Orientation	0.36	14
Structuring	0.36	35
Questioning	0.34	12
Teacher-modelling	0.41	35
Application	0.18	27
Classroom learning environment	0.45	78
Management of time	0.35	30
Assessment	0.34	30

[17] Note Kyriakides' effect sizes refer to Fisher's Z transformation of the correlation coefficient. For small values of the correlation coefficient Z and r do not differ significantly. In this instance, for example, the r-scores for each factor would be; orientation = 0.35, structuring = 0.35, questioning = 0.33, teacher-modelling = 0.39, application = 0.18, CLE = 0.42, management of time = 0.34, assessment = 0.33. The r-squared scores of these values are therefore; orientation = 0.12, structuring = 0.12, questioning = 0.11, teacher-modelling = 0.11, application = 0.15, CLE = 0.18, management of time = 0.11, assessment = 0.11.

These effect sizes imply that the instructional behaviours from the Dynamic Model have a moderate level of association with student outcomes. Most exert a comparable level of influence (average effects of 0.34 to 0.36), with only the teachers' role in maintaining an effective classroom learning environment (average effect = 0.45) and teacher-modelling activities (average effect = 0.41) standing out as being of particular importance, and teachers' use of application tasks as less consequential (average effect = 0.18). This is supportive of the model's validity. As is the fact that seven of the eight factors had a greater influence than computer use (average effect = 0.20), interpersonal behaviour (average effect = 0.16) and classroom organisation (average effect = 0.05). These are teaching behaviours that have been discussed in the academic literature, but were excluded from the model.

In should be noted, though, that concept-mapping was recorded as having an effect size of 0.75 and self-regulation an effect of 0.47. On the surface, both findings suggest that meaningful influences are absent from the classroom level of the Dynamic Model. Kyriakides *et al.* (2013), however, assert that the former may have been a statistical artefact brought about by the low number studies that reported upon concept-mapping (n=3) and the type of studies reviewed (experimental designs – see latter discussions). What is more, the authors argue there is likely to have been an overlap between the concept of self-regulation and the problem-solving skills that are developed through teacher modelling. Neither finding is therefore viewed as presenting a serious challenge to the Dynamic Model.

In terms of moderating influences, the analysis confirmed that there were relatively large variations in effect sizes of classroom behaviours within and across studies[18]. For the most part, however, these defied explanation. Only in a few instances could the differences be explained by methodological or contextual factors and no moderator had a meaningful relationship with the effect size of all behaviours.

It was nevertheless apparent that teacher-modelling had a greater impact upon secondary school students (average effect size 0.22 greater than within primary education), whereas application tasks were more influential amongst younger pupils (average effect size 0.15 lower within secondary education). This may be because constructivist approaches, which use students' existing experience to develop new understanding, rely upon higher-order skills that take time to develop or because the curriculum of older students places greater emphasis upon these skills. Higher effect sizes were also present in the results of longitudinal (+0.12 in the case of the structuring variables and +0.11 in the case of classroom assessments), quasi-experimental (+0.19 in the case of teacher-modelling) and experimental studies (+0.12 in the case of teacher-modelling and +0.12 in the case of the CLE), than within cross-sectional studies (the control group for previously cited figures). This suggests that robust designs may be more proficient at detecting educational effects. The fact that the remaining moderators (the type of learning outcome utilised, the country in which the research was conducted and whether single or multi-level statistical techniques were employed) could not account for a substantial portion of the variance in scores, however, supports the supposition that the classroom level of the Dynamic Model refers to generic factors that are neither context or outputs specific.

7.4.2. The effect of school-level factors

Similar meta-analyses have been used to validate the school-level of the Dynamic Model. Creemers and Kyriakides (2008), for example, used the results from 67 effectiveness studies to examine the impact of school policies.

[18] The standard deviation of effect sizes was not reported.

Their investigations can be broken down into three parts. The analysis began by calculating the mean effect of each school-level effectiveness factor. These were then compared with the influence of factors that were purposefully excluded from the model (see Table 7.4.2a).

Table 7.4.2a: The average effect size of school-level factors within Creemers and Kyriakides (2008) meta-analysis of school-level policies

School-level factor	Average effect (z-score[19])	Number of studies
1. Policy on teaching		
(a) Quantity of teaching	0.16	18
(b) Opportunity to learn	0.15	13
(c) Quality of teaching	0.17	26
- assessment	0.18	12
2. Policy on the school learning environment		
(a) Collaboration	0.16	31
(b) Partnership policy	0.17	21
3. Evaluation of policy on teaching	0.13	6
4. Evaluation of policy on the school learning environment	-	0

These figures suggest that schools' policies have a modest impact upon student outcomes (average effect sizes = 0.13 to 0.18). This helps to validate the school level of the Dynamic Model as these influences were envisaged as having a small but meaningful effect.

The official school-level factors also exerted a greater influence than six of the school-level factors that were not included in the model. The importance of school leadership, for instance, is often cited in the literature but has little impact upon students' performance (average effect = 0.07). Nor do the resources, salary and working conditions of schools (average effect size = 0.14), the school climate (average effect = 0.12), job satisfaction (average effect = 0.09), the experience of school staff (average effect = 0.08) or teacher autonomy (average effect = 0.06). This implies that the factors highlighted in the Dynamic Model are of differential importance.

It should be noted, though, that there is less evidence to justify the status that is afforded to schools' evaluation policies. In fact, only six of the 67 studies analysed the impact of schools' evaluation mechanisms and all of these focused upon the evaluation of schools' teaching policies. Schools' procedures for evaluating the school learning environment were therefore included as effectiveness factor in the Dynamic Model because of their presumed influence. This decision needs to be validated in future research.

Moreover, the study suggests that teacher empowerment may have a comparable impact to the school-level factors outlined in the model (average effect size = 0.17). The influence of this factor, however,

[19] Creemers and Kryriakies' effect sizes also refer Fisher's Z transformation of the correlation coefficient. In this instance the Z and r scores of each factor are identical to 2dp. The r-squared scores of factors were therefore; quantity of teaching = 0.03, opportunity to learn = 0.02, quality of teaching = 0.03, assessment = 0.03, collaboration = 0.03, partnership policy = 0.03, evaluation of the school teaching policy = 0.02.

was calculated from the results of only two studies. So, whilst this outcome suggests that it might be possible to amend the school level of the model to provide a more comprehensive account of students' learning, further evidence is warranted before any adaptations are considered.

The next step in the analysis used multilevel modelling to evaluate the impact of three over-arching school-level factors; the school teaching policy, the school's policy on collaboration (SLE Component 1) and the school's partnership policy (SLE Component 2). The collective effect of schools' evaluation procedures was not considered because of the limited number of studies available.

The results of this assessment are reported in Table 7.4.2b.

Table 7.4.2b: The average effect sizes for each over-arching school-level effectiveness factor and the standard deviation of effect sizes across and within replications

Over-arching school-level factor	Mean effect size across replications (z-score[20])	Standard deviation within replications	Standard deviation
1. Policy on teaching	0.179	0.033	0.036
2. Policy on the school learning environment			
(a) Collaboration	0.158	0.043	0.040
(b) Partnership policy	0.172	0.032	0.031

*Figures cited from Creemers and Kyriakides (2008)

Two observations can be made from this table.

The first is that the mean effect size of the schools' teaching, collaboration and partnership policies reemphasises the importance of these procedures and justifies their inclusion in the Dynamic Model. Especially since their influence comfortably exceeded the impact of factors that were excluded from the model (see previous discussion).

The second observation is that the effect sizes of the over-arching factors deviated substantially, both within and across studies (see standard deviation values in Table 7.4.2b). This has important implications for any researcher that plans to use the model within their research, as it signifies that the effect sizes attributed to each factor (or over-arching factor) are unlikely to be replicated precisely within individual studies and that the rank-order of variables' influences may deviate.

The final section of the analysis attempted to explain this variation. More specifically, Creemers and Kyriakides (2008) codified the type of outcome variable that was utilised in each study (cognitive,

[20] These effect sizes refer Fisher's Z transformation of the correlation coefficient. For small values of the correlation coefficient Z and r do not differ significantly. In this instance the r scores for each over-arching factor are; policy on teaching = 0.177, collaboration = 0.157 and partnership = 0.170, and their r-squared scores are 0.031, 0.025 and 0.029 respectively.

affective, psychological), the educational level of the institutions (primary, secondary, tertiary), the country in which the research was conducted (USA, European countries, Asian countries, other), the study design (cross-sectional, longitudinal, quasi-experimental, experimental, outlier), the type of statistical techniques employed (single, multi-level), and the grouping of factors into over-arching factors (grouping, no grouping), and used each characteristic as the basis for predicting differences in the reported effects.

For the most part these variables were unhelpful in explaining the deviation in factor's influence. This suggests that the reported effect sizes were not unduly influenced by the context of studies or researchers' methodological decisions. The only exceptions were that the two components of schools' policy for establishing an effective learning environment, i.e. their partnership and collaboration policies, had a closer relationship with student outcomes in Asian countries (+0.05 and +0.04 relative to studies in the USA), whilst the effect attributed to schools' teaching policies was higher in longitudinal studies (0.02 higher, on average, than in cross-sectional studies), and the effect attributed to the schools' partnership policy was higher in experimental studies (0.03 higher, on average, than in cross-sectional studies). No moderator, however, was found to have a meaningful relationship with the effect size of all five overarching factors.

To summarise then, Creemers and Kyriakides' (2008) meta-analysis reported that school-level procedures have a modest association with student outcomes. Moreover, whilst the effect attributed individual policies varied substantially from study to study, the influence of contextual and methodological variables was minor. The bulk of results were therefore in-line with the authors' expectations and provide support for the major assumptions of the Dynamic Model.

It should be noted, though, that only one study found evidence of there being a non-linear association between a school-level effectiveness factor and student performance. What is more, the relationship in question described the influence of resources upon student attainment, an interaction that is downplayed within the Dynamic Model. This is a slight inconsistency. One possible explanation is that the majority of educational effectiveness research has utilised cross-sectional or longitudinal designs. These approaches are less adept to detecting non-linear relationships because they cannot guarantee that there will be sufficient variation within the independent variables (Creemers and Kyriakides, 2008). Future studies could therefore explore the issue using alternative methodological approaches, such as experiments.

It should be likewise be acknowledged that the studies reviewed in this meta-analysis were almost exclusively concerned with the frequency dimension of factors (almost 94.2% of studies only considered this aspect of policies). Whilst this bias reflects the features of current effectiveness literature rather than the study's inclusion criteria, it may have impacted upon the analysis' results. Creemers and Kyriakides (2008) therefore point out that the two studies that evaluated the stage dimension of factors elicited comparable figures. This suggests that the cited relationships will apply irrespective of which dimension of effectiveness is considered, but further evidence is warranted. The authors also note that the only study to investigate the consistency dimension of effectiveness, a dimension that was included in previous versions of the model and later excluded (see, the Comprehensive Model (Creemers, 1994)), did not uncover a meaningful relationship between the characteristic and student achievement (see Ressight *et al.*, 1999). This justified the decision to remove this dimension from the measurement framework.

7.4.3. The validity of the measurement framework

Creemers and Kyriakides (2008) also validated their measurement framework; first, by demonstrating that instructional practices are multidimensional constructs that can be measured in relation to five dimensions, and second, by demonstrating that there is added value in evaluating effectiveness from multiple perspectives.

Both investigations utilised a stratified sample of 50 Cypriot primary schools, 108 Year 5 classes and 2503 students.

Part 1: Testing the validity of the framework used to measure each effectiveness factor

Multi-trait multi-method (MTMM) matrices are a useful tool for evaluating construct validity. The approach factorally combines sets of traits and measurements, so that the variance attributable to traits, methods, and unique or error variance can be identified. In the first analysis the authors used this methodology to test the assertion that classroom-level effectiveness factors are multi-dimensional constructs that can be assessed by five interrelated, but conceptually distinct, dimensions of effectiveness.

Detailed information on teachers' instructional practices in maths, Greek language and religious education lessons was collected using four research instruments. Specifically, two types of low-inference observation, a high-inference observation and a student questionnaire[21]. 24 MTMM matrices were then created to depict the variation in teachers' scores for each classroom behaviour, in each subject area. Since each instrument was intended to evaluate the frequency, focus, timing, quality and differentiation of effective behaviours, in accordance with the measurement criteria set out within the Dynamic Model of Educational Effectiveness, the differences in teachers' scores for each behaviour should, in theory, be explained by five traits (i.e. the five dimensions). Method effects, that is to say, differences between the ratings of each instrument, may also occur but should be relatively small if the measurement instruments and framework are valid.

Confirmatory factor analysis (CFA) was used to assess whether this was the case. Six models were posited and their goodness-of-fit was evaluated[22]. More specifically, four first-order models and two second-order models. The first model (null model) was the most restrictive. It presumed that were no trait or method effects and that teachers' scores for a given classroom-level factor would therefore act as 20 uncorrelated variables. Model 2 contained five correlated traits and no methods. Model 3, five correlated traits and four correlated methods. Model 4, five correlated traits and three correlated methods. Model 5, contained one second-order general trait and three correlated methods. Whereas, Model 6, contained two correlated second-order general traits and three correlated methods. Once the best-fitting model had been determined, the amount of variance attributable to each trait and method effect was calculated by squaring their respective loadings.

The results provide support for construct validity of Creemers and Kyriakides' (2008) framework. The measures of most classroom-level behaviours (orientation, structuring, application, and assessment)

[21] The outputs from these instruments were standardised and independently validated before the analysis commenced.

[22] For those that are interested in the technical details, the analysis was conducted using the EQS program (Bentler, 1989) with maximum-likelihood estimation. Scaled chi-squared, Bentler's comparative fit index, the root mean squared error of approximation, the chi-squared to degrees of freedom ratio and the parameter estimates were used to assess models fit. The chi-squared difference test was used to evaluate the improvement in fit among hierarchically nested models.

were best explained by five factors or 'traits', which correspond to the five dimensions of effectiveness outlined in the Dynamic Model (i.e. the frequency, focus, timing, quality and differentiation of teacher behaviours). Each of these factors had a strong positive loading (>0.6), which demonstrates the model's convergent validity. Whereas, the correlations among factors were consistently positive but relatively low (<0.4). This can be interpreted as a sign of divergent validity. That is to say, that each dimension assesses a different aspect of teachers' behaviours.

The few exceptions that were identified reveal the difficulty of defining the quality dimension. This aspect of questioning tasks, for example, was separated into two parts; measures concerned with the quality of teachers' questions and measures that evaluated teachers' follow-up. Whereas, the measures that were intended to evaluate the quality and differentiation of teacher-modelling activities were grouped into one factor, suggesting that there was a lack of distinction between the two elements. The fact that the management of time variables were influenced by four factors, though, is not surprising as the focus dimension of this behaviour is not intended to be operationalised. The only consideration is whether students are or and not on-task. Similarly, the finding that the classroom learning environment was best described by two over-arching factors, i.e. teacher-student interactions and student relations, suggests that the model could be made more parsimonious in places, but does not contradict any of the underlying assertions.

Method effects were present in all analyses. In most cases, a three-factor model tended to account for more variance than a four-factor model, after the first-order trait groupings had been taken into account (see discussion above). This was the case for the orientation, structuring, teacher-modelling, application and management of time scores. Such a result implies that it was not the choice of research instrument that influenced the measurement scores per say, but the type of data collection (i.e. the use of low-inference observations, high-inference observations or questionnaires). Each method has its advantages and disadvantages that impact upon its measures. The authors therefore suggest that utilising more than one form of data collection would strengthen the reliability and validity of the classroom-level constructs.

It should be acknowledged, though, that the proportion of variance explained by trait factors (i.e. the 5 dimensions) was far greater than the percentage of variance explained by method factors. This suggests that method effects did not have undue influence upon the measures. Moreover, there was no evidence of systematic method bias across or within traits for student questionnaires, high-inference observations and low-inference observations. This provides further support for the convergent validity of the measures.

Part 2a: The effects of classroom-level factors on achievement in four outcomes of schooling

The second part of the study confirmed that each dimension of effectiveness was useful in predicting student performance.

Four sets of multilevel models were created. Sets 1-3 attempted to explain the variation in students' cognitive attainment in maths, Greek language (students' first language) and religious education respectively. Set 4 accounted for the variation in students' affective outcomes[23] within religious education.

[23] Affective outcomes are incorporated into the Cypriot curriculum for religious education. Students' initial and final attainment on affective outcomes could therefore be evaluated using the same form of subject-specific written test that was used to assess their cognitive knowledge.

Each analysis began with an empty model that evaluated the variation that occurred at student, classroom and school level. Explanatory variables were then added in stages. Model 1 contained only background factors, specifically; students' prior attainment, socio-economic status and gender, as well as the classroom- and school-level aggregates of these variables. Model 2a-2e contained these background factors, plus measures of the frequency, focus, stage, quality or differentiation of the eight classroom behaviours. The influence of each dimension of teaching behaviours was then be judged by the change in model fit and the z-scores of individual variables.

The results from this section of analyses are discussed below:

Empty models

The first stage of the analysis confirmed that most of the variance in student attainment occurs at the pupil level. In this instance, 73.1% of the variation in maths attainment, 75.3% of the variation in Greek language, 78.8% of the variation in religious education (cognitive outcomes), and 82.1% of the variation in religious education (affective outcomes). The effect of classroom and school, however, was more pronounced in mathematics (where 15.4% of the variance in cognitive attainment occurred at classroom level and 11.5% occurred at school level) and Greek language (where 15.4% of the variance in cognitive attainment occurred at the classroom level, 9.5% at school level) than in religious education (where 13.2%/10.4% of the variation in cognitive/affective outcomes occurred at the classroom level, and 8.0%/7.5% of the variation in cognitive/affective outcomes occurred at school level). Moreover, the classroom effect was found to be higher on achievement of cognitive rather than affective aims of religious education.

Modelling of background factors (Model 1)

Background factors accounted for roughly 50% of the variation in students' performance, with most of the explained variation occurring at the student level. Prior-attainment was shown to be the most important characteristics. This factor had a positive association with performance. In fact, it was the only attribute to have a consistent and statistically significant relationship with achievement at student, classroom and school level[24]. The remaining results suggest that socio-economic status and gender interacted with performance in the anticipated manner, though some of the interactions were not statistically significant and were not reported by the authors (see Table 7.4.3a). That is to say that being economically privileged and/or educated alongside privileged students was beneficial in most instances, whereas being female and/or educated alongside a high proportion of female students was advantageous in all subjects except mathematics, where boys outperformed girls. Overall, though, socio-economic status had the closer association with attainment, except in religious education where its influence was limited.

[24] This thesis has challenged the legitimacy of inferential statistic, especially when they have been applied to non-random samples. They are referred to in this section, however, because Creemer and Kyriakides (2008) did not report upon non-significant associations.

Table 7.4.3a. Parameter estimates for background factors from Creemers and Kyriakides (2008) analysis of achievement in mathematics, Greek language, cognitive outcomes in RE and affective outcomes in RE.

Subject/Dependent variable	Background factor	Student-level	Classroom-level	School-level
Mathematics (cognitive attainment)	Prior attainment	0.71	0.31	0.11
	SES	0.60	0.15	NSS
	Gender	-0.18	-0.05	NSS
Greek language (cognitive attainment)	Prior attainment	0.49	0.15	0.13
	SES	0.32	0.09	NSS
	Gender	0.23	NSS	NSS
RE (cognitive attainment)	Prior attainment	0.51	0.25	0.13
	SES	0.12	0.09	NSS
	Gender	0.23	NSS	NSS
RE (affective attainment)	Prior attainment	0.41	0.21	0.08
	SES	NSS	NSS	NSS
	Gender	0.18	0.05	NSS

* Figures cited from Model 1, Table 8.3a-e, Creemers and Kyriakides (2008).

** Cell values are Fisher's Z scores and therefore refer to the change in dependent variable associated with a one unit change in each factor. For small values of the correlation coefficient Z and r do not differ significantly.

***NSS = Not statistically significant. These associations were not reported within the original text.

Modelling of classroom-level behaviours (Models 2a-2e)

The addition of classroom-level effectiveness measures increased the percentage of variation that the models could explain. This occurred in each set of analyses, no matter which dimension of the constructs were considered (see Table 7.4.3b). Such a result implies that all dimensions of classroom-effectiveness factors are useful in predicting student achievement.

The quality of teaching behaviours was shown to be the most informative aspect of instruction. On average, the models that included this dimension of constructs explained 56.1% of the variation in students' scores. Whilst the models that considered the level of differentiation in instructional activities accounted for 55.5% of the variation, their timing 54.9%, frequency 54.9%, and focus 54.6%.

It should not be forgotten however that all models contained background factors which accounted for roughly 50% of the variation and that the effect of classroom practices was therefore modest.

Table 7.4.3b: The average percentage of variance explained by models that included each dimension classroom-level factors

Dimension of behaviours included in model	Mean percentage of variance explained across the four sets of analyses	Rank based on average percentage of variance explained
Frequency	54.9%	4
Focus	54.6%	5
Stage	54.9%	3
Quality	56.1%	1
Differentiation	55.5%	2

*Figures were calculated using z-scores, and are cited from Creemers and Kyriakides (2008)

It should likewise be noted that all variables did not have a meaningful impact in all analyses (see Table 7.4.3c).

The frequency dimension of most behaviours had a statistically significant associations student performance (68.4% of relationships assessed). The only factor which did not correlate with any output was teacher-modelling. Structuring and teachers' management of time, on the other hand, had statistically significant association with all four measures of student performance.

The focus dimension of at least four factors had significant associations with each attainment measure. Or, to put it another way, 50.0% of relationships evaluated across the four sets of analyses were statistically significant. Moreover, whilst no behaviour had a meaningful relationship with all four performance measures, no factor failed to correlate with at least one outcome.

The stage dimension of behaviours had a more tenuous link with performance (42.1% of factors had a statistically significant association with student attainment, across the four outcome measures). The divide, however, was reasonably clear cut and suggests that the timing of orientation, structuring and application tasks is important. These factors all had a statistically significant relationship will three or more of the outcome measures. The remaining factors had little to no association with the stage dimension of classroom variables.

65.8% of quality variables had a statistically significant association with performance. That is to say, to at least 6 variables per outcome.

Finally, 51.2% of the differentiation measures correlated with student outcomes. Questioning, application and classroom learning environment were associated achievement gain on all four outcome measures. The differentiation dimension of the remaining factors however did not have statistically significant relationship with student performance.

All of the aforementioned relationships were positive, though there was some evidence of non-linear associations. Specifically, both questioning and classroom assessment had curvilinear relationships with attainment in Greek language.

Table 7.4.3c: Overview of the impact that the five dimensions of classroom-level factors had upon student outcomes in maths, Greek language and religious education

Factor	Mathematics					Greek language					RE (cognitive)					RE (affective)				
	Fr	Fo	St	Qu	Di	Fr	Fo	St	Qu	Di	Fr	Fo	St	Qu	Di	Fr	Fo	St	Qu	Di
Orientation	+	+	+			+	+	+					+				+		+	
Structuring	+	+	+			+		+	+		+	+	+	+		+	+	+	+	
Questioning	+			++	+	Curvi			++	+	+	+		++	+	+	+		++	+
Teacher-modelling		+		+			+		+					+					+	
Application	+	+	+		+	+		+		+	+	+	+	+	+		+	+		+
Management of time	+	n/a				+	n/a				+	n/a				+	n/a			
CLE Part 1: Teach-stu relations	+		+	+	+				+	+			+	+	+	+	+	+	+	+
CLE Part 2: Student relations	+				+	+	+	+	+	+	+	+	+	+	+	+				+
Assessment		+		+		Curvi	+		+										+	

Note: The trait factors which emerged from Creemers and Kyriakides' (2008) CFA are presented here. For this reason, teacher-modelling is depicted as having four dimensions, as the variables that operationalised the quality and differentiation of this behaviour were shown to measure the same construct. Similarly, the quality of questioning was assessed using two different factors, as the measures associated with the quality of teachers' questions and the quality of teachers' responses were conceptually distinct. The remaining behaviours were modelled as having five dimensions; frequency (Fr), Focus (Fo), Stage (St), Quality (Qu) and Differentiation (Di). Light grey shading signifies that a statistically significant association with student achievement was identified ($p<0.05$). Dark grey shading signifies that both sub-divisions of the factor associated with quality dimension the questioning factor had a statistically significant association with student performance ($p<0.05$).

After observing the inconsistencies in variables effects, Creemers and Kyriakides (2008) tentatively classified the factors into three groups. The first group, which consisted of structuring, questioning, application and the classroom learning environment, had consistent impact upon students' performance no matter which dimension of the actions was considered. These factors were therefore considered to be generic. The second was comprised of teacher-modelling and management of time. Only certain aspects of these variables had a persistent relationship with performance. Specifically, the quality dimension of teacher-modelling and the frequency dimension of management of time. These findings need to be confirmed by additional research but may indicate that the model could be made more parsimonious. It should be noted, though, that former statement appears to be inconsistent with the results Kyriakides *et al.* (2013), a meta-analysis that collated evidence from 167 effectiveness studies and concluded that teacher-modelling had the second largest effect upon pupil attainment (see Section 7.4.1). This is surprising when one considers that the overwhelming majority of classroom-level effectiveness studies have focused upon the frequency dimension of effectiveness. Finally, the impact of the last two factors, i.e. orientation and assessment, was subject-specific. That is to say, that several dimensions of these behaviours were associated with attainment in mathematics and Greek language but almost none of them were related to achievements in religious education. This finding is also inconsistent with Kyriakides *et al.* (2013) conclusion, that effect sizes are not unduly influenced by the type of outcome that is evaluated. There is no outright contradiction however as the aforementioned meta-analysis did not assess performance within religious education specifically, only achievement in maths, language, affective outcomes and 'other' measures. In addition to demonstrating that differences in student attainment can be explained by more than just the frequency of effectiveness behaviours. These results therefore suggest that studies which only consider a single aspect of effectiveness could lead the researcher to draw spurious conclusions.

Part 2b: The amount of variation that can be explained when researchers account for the frequency dimension of classroom-level factors and at least one other dimension of effective behaviours

The final section of the analyses showed that there was added value in evaluating effectiveness factors from multiple perspectives.

Five multi-level regression models were created for each outcome variable (i.e. maths, Greek language, religious education (cognitive), religious education (affective)). The first four (models 2f-2i) contained measures which assessed the frequency dimension of classroom behaviours and one other dimension of teachers' actions. The fifth model contained measures for all five dimension of each classroom-level behaviour. These were compared with Model 2a (i.e. frequency model), to quantity the benefit of evaluating classroom effectiveness from multiple perspectives. All models also took into account the background factors that were considered in the previous analyses

The results of the analysis are depicted in Table 7.4.3d.

Table 7.4.3d: Percentage of explained variance in student achievement for each student outcome provided by each alternative model testing the effect of the frequency dimension of the classroom-level factors and the effect of combinations of frequency dimension with each of the other dimensions

Alternative model	Maths	Greek language	RE (cognitive)	RE (affective)
Model 2a (frequency dimension of classroom-level factors)	55.5%	55.3%	53.3%	55.3%
Model 2f (frequency and focus dimensions)	56.8%	58.7%	55.9%	57.9%
Model 2g (frequency and stage dimensions)	57.8%	59.2%	56.7%	58.7%
Model 2h (frequency and quality dimensions)	59.1%	59.7%	57.1%	59.1%
Model 2i (frequency and differentiation dimensions)	58.1%	58.9%	56.2%	58.9%
Model 3 (all five dimensions of classroom-level factors)	60.1%	60.9%	59.0%	59.8%

*Figures were calculated using z-scores and are cited from Creemers and Kyriakides (2008)

Accounting for an additional dimension of effectiveness constructs always increased the percentage of variation that classroom-level factors could explain, irrespective of the outcome considered. The combination of the frequency and quality dimension accounted for the more variation than any other two-dimension combination, which re-emphasises the importance of teaching quality.

The best fitting model though was the Model 3. This model was able to account more than 70% of the classroom-level variance in student attainment in each outcome, which implies that all five dimensions of effectiveness should be taken into account when evaluating instructional behaviours.

No model, however, accounted for more than 60.9% of the total variance. The authors attribute this to the fact that no school-level factors were operationalised. Moreover, only three student-level factors were considered. There is nevertheless a need for additional research that investigates whether greater variation can be explained when all five dimensions of school-level factors are assessed. For the time being this benefit remains a theoretical one.

Overall Creemers and Kyriakides (2008) analyses demonstrated that classroom-level effectiveness factors can and should be evaluated from multiple perspectives. Their results suggest that the five dimensions outlined within the Dynamic Model have convergent and divergent validity. Moreover, each perspective helped to predict student performance and in combination accounted for more variation than the frequency of effective behaviours can explain.

The study also found evidence of their being curvilinear relations between classroom behaviours and attainment. However, only two factors exhibited this type of association which is less than expected. It is conceivable that this was because there was insufficient variation in the functioning of the other classroom-level factors. In support this assertion the authors point out that both of the non-linear associations occurred within the teaching of Greek language, where the frequency of questioning was most varied. The pair therefore suggests that international and experiment studies may be a needed to evaluate this aspect of the model. What is clear, however, is that one should not expect to find curvilinear associations between the frequency and focus of effective behaviours in all analyses, especially when the research has been conducted within a single country using non-experimental designs.

8. Overview of the Empirical Sections

8.1 Chapter Introduction

The preceding chapters have identified several areas where Progress 8 may fall short. These problems include the difficulty of operationalising students' aptitude, the risk of judging schools for variation that is outside of their control, the effect of measurement error, missing and erroneous data, the consistency of schools' internal results (especially inter-cohort and inter-departmental ratings) and the stability of the scores over time. Whilst these issues may seem disconnected they are all underpinned by a concern that the statistical controls that this form of model uses are ill equipped to negate the multitude of extraneous influences that impact upon students' examination performance. It may therefore be that much the volatility in value-added scores is not indicative of genuine changes in school effectiveness but of the models' failure to recognise differences between consecutive cohorts of students.

The remaining sections of this thesis address these concerns by establishing whether the differences in schools' annual performance ratings and the change in schools' ratings over time can be explained by the kinds of factors that educational effectiveness is normally attributed to and perhaps more importantly, whether these factors are under the control of schools.

8.2 The Structure of This Report

The empirical investigations of this thesis can be divided into four parts. Each is presented in a separate chapter.

Chapter 9: Prediction Analysis

The first analysis asked school leaders to predict their schools' value-added score in advance, based on their in-depth knowledge of their school. Since these individuals are the ultimate authority on their institutions, one would anticipate that if the variation in value-added ratings were indicative of genuine changes in school effectiveness, then the endeavour should be reasonably successful. Even after we take into account that value-added ratings are a relative construct and that the precise Key Stage 4 attainment level required to achieve a particular ratings will vary slightly year to year, it stands to reason that if the measure has any pragmatic value, those with the most informed opinion should be able to foresee, at the very least, any dramatic changes in their school 'effect'. The foresight of school leaders was thus evaluated and the implications for the practical application for Progress 8 considered.

Chapter 10: Thought Experiment

In the second empirical section a thought experiment was conducted to assess the implications of there being inaccuracies in students' Key Stage 2 and Key Stage 4 data. More specifically, the DfE national attainment averages from 2019 were used to evaluate how a 10% measurement error in students' KS2 fine-levels would impact upon students' progression scores. The magnitude of these inaccuracies was then compared to the error that would occur if students' Attainment 8 score were over-stated by 10%.

The more distinct the former is from the latter, the more differential the two effects were assumed to be.

The analysis therefore overlaps with the work of researcher such as Gorard (2010a), Reynolds *et al.*, (2012) and Perry (2019) which were discussed at length earlier in the thesis. To our knowledge, though, the argument that error in students' prior-attainment data might be more influential than error in the final-attainment data, has never been made before.

Chapter 11: Shallow Regression Analysis

The third assessment evaluated the performance of 125 state-funded schools using a well-established model of educational effectiveness. This model was intended to map the most important influences on students' learning including differences in students' characteristics and background, their schools' policies and teaching practices. If the results of Progress 8 assessments are valid then this model should account for a sizeable portion of the variation in schools' Progress 8 results, both at specified points in time and over time. Regression analysis was used to test whether this was the case.

Furthermore, since the model purports to distinguish between the effects of school-related and non-school factors, deductions are made about the percentage of variation that could be explained by genuine school effects.

Whilst the scope of this analysis was deliberately restricted and neglects some of the more recent additions to school effectiveness theory, including some of the dimensions upon which school effectiveness can be measured, the most influential and frequently referenced factors were considered. The compromise, however, encouraged a higher rate of participation and thereby permitted the use of more informative statistical techniques which can only be utilised when a large sample is available.

Chapter 12: Detailed Regression Analysis

The second analysis examined the performance of 9 schools in greater detail.

Once again, regression analyses was used to evaluate whether established effectiveness correlates could account for the differences in schools' performance ratings at specified points in time, and the change in schools' scores over time.

This time, though, a far wider range of variables was operationalised. This reduced the risk of omitted variable bias substantially. The enhanced coverage however came at a price. Since the number of explanatory variables that can be justifiably included in multiple-regression models is contingent upon sample size, the complexity of the analysis had to be reduced. This section therefore assessed the relationship between each factor and schools' performance ratings individually, without controlling for the effect of other factors.

After establishing which factors best explained the differences in schools' Progress 8 ratings, and whether these were under the control of schools, logical conjectures were made about how the variables are likely to interact as a group. Whilst these judgements were informed by evidence from the previous chapter and comparable effectiveness studies, they are speculative and require validation from future research.

The analysis nonetheless provides an interesting case study as to the effect that observable changes in schools' policies and practices have upon schools' performance ratings.

8.3. Adaptions Necessitated by the Covid-19 Pandemic

Within educational effectiveness research it is generally accepted that school effectiveness should be evaluated based upon student-level outputs (Aitkin and Longford, 1986; Raudenbush and Willms, 1995; Woodhouse and Goldstein, 1988). That is to say, that one should judge school performance based on the difference between each student's final attainment level and that predicted by ones value-added model. This helps the researcher to create a more precise representation of the relationship between attainment and any factors included within the model (Raudenbush and Willms, 1995), it allows a wider range of interactions to be studied (Aitkin and Longford, 1986; Bryk and Raudenbush, 1992) and helps to prevent misinterpretations of the data (Dettmers *et al.*, 2009; Snijders and Bosker, 2011). The intention was therefore for this thesis to utilize student-level performance data from the National Pupil Database within Sections 3 and 4 of the empirical analyses. Unfortunately, Covid-19 interfered with our ability to access a secure research environment. The methodology of these analyses therefore had to be adapted so that they could be performed with publically available cohort-level aggregates.

Theoretically, there is nothing wrong with analyses that utilise aggregated data provided that the researcher is only interested in macro-level relations (Creemers and Kyriakidies, 2008). Problems such as the 'shift in meaning' (Huttner and van de Eeden, 1995) or the 'ecological fallacy' (Alker, 1969) can occur, however, if such data is used to interpret micro-level or cross-level interactions as the data no longer refers directly to the micro-level units. The first of these pitfalls describes a scenario where the meaning of a collective variable differs from that of the individual-level metric. Student aptitude could, for example, be viewed as an indication of a student's ability to perform at task. The average ability of a class however is indicative of the level at which their curricular should be set. Thus, the individual and collective measures of attainment can have different implications. An ecological fallacy, on the other hand, occurs when the interactions between micro-level units and macro-level units differ. An example of this is the relationship between praise and performance. Whilst it is generally accepted that the receipt of positive feedback can have a beneficial effect upon students' performance (Walberg, 1986) there will sometimes be a negative relationship between the frequency of praise and the attainment if the relationship is evaluated at classroom-level. This is because teachers tend provide more encouragement to low-ability groups (Brophy, 1992). Researchers must therefore be cautious of drawing conclusions about individual-level interactions based on correlations within aggregate-level data as these relationships sometimes conflict.

Within the empirical sections the potential for misinterpretations was reduced by only reporting upon cohorts' performance and the interaction between cohort-level variables. That is to say, that the models identify, for example, whether schools' with high percentage of disadvantaged Progress 8 entrants tended to perform above or below the national average. It is recognised, however, that the interpretation of these relationships was informed by educational effectiveness literature and that much of this will have utilised student-level data. There remains therefore a risk that some of the results that were deemed unexpected may in fact be consistent with past observations and merely viewed from a different perspective. That being said, it is argued that such instances will be in the minority and the author was alert to the risk. The probability of such phenomenon affecting the substantive findings of the study is therefore very slim.

These extenuating circumstances also prevented the analyses from using multi-level modelling techniques, something that is expected by many educational effectiveness researchers (Teddlie and Reynolds, 2000). This form of modelling would have provided a preferable means of evaluating collective influence of effectiveness factors because it acknowledges the clustering of educational data. In other words, the fact that students are educated in classrooms and classrooms are located within schools (Goldstein, 2011; Heck and Thomas, 2000; Raudenbush and Bryk, 1986). The models would also have broadened the range of analytical possibilities by allowing the paper to directly report upon the percentage of variance that took place at each level of analysis, something that was reported indirectly within the assessments. Perhaps the most influential shortfall though is the resulting inability to distinguish between the individual-level effect of background factors and any compositional effects. Within all of the subsequent analyses any relationship that non-school factors such as socio-economic status, gender and SEN status have with students' performance was interpreted as bias. However, since these variables are all assessed as cohort-level averages it is possible that compositional effects were also at play. That is to say, that the effect of being educated alongside students with advantageous characteristics may have had a beneficial influence upon students' performance that should count towards schools' Type A effect (see Section 2.7). That being said, as peer effects are a secondary effect of differences in intake and there is debate as to the legitimacy of the effects (see Section 4.3.3), these types of factors are unlikely to have been the predominant difference between institutions. Whilst our solution is not ideal we are therefore confident that it is adequate. Strictly speaking, however, the shallow- and detailed-regression analyses provide a more valid report upon the quality of schools' provisions than they do of the legitimacy of basing parental decision upon Progress 8 results.

Finally, though it has little impact upon the study, it be should acknowledged that the proponents of the multi-level methodologies argue that one of the principle reasons for constructing multi-level models is that the calculations produce more accurate and cautious estimations of standard errors (Snijders and Bosker, 2011). Logically, this should increases the validity of the significance tests that are used to identify influential variables. In previous sections, however, it argued that inferential statistic do not report the type of information that is needed to make such decisions, especially when they are applied to non-random samples (see Section 6.3). Within the empirical sections of this thesis, the most important variables were therefore identified based on their theoretical importance and the percentage of variance that they explain. This position is consistent with that assumed by Gorard (2007; 2014) and several other researchers (e.g. White, 2014).

9. Prediction Analysis

9.1. Chapter Introduction

School leaders are a key authority for their schools. They are instrumental in developing their school's mission, structure, policies, culture, resources and strategies for improvement (Leithwood, *et al.*, 1998). It is therefore reasonable to assume that these individuals will have a detailed knowledge of the factors that differential school effectiveness is normally attributed to both in policy and research. If Progress 8 provides a valid and reliable measure of school performance, then knowledge of such factors and how they might have changed would help leaders to anticipate their schools' ratings. This section assesses whether this was the case.

9.2. The Department for Education's Stance on Statistical Projections

How legitimate is it to ask school leaders to predict their school's Progress 8 scores? Whilst the DfE have acknowledged that statistical projections of school performance have their place, their overall stance is one of caution. Specifically they emphasise that "care should be taken when using a previous year's attainment averages as a guide to potential future Progress 8 results" (DfE, 2020, p. 19, ln 13-14). Their concern stems from the fact that Progress 8 is a relative measure. Each year pupils' progress scores are calculated by comparing their Key Stage 4 results with the performance of other students in their national cohort. Since subject entry patterns and the performance of each prior-attainment group can change annually, the Key Stage 4 point score that a student requires to achieve a particular Progress 8 rating will fluctuate. Any predictions that are made about schools' Progress 8 ratings will be based upon the attainment averages from previous year's assessments, which will be slightly different from the averages in the specified year. This however, does not necessarily make predictions invalid. The legitimacy of predictions merely depends on the nature of this variation.

The methodology of the new analysis in this thesis was based upon two assumptions. The first was that, over a period as short as one year, the average quality of the education offered by most publicly-funded schools in England should remain fairly stable. Some, individual schools would be expected to improve and others would inevitably suffer setbacks. However, the policies and practices of the majority of institutions were expected to change only gradually, if at all. Any changes that did occur were also expected to take time to have their full effect (Creemers and Kyriakides, 2008).

Note that in making these statements I am distinguishing between the quality of schools' instructional practices and the alignment between the taught and assessed curriculum. The latter may well have been influenced by the reform to GCSE grading which took place during the study (although these reforms would affect all schools). To control for this effect I highlighted the issue to school leaders and asked them to take the assessment protocols into consideration when making their estimates. The volatility of school attainment averages was also assessed and deemed insufficient to interfere with the analysis. Specifically a 0.987 correlation was found between the 2015 and 2019 attainment averages (DfE, 2020), showing that even over a period of four years the raw scores figures remain comparable. The DfE make similar assumptions about the stability of school performance by using value-added models within the context of a market-based accountability system. If the quality of schools' provisions are not comparable one year later, then Progress 8 scores cannot help parents to select the best school for their child six years in advance. Similarly, if the computation of Progress 8 is assumed to be so volatile that the same student making the same progress would get markedly different value-added scores in different years, what meaning can practitioners derive from their score? How is one to discern whether an improvement strategy has worked, or even whether the scores are valid, if any variation can reasonably be attributed to the performance of other schools?

A second assumption was that much of the year-to-year variation in pupils' national attainment averages would be due to changes in subject entry patterns. This was expected to be reasonably predictable as students' entries are deliberately guided by Ofqual's list of approved qualifications, their point scores and the weighting attached to each subject in the Attainment 8 calculation. It was therefore assumed that school leaders would be aware of the qualifications that receive the most recognition, and the extent to which they had encouraged pupils to study those subjects in comparison to other years and could adjust their expectations accordingly. During the pilot study and the initial discussions with school leaders the participants made regular and astute comments about how they expected their schools' examination entry practices would or could alter their performance ratings. This supports the assumption that school leaders are aware of such issues and appear able to make logical inferences about the effect of their curricular and assessment decisions. To ensure that they had considered the matter fully the questionnaire purposefully drew attention to these factors and enquired about key aspects of the schools' provision.

Validity is not intended to be a general concept. It is specific to the intended use of the data (Brualdi, 1999; Messick 1996b). Whilst the variation in national attainment averages may well be substantial enough to hinder precise predictions being made, this was not the sole concern of the study. The research was predominately interested in whether school leaders could anticipate major shifts in their upcoming Progress 8 score. This level of prediction accuracy may not be sufficient to satisfy parents or inform schools' improvement efforts. It would, however, suggest that the year-to-year variation in Progress 8 scores is indicative of genuine changes in schools' effectiveness and not merely the influence of unacknowledged variables or random fluctuations.

9.3. Method Section

9.3.1. Research sample

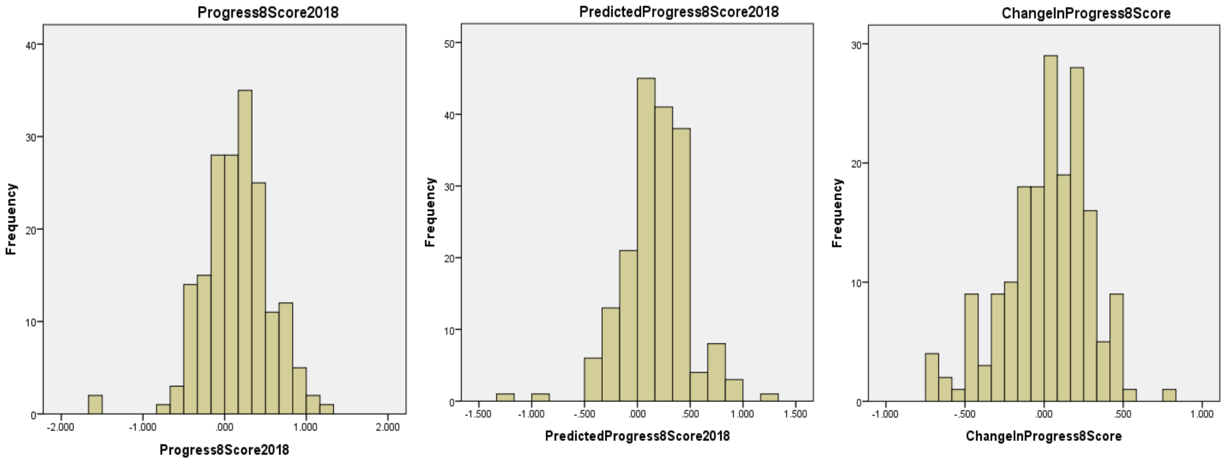
Participation in Progress 8 is mandatory for all state-funded mainstream secondary schools in England. In 2018 a total of 3659 schools fell into this category (EduBase, 2018). However, after restricting the sampling frame to exclude pupil referral units and schools that do not educate students from the beginning of Key stage 3 to the end of Key Stage 4, the population then became 2991 schools. This research uses a convenience sample of 192 schools from this population. More specifically, all of the schools in the sampling frame were identified using Edubase. Each was contacted via email to request that they take part in the study. Non-responses were followed up with a second invitation. All schools took part on a voluntary basis.

From these 192 schools, 196 predictions were received, 182 of which were included in the analysis. Four schools were excluded because they did not have Progress 8 results in specified academic years. Two responses were omitted because there were identical predictions from the same schools (in both instances, two deputy heads responded to the questionnaire, despite the clear instructions that one response was requested from each institution). Two responses were removed because they were from schools not intended to feature in the sampling frame and three predictions were excluded because the information that the respondent provided could not be matched to a specific school. Finally, three school leaders declined to make a prediction and could not be included within the analysis (this is significant as anyone that refused to participate must weaken the evidence of predictability).

Two additional leaders submitted predications for the same school. As these estimates were different both were included in the analysis. Again, in these instances the two estimates came from two deputy head-teachers who presumably made different qualitative judgements about the quality of the schools' provisions or the effect of external changes. And again the fact that the predictions differed weakens the evidence of predictability.

An obvious limitation of this approach is that convenience samples are particularly vulnerable to selection bias. In other words there is an increased likelihood of there being differences between the achieved sample of schools and the population they are intended to represent. Under most conditions this reduces the certainty with which researchers can generalise their findings. In this instance however the primary objective was not to describe the distribution of scores within the population. That is to say, that the intent was not to infer how accurate school leaders' estimations would be within other schools. Instead, the principal concern was to establish whether leaders' appraisal of their schools' performance were in line with their official Progress 8 ratings. The priority was therefore to ensure that a wide range of Progress 8 scores and estimations were represented. This necessitated a large sample, which the adopted sampling procedures made possible. The sample might still provide a reasonably accurate representation of the population because the Progress 8 scores, predicted Progress 8 scores and the changes in Progress 8 scores were approximately normally distributed (see Figure 9.3.1a). The sample also consisted of schools from all over the country (84 different local authorities) and included the most common school types. The sample consisted of 10.4% sponsor led academies, 60.4% converter-mainstream academies, 13.7% community schools, 8.2% foundation schools, and 7.1% voluntary aided schools. See Section 11.2.1 for further detail on the population of state-funded schools.

Figure 9.3.1a: Histograms of the sampling distribution of Progress 8 scores, predicted Progress 8 scores and the changes in Progress 8 scores



A potential bias was that the mean of the aforementioned variables were all fractionally above zero (at 0.149, 0.181 and 0.019 respectively), despite the average Attainment 8 scores being very close to the national average (mean Attainment 8 score in sample in 2018 = 50.967, mean Attainment 8 score in population in 2018 = 50.869). This suggests that schools were slightly more likely to participate in the study if their leadership team anticipated favourable results. The deviation was minor.

9.3.2. Data collection

In order to assess whether changes in schools' Progress 8 ratings can be anticipated by individuals who have a detailed knowledge of what has been going on within the institutions, this study asked school leaders to predict their school's rating in advance. More specifically, school leaders were instructed to provide a point estimate of their school's 2018 Progress 8 score based upon a personal appraisal of their schools' provisions and changes to the external environment.

The predictions were collected using an electronic questionnaire that was distributed between March 2018 and July 2018 (see Appendix A). All estimates were therefore made at the end of students' Key Stage 4 education but before the school had knowledge of their KS4 attainment outcomes. This prevented *post hoc* interpretation of the results. Checks were carried out to verify that that all respondents were suitably placed to be regarded as experts on their school, and the overwhelming majority of questionnaires were completed by the head-teachers themselves. The remaining schools allocated the task to suitably senior and well placed individuals such as a deputy head-teacher or the progression leader. The completed forms were later matched to school performance data from the National Pupil Database.

In addition to providing an avenue for school leader's predictions, the questionnaire also asked about key areas in the schools' provisions. These questions were predominantly based around the effectiveness factors identified in the Dynamic Model of Educational Effectiveness (Creemers and Kyriakides, 2008), with additional items to evaluate whether there had been any changes to the school's exam entry patterns or the allocation of lesson time between subjects. The responses to these questions are presented in the latter sections of this thesis. It was hoped that issuing the two research instruments together would ensure that school leaders considered all important factors before making their prediction. A terminology sheet was also provided alongside the questionnaire to ensure that any technical terms were fully understood.

9.3.3. Data analysis

The level of agreement between school leaders' expectations and schools' Progress 8 scores was evaluated in four stages:

Part 1: Descriptive statistics of the accuracy of school leaders' predictions

In the first stage of the analysis some simple computations were performed. More specifically, the average deviation between school leaders' estimates (2018) and their schools' Progress 8 scores (2018) was calculated, along with the range of prediction errors and the standard deviation of errors.

The extent of any inaccuracy was expressed in absolute terms and relative to size of schools' Progress 8 scores, with the relative error being defined as the difference between schools' official Progress 8 ratings and leaders' predictions, divided by the official Progress 8 score of the school.

These figures established how far school leaders' estimates were from the schools' actual performance ratings. A high degree of consistency between actual and predicted scores was viewed as evidence of Progress 8's validity, whereas large discrepancies were treated as a cause for concern. Small to moderate deviations, however, were expected and were not deemed sufficient cause to question the validity or reliability of Progress 8 assessments.

Part 2: The association between school leaders' estimations and schools' progress scores

The correlation between school leaders' predictions (2018) and schools' Progress 8 scores (2018) was then calculated. This established whether higher than average predictions were associated with higher than average scores (r-score). It also quantified the proportion of the variation in Progress 8 scores that school leaders' anticipated (r-squared score).

Part 3: The unique information that predictions provide

In the third section of the analysis a multiple-regression model was created, with schools' Progress 8 scores (2018) as the dependent variable, and the schools' former (2017) and predicted ratings (2018) entered as independent variables.

These variables were entered in stages, so that the relationship between schools' Progress 8 ratings (2017) and Progress 8 ratings (2018) could be reported, followed by the additional variation that school leaders' predictions were able to explain by themselves.

This model was intended to acknowledge that educational stakeholders would have access to schools' 2017 ratings, and the worth of leaders' predictions must therefore be judged in relation to the variation that the 2017 scores could explain.

Whilst a high degree of consistency between the 2017 and 2018 performance ratings was interpreted as a sign of Progress 8's reliability and validity (see Section 5.2.1), it was also expected that school leaders' insight would account for additional variation that the preceding ratings could not. The more unique information that school leaders' provide, the more useful the information they provide.

Part 4: Leaders' ability to predict changes

In the final section the issue was approached from a slightly different perspective.

This section reports the correlation that existed between school leaders' estimate of 2017-18 changes and the change in Progress 8 ratings that actually took place. It also calculated the percentage of the latter that the former could explain.

Finally, the percentage of school leaders that correctly anticipated whether their school's performance would improve upon, remain the same, or decline between summer 2017 and summer 2018 was computed.

These analyses provide comparable information to the preceding sections, but are intended to deliver a more intuitive and pragmatic interpretation of the predictions' value.

One methodological point to note is that at several points in the aforementioned analyses the average error within a particular type of prediction was calculated. In these situations both the mean and median error are reported. This is because the data was positively skewed. The mean values might therefore give an exaggerated impression of the error within school leaders' predictions.

9.4. Results[25]

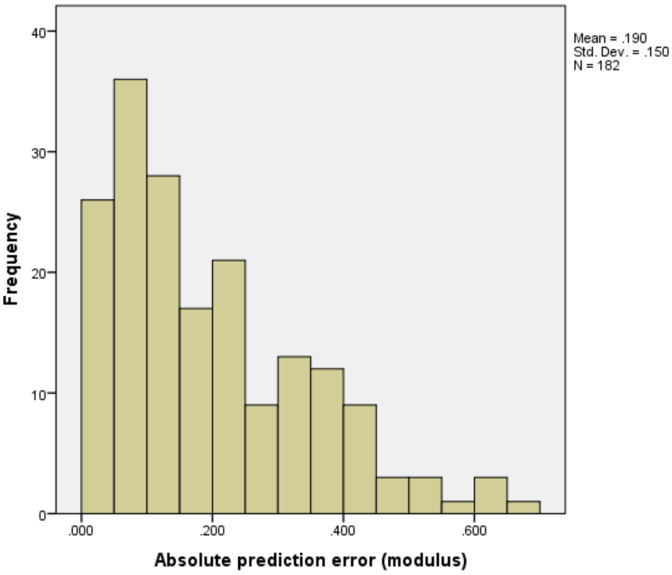
Part 1: Descriptive statistics of the accuracy of school leaders' predictions

i. The absolute error in leaders' predictions

Within the sample, the mean prediction error was 0.190 Progress 8 points. This signifies that, on average, school leaders' estimates of cohorts' relative progress were out by 0.19 of a GCSE points per subject area. Though given that the data was highly skewed (skewness = 0.946, see Figure 9.4a) it is perhaps more appropriate to state that the median error was 0.150.

The remaining errors, however, were only loosely clustered around this point (standard deviation = 0.150), so far greater deviations were common (minimum error = 0; maximum error = 0.660).

Figure 9.4a: Histogram of the absolute error within leaders' predictions



[25] Table of results available in Appendix B.

ii. The relative error in leaders' predictions

From the above, one gets the impression that leaders' predictions were somewhat inaccurate. It is more informative, however, to view these errors in relation to the size of schools' value-added residuals (see Figure 9.4b).

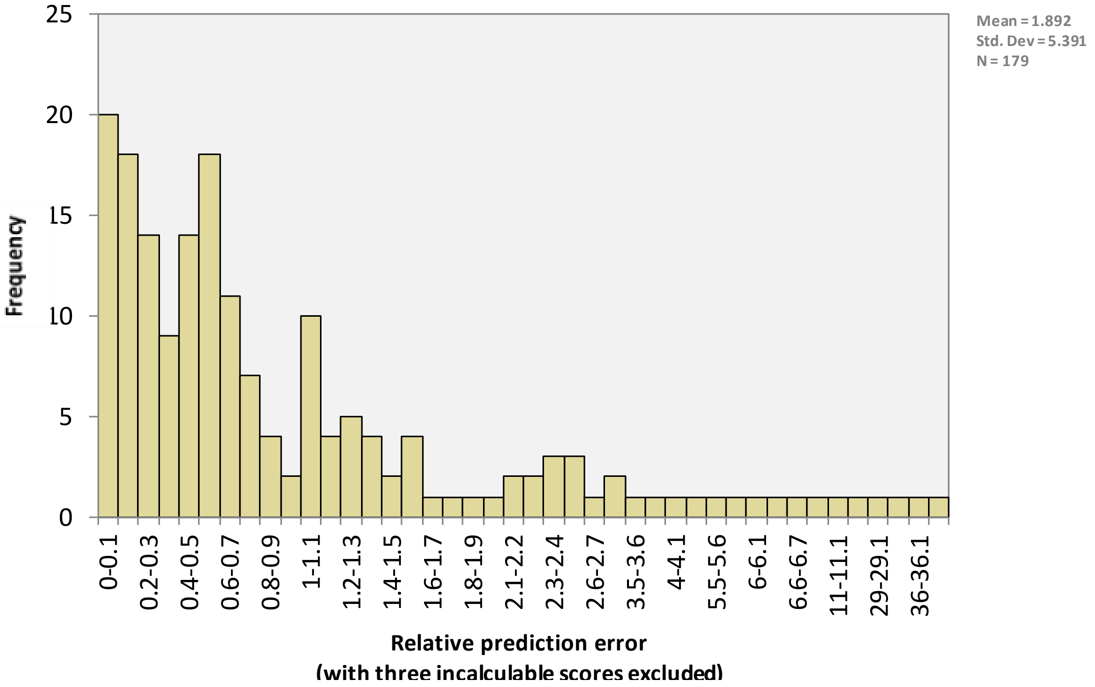
In the median case, the relative error in leaders' measurements was 0.583, or 58.3% of the size of the progress score that the individual was trying to predict (mean relative error = 1.892; Skewness = 5.717).

In fact, 36% of predictions had an error component that was larger than the Progress 8 score itself.

Far larger relative errors were present, though it should be noted that the most extreme occurred when Progress 8 scores were very close to zero (minimum relative error = 0.0; maximum relative error = 43.0; standard deviation = 5.391). The absolute error within these scores was therefore not atypical.

It should likewise be noted, that the results of the three schools with Progress 8 scores of zero had to be excluded from these calculations as the relative error in their score was incalculable.

Figure 9.4b: Histogram of the relative error within leaders' predictions



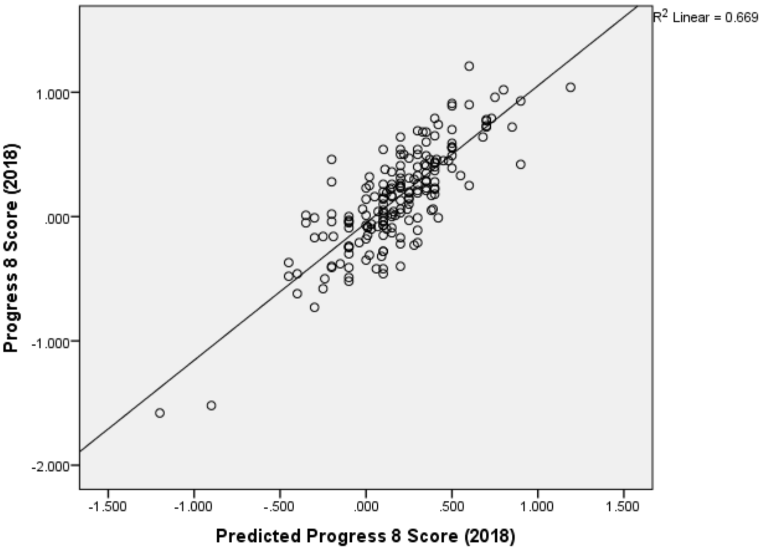
These statistics illustrate that leaders' predictions were loosely in-line with schools' official performance ratings. Substantial deviations however occurred on a regular basis, with many dwarfing the magnitude of schools' Progress 8 score. This implies that leaders will often struggle to predict whether their school would receive a positive or negative rating.

Part 2: The association between school leaders' estimations and schools' progress scores

i. The relationship between school leaders' predictions (2018) and their schools' Progress 8 scores (2018).

A strong positive correlation (r = 0.818) was found between school leaders' predictions and school's progress scores. Higher than average predictions were therefore associated with higher than average ratings, and vice versa (see Figure 9.4c). Moreover, the strength of this association demonstrates that roughly two thirds of the variation in schools' ratings (66.9%) was anticipated by school leaders.

Figure 9.4c: Scatter graph of the relationship between predicted Progress 8 scores (2018) and Progress 8 scores (2018)



A high proportion of the variation in schools' Progress 8 ratings therefore appears to be explicable, though it should not be forgotten that school leaders' failed to account for almost 1/3 of the variation in schools' ratings.

Part 3: The unique information that predictions provide

Step 1:

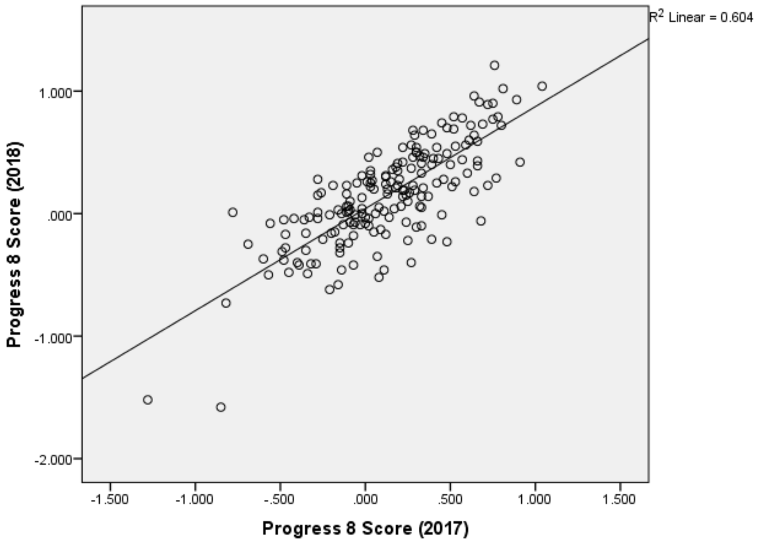
In Step 1 of the Multiple Regression Model, a moderately strong positive linear correlation was found between schools' previous progress rating (2017) and their current Progress 8 scores (2018) (r = 0.777). This means that 60.4% of the scores from consecutive evaluations were consistent (see Figure 9.4d).

Within the context of this analysis, this implies that a portion of the variation in Progress 8 ratings that school leaders' foresaw could also be predicted by consulting schools' existing performance data.

School leaders' insights are considerably less useful, therefore, than the correlation in Part 2 suggests. The extent of this overlap is calculated and reported in the latter half of this model (see Step 2).

From the perspective of school accountability a moderate level of inter-year stability is less than many researchers and/or practitioners would desire, as it limits the pragmatic worth of value-added ratings and cast doubt upon their medium-term validity. See next section for further discussion.

Figure 9.4d: Scatter graph of the relationship between Progress 8 scores (2017) and Progress 8 scores (2018)



Step 2:

In Step 2 of the multiple regression model the combined explanatory power of schools' 2017 Progress 8 ratings and school leaders' predictions was evaluated.

The assessment yielded the following equation:

$$E(y) = -0.031 + 0.307x_1 + 0.772x_2$$

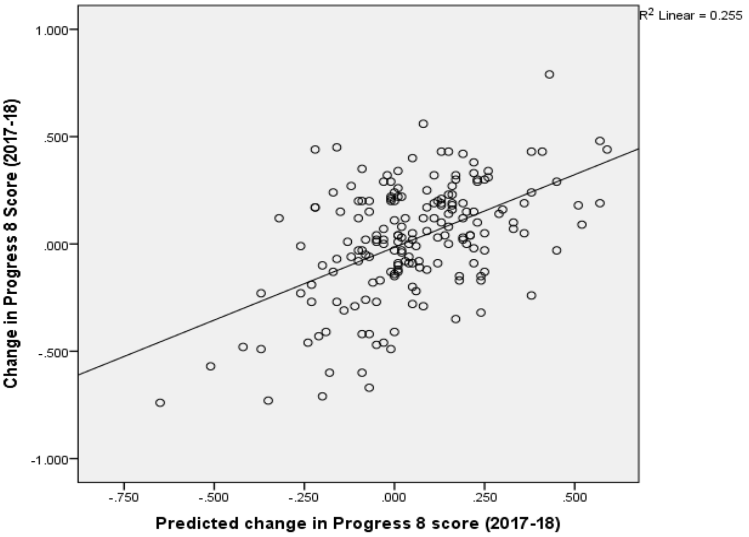
This function had a close association with schools' Progress 8 results (r = 0.831), that explained 69.0% of the variation in Progress 8 scores (2018). This was only slightly higher than the percentage of variation explained by the 2017 progress scores. Therefore, whilst the two independent variables accounted for a slightly higher percentage of the variation in schools' Progress 8 results (2018) than either could in isolation, the percentage of variance that could only be explained by leaders' predictions was small (8.6%). It thus follows that school leaders' insight does not extend far beyond the knowledge of their school's previous performance rating.

Part 4: Leaders' ability to predict changes

A similar picture emerged in the fourth part of the analysis.

A modest correlation (r=0.505) was found between the change in Progress 8 ratings that school leaders' anticipated and the change that actually occurred (2017-2018). Whilst this suggests that school leaders' had a foresight, the relationship only accounted for 25.5% of the variation in Progress 8 scores. There was therefore a great deal of noise or something else within the relationship, and the manifest changes in schools' ratings varied substantially from the projected amounts (see scatter is Figure 9.4e). This is consistent with the finding that school leaders' insight is minimal after differences in the schools' previous ratings have been taken into account.

Figure 9.4e: Scatter graph of the relationship between the predicted and manifest changes in schools' Progress 8 scores (2017-2018)



A useful way of understanding the impact of this 75% of unexplained variation is to consider the fact that only 62.1% of school leaders were able to anticipate whether their school's rating would improve, remain the same, or decline between the 2017 and 2018 Progress 8 assessments. More specifically, 39/74 (52.7%) of the leaders from schools' with declining scores, 0/3 leaders (0%) from schools with the same score (0%), and 74/104 (71.2%) of leaders from improving schools foresaw their school's outcome.

Two statements should be made about these figures. First, there is a need to acknowledge that the percentage of correct predictions was notably higher amongst improving schools than schools with lower ratings. The intuitive explanation for this is that it reflects bias within the sample. All school leaders' in the study took part on a voluntary basis. It would therefore be understandable if these individuals were more likely to participate if they expected the study to reflect favourably upon their institution. Though this is impossible to prove, the fact that the average Progress 8 score, the average predicted Progress 8 score and the average 2017-2018 change in Progress 8 scores were all above average within the sample, supports the assertion (see Section 9.3.1). It is also possible that school leaders have a tendency to be slightly optimistic about the impact of their improvement efforts.

The second point to note is that it is unsurprising to find that none of the leaders from schools' with stable Progress 8 scores foresaw this outcome. This is because of the way 'correct' and 'incorrect' decisions were classified. Specifically, the fact that these individuals needed to provide estimates that were accurate to two decimal places in order to be considered correct. Whereas the leaders from improving (or declining) schools could have anticipated that their school would receive any rating, as long as it was higher (or lower) than the schools' previous score. The results reflect this disparity.

9.5. Discussion

The evidence collated in this section cannot provide incontrovertible proof that the Progress 8 does or does not provide a valid measure of school performance. That being said, there is sufficient reason here to question whether it can adequately perform all of the functions that it is assigned.

Schools' performance ratings, for instance, were volatile. In fact, the association observed between consecutive Progress 8 ratings (r=0.777) was comparable but slightly lower than the correlation between consecutive Best 8 value-added ratings (see Section 5.2.1). This may suggest that some form of construct irrelevant variance was interfering with the measure. Even if this were not the case, this level of association means that even one year apart only 60% of the variation in schools' scores was consistent. Were this level of association to continue, schools' scores would be largely unrelated after only a few years. The same correlation between scores, for example, would result in 36% of the variation in ratings being consistent over two years, 22% over three, 13% over four, and less than 8% after five years. This is concerning as parents have been actively encouraged to select their child's school based on value-added ratings that were calculated six years before their child would sit their GCSE examinations. After such a prolonged period, however, it is doubtful that the ratings would tell them anything about the education their child will receive. Progress 8 scores may not therefore be a dependable method of selecting the best secondary school for one's child. A wider concern is that the organisation and funding of all state-maintained education is currently dependent upon Progress 8 ratings and the underlying assumption is that these assessments provide a stable indicator of schools' future performance (see discussion of market-based accountability and funding within Section 2.2). If, instead, school performance is found to vary dramatically from year to year then the logic of, for example, Ofsted inspecting schools less frequently if they receive a highly positive rating, should be reconsidered.

What is more, if a high proportion of the variation in school effectiveness ratings was genuine it is rational to expect that school leaders would have been able to anticipate changes in their schools' scores. As experts of their institutions these individuals have an intricate knowledge of the factors that school effectiveness is normally attributed to, and this information should provide a degree insight. The evidence amassed in this section however suggests that the foresight of school leaders was very limited. Whilst there was a reasonable correlation between school leaders' estimates of their school's Progress 8 scores and their school's official value-added ratings, which suggests that the scores were influenced by the changes that occurred within institutions, school leaders were only able to account slightly larger proportion of the deviations in scores (8.6%) than the 2017 ratings. In fact, due to the r=0.858 correlation that existed between school leaders' estimates and the schools' previous performance ratings, it is likely that most of their ability to predict scores came directly from their knowledge of the 2017 results and/or them having access to the 2017 national attainment averages.

Sizeable prediction errors were also common, with many dwarfing the size of the schools' value-added residuals. This ultimately meant that nearly 40% of leaders were unable to specify whether their schools' rating would improve, remain the same or decline in their next evaluation. Given the level of information that school leaders have at their disposal, this is not a high success rate. The unpredictability of Progress 8 ratings increases the evidence for concluding that they may not provide an accurate and reliable measure of school performance.

It also is worth acknowledging though that even if reported correlations are interpreted in the best possible light and it is assumed that the inability of school leaders' to predict their schools' ratings is hindered only by the changes in the performance of other schools (the zero-sum problem), this would still be concerning. This is because the ratings would only allow school leaders to respond to the evaluations in a retrospective manner, and to act based on the performance of students that have already left their school. Whilst this does not necessarily imply that the rating would not be helpful, it does raise the question of whether a relative measure of between-school performance is the best choice of performance indicator. An absolute measure such a regression discontinuity or a within-school measure may therefore have greater utility.

9.6. Possible Amendments to the Methodology

Whilst the evidence considered in this section would always be circumstantial it is worth noting the methodology of this analysis could have been improved by asking school leaders to specify in detail how they derived their predictions. Though it was assumed that school leaders would base their prediction either upon their schools' 2017 scores or the 2017 attainment averages there are a number of ways in which they could have arrived at their estimate. School leaders could, for example, have produced estimates in an entirely mechanical manner. That is to say, that they may have predicted each student's progression as being equal to the differences between their predicted Attainment 8 scores (based on historical performance data, mock examinations and/or teachers' assessments) and the Attainment 8 scores that would have been projected if the students' Key Stage 2 assessment data were interpreted using the preceding years' KS2-KS4 attainment averages. Others may have modified this figure, or produced an entirely subjective estimate. School leaders may also have taken into account recent changes in the specification of the model, for example, the fact that extremely low scores were capped in 2018 but were not in 2017, or failed to consider such matters. Knowledge of these differences would have helped to draw more informed conclusions.

9.7. Conclusion

This analysis scrutinised the validity and reliability of Progress 8 ratings. In particular, it tested whether schools' value-added residuals provide a meaningful indicator of institutional effectiveness. It did this by first establishing the stability of schools' scores and then testing whether the year-to-year change in schools' performance ratings could be predicted by school leaders. Theoretically, if all deviations reflected the changes that occurred within schools, then the expert knowledge of these individuals should have ensured that the endeavour was successful.

Despite a base level of agreement between leaders' estimates and schools' official ratings, the analysis concluded that school leaders' insight was minimal. This suggests that some form of construct irrelevant variance may have impacted upon schools' ratings, but does not definitely prove that this was the case. More detailed and persuasive evidence would require a more robust research design that controlled extraneous influence upon schools' performance data (see latter empirical sections).

On its own however the volatility observed in Progress 8 ratings provided sufficient basis for questioning whether some applications of the scores are appropriate, particularly the notion that schools' residuals can act an effective basis for parental decisions about school choice.

10. Thought Experiment

10.1. Chapter Introduction

This section considers whether errors in Key Stage 2 (KS2) and Key Stage 4 (KS4) datasets have a comparable impact upon Progress 8 ratings.

10.2. Method Section

All value-added calculations are underpinned by at least two sources of information, data on students' prior- and current-attainment. These datasets contain errors that will impact students' progression scores (Gorard, 2010a). In this section a thought experiment is presented which considered the impact of these inaccuracies. Specifically, the DfE national attainment averages from 2019 were used to evaluate how a 10% measurement error in students' KS2 fine-levels would impact upon students' progression scores. The magnitude of these inaccuracies was then compared to the error that would occur if students' Attainment 8 score were over-stated by 10%. The more distinct the former is from the latter, the more differential the two effects. The analysis therefore overlaps with the work of researcher such as Gorard (2010a), Reynolds *et al.*, (2012) and Perry (2019) which were discussed at length earlier in the thesis. To our knowledge, though, the argument that error in students' prior-attainment data might be more influential than error in the final-attainment data, has never been made before.

10.3. Results

The findings of the analysis are described in Table 10.3a.

Table 10.3a: Comparison of the error that will result in Attainment 8 estimates because of ten percent errors in students' Key Stage 2 and Key Stage 4 scores

KS2 fine-Level	Corresponding Attainment 8 estimate	10% over-statement of KS2 score	A8 prediction that would result from the 10% KS2 error	Absolute error in A8 prediction that transpires because of inaccurate KS2 scores	The relative error in A8 estimate that transpires because of 10% error in KS2 score	10% over-statement of A8 score	Absolute error in in A8 estimate that results from 10% error in KS4 dataset	Deviation in impact of KS2 and KS4 error ***
1.5	15.13	2.0*	17.24	2.11	13.95	16.64	1.51	0.60
2	17.24	2.5*	17.49	0.25	1.45	18.96	1.72	-1.47
2.5	17.49	2.8*	17.49	0	0.00	19.24	1.75	-1.75
2.8	18.29	3.1	20.65	2.36	12.90	20.12	1.83	0.53
2.9	19.81	3.2	22.44	2.63	13.28	21.79	1.98	0.65
3	20.65	3.3	23.12	2.47	11.96	22.72	2.07	0.41
3.1	21.63	3.4	23.97	2.34	10.82	23.79	2.16	0.18
3.2	22.44	3.5	24.87	2.43	10.83	24.68	2.24	0.19
3.3	23.12	3.6	25.66	2.54	10.99	25.43	2.31	0.23
3.4	23.97	3.7	26.54	2.57	10.72	26.37	2.40	0.17
3.5	24.87	3.9	28.97	4.1	16.49	27.36	2.49	1.61
3.6	25.66	4.0	30	4.34	16.91	28.23	2.57	1.77
3.7	26.54	4.1	31.27	4.73	17.82	29.19	2.65	2.08
3.8	27.43	4.2	32.88	5.45	19.87	30.17	2.74	2.71
3.9	28.97	4.3	34.2	5.23	18.05	31.87	2.90	2.33
4	30	4.4	36.02	6.02	20.07	33.00	3.00	3.02
4.1	31.27	4.5	37.68	6.41	20.50	34.40	3.13	3.28
4.2	32.88	4.6	39.76	6.88	20.92	36.17	3.29	3.59
4.3	34.2	4.7	41.93	7.73	22.60	37.62	3.42	4.31
4.4	36.02	4.8	44.25	8.23	22.85	39.62	3.60	4.63
4.5	37.68	5.0	49.19	11.51	30.55	41.45	3.77	7.74
4.6	39.76	5.1	52.05	12.29	30.91	43.74	3.98	8.31
4.7	41.93	5.2	54.85	12.92	30.81	46.12	4.19	8.73
4.8	44.25	5.3	58.09	13.84	31.28	48.68	4.43	9.42
4.9	46.51	5.4	61.6	15.09	32.44	51.16	4.65	10.44
5	49.19	5.5	65.28	16.09	32.71	54.11	4.92	11.17
5.1	52.05	5.6	69.67	17.62	33.85	57.26	5.21	12.42
5.2	54.85	5.7	74.31	19.46	35.48	60.34	5.49	13.98
5.3	58.09	5.8**	70.19	12.1	20.83	63.90	5.81	6.29
5.4	61.6	5.8**	70.19	8.59	13.94	67.76	6.16	2.43
5.5	65.28	5.8**	70.19	4.91	7.52	71.81	6.53	-1.62
5.6	69.67	5.8**	70.19	0.52	0.75	76.64	6.97	-6.45
5.7	74.31	5.8**	70.19	-4.12	-5.54	81.74	7.43	-11.55
5.8	70.19	5.8**	70.19	0	0.00	77.21	7.02	-7.02

* Values rounded up due to the grouping of low KS2 fine-levels

**Values reported as if there contained less than a 10% error due to the 'ceiling effect'

***Calculated by deducting the absolute error in KS4 ratings from the absolute error in KS2 rating.

The table above illustrates that errors within students' KS2 data tend to translate into larger discrepancies in the Attainment 8 estimates, both in absolute terms and relative to the original KS4 value. Were a student with an actual KS2 prior-attainment level of 4.0 to receive a fine-level rating of

4.4 for example this would manifest as a 6.02 point error in the students' Attainment 8 estimate, which amounts to 20.1% over-statement of their true Attainment 8 score.

The only exceptions to this are values that are affected by the floor/ceiling effect or the grouping of very low KS2 fine-levels (see Section 4.3.2 for an explanation of these effects).

The effect also occurs in the opposite direction, when students' aptitude is underestimated[26].

10.4. Discussion

The results of this analysis suggest that errors in students' prior-attainment data may have a substantial impact upon the validity of their value-added scores. Two questions therefore need to be addressed, why do errors in students' prior-attainment data have a greater impact than errors in their final attainment score and what are the implications for school-level ratings?

It is argued here that the most likely explanation for the discrepancy is the differential effect that schools' have upon students with differing prior-attainment levels. The fact that students with favourable starting points will often pull further ahead during their education, and vice versa (Ready, 2013). This 'fanning out' of scores is illustrated in Figure 5.3.3a. From this perspective, it therefore makes sense that the difference between the mean attainments of two dissimilar groups of students would expand between Key Stage 2 and Key Stage 4, making the consequences of misjudging a student's initial performance level graver than misjudging their final attainment by the same amount. [27]

To address the second question we considered the debates in Chapter 6. We know therefore that any unexplained variance in student performance will be attributed to the school effect by prior-assumption. The critical question then is whether these errors transfer into schools' progress ratings. Intuitively it seems likely that the vast majority of measurement errors will occur at random as Reynolds *et al.*, (2012) argued. Assuming this is the case, then a large proportion of the variance that results from the above phenomenon would cancel itself out when students' individual value-added scores are aggregated to the school-level. That being said there are undoubtedly cases where measurement errors are introduced in a systematic manner that makes them more common in certain types of institution and these error can propagate through the computation (Gorard, 2010a). Students that speak English as a second language, for example, have a tendency to receive favourable progress ratings as their English speaking proficiency often leads to their initial aptitude being under-reported (Thomas *et al.*, 1997a). It follows that this effect will lead to some schools being systematically advantaged or disadvantaged. Further research, however, would be needed to assess the magnitude of any bias that is introduced.

[26] All observations were confirmed by repeating the exercise with the 2015 attainment averages. These tables are not presented as they do not contribute any new information to the discussion.

[27] The astute reader will have noticed that the absolute/relative error in KS4 ratings is higher amongst high-achieving pupils. This is solely because of the way in which our errors were represented. The fact that in all cases it was assumed that there would be a 10% error within students' prior-attainment levels. A 10% error in a KS2 fine-level of 5.8, however, is far greater than a 10% error in a KS2 fine-level of 1.5. In absolute terms these errors would be 0.58 and 0.15 respectively. The readers should bear this in mind when interpreting the results. Likewise when the exercise was repeated to observe the effect of understating students' scores by 10%, the absolute error peaked at a higher amount and the relative error at slightly lower level than that in Table 10.3a (highest absolute error: -22.26, highest relative error: -29.96%).

10.5. Conclusion

This section evaluated whether errors in students' Key Stage 2 and Key Stage 4 attainment data have the same effect. The results suggest that this is not the case. In fact, the impact of KS2 error could be up to 2.5 times greater in some cases. If any of these errors were to be non-random, then the capacity for them to impact upon schools' Progress 8 ratings is substantial. To our knowledge, this is a unique observation that has not been addressed in past research. Further research is therefore warranted to explore the implications.

11. Shallow Regression Analysis

11.1. Chapter Introduction

Since the late 1960's effectiveness researchers have sort to explain why some students and/or schools perform better than others. A popular approach has been to identify factors that correlate with students' raw-attainment and to integrate these into conceptual models of educational effectiveness. In this analysis, one of these frameworks was adapted and used it to identify the factors that have the greatest impact upon schools' Progress 8 results. The validity of the Progress 8 assessments was then judged by whether the differences in schools' outputs were explicable and under the control of schools.

11.2. Method Section

11.2.1. Research sample

i. Characteristics of the sample

Participation in Progress 8 is mandatory for all state-funded mainstream secondary schools in England. In 2018, 3659 schools fell into this category (EduBase, 2018). However, after restricting the sampling frame to exclude Pupil Referral Units and schools that do not educate students from the beginning of Key Stage 3 to the end of Key Stage 4, the population then refers to 2991 schools. From this population we took a convenience sample of 187 schools. More specifically, all of the schools in the sampling frame were identified using Edubase. Each was contacted via email to request that they take part in the study. Non-responses were followed up with a second invitation. All schools took part on a voluntary basis.

62 of these schools, however, had to be excluded from the analysis. Four schools were excluded because they did not have Progress 8 results in specified academic years. Three were omitted because we received multiple responses from members of the same institution. Two schools were removed because they were not intended to be in the sampling frame and three were excluded because the information that the respondent provided could not be matched to a specific school. The remaining 50 schools were excluded because they failed to answer key questions in a questionnaire that was integral to the study's design. This left 125 schools in the study.

17 schools submitted partially completed questionnaires that were included in the analysis. These schools were retained because the extent of missing data was minimal. That is to say, that they contained two or fewer missing data items. All missing data items were replaced with the mean score for the variable to prevent the data-item from affecting the associated regression co-efficient. This is the recommended practice for dealing with small quantities of missing data in regression models (Agresti and Franklin, 2014).

The limitation of convenience samples such as this is that they are particularly vulnerable to selection bias. In other words there is an increased likelihood that there will be differences between the achieved sample of schools and the population they are intended to represent. Under most conditions this reduces the certainty with which researchers can generalise their findings. In this analysis, however, the primary objective was not to describe the distribution of a variable within the overall population. The intention was to test whether the differences in schools' annual Progress 8 ratings and the change in schools' performance ratings over time could be explained by established effectiveness correlates. It was therefore more important to achieve a large sample that would maximise the variation in the assessed variables. Convenience sampling helped to achieve this. The coverage and representativeness of the sample was nonetheless tested and is discussed below.

ii. The representativeness of the sample

In 2018, 150 local authorities were included in the DfE Secondary School Performance Tables. 60 of these were represented in the sample. After restricting the population to exclude independent schools, special schools, pupil-referral units, forms of alternative provision and any establishments that did not educate students from beginning of KS3 to the end of KS4, the remaining institutions displayed the following characteristics (see Table 11.2.1a).

Table 11.2.1a: The types of schools included in the Shallow Regression Analysis

Type of School	Percentage of population	Percentage of sample
Converter-mainstream academies	46.4%	61.6%
Sponsor-led academies	20.8%	6.4%
Community schools	14.1%	16.0%
Voluntary aided schools	8.1%	7.2%
Mainstream foundation schools	7.0%	7.2%
Maintained free schools	2.5%	1.6%
Voluntary controlled schools	1.0%	0.0%
City technology colleges	0.1%	0.0%

The predominant differences were therefore that converter-mainstream academies were slightly over-represented within the sample and sponsor-led academies were under-represented.

These institutions varied in size with the average school entering 156.3 students into the Progress 8 calculation (sd = 60.6). Within the sample the average number of students was 168.6 (sd = 57.5). These figures were therefore comparable, as were the average entry rates which were reported as 95.5% and 96.0% in the population and sample respectively.

The composition of the schools' intakes was also typical (see Table 11.2.1b).

Table 11.2.1b: The composition of school cohorts in 2018

Pupil characteristic	Percentage of population	Mean percentage per school within sample
Disadvantage[28]	27.4%	20.1%
Female	49.7%	52.2%
Non-mobile	96.9%	97.5%
English as additional language	16.3%	11.7%
Special educational needs (and Statement or EHC plan)	2.0%	1.9%
Special educational needs (but no Statement or EHC plan)	11.0%	10.0%

It was therefore concluded that the sample was not biased by the characteristics of the institutions that were included within the study or their intakes.

A more pertinent distinction was that high achieving schools (those with high Attainment 8 scores) and 'effective' schools (those with high progression scores) were slightly over-represented within the sample. The mean average Attainment 8 score of the sampled schools was 50.869, whilst the mean average Attainment 8 score within the population was 47.334. Similarly the mean Progress 8 score was 0.164 (sd = 0.338) within the sample, and 0.013 (sd = 0.449) nationally. As a wide range of Progress 8 scores were represented (range in sample = -1.52 to 1.21, range in population = -1.58 to 1.9), however, the disparity is not believed to have had any substantive implications.

11.2.2. Research design

There are three strands to the analysis; an assessment of schools' 2017 Progress 8 results, an assessment of schools' 2018 Progress 8 results and an assessment of the 2017-2018 change in schools' Progress 8 ratings.

The first two assessments evaluated the relationship between established effectiveness factors and schools' performance ratings at specified moments in time. The results were interpreted based upon the direction of the associations, their magnitude and whether the relationships were consistent with the interactions theorised in academic research. That is to say, the impact that the variables have upon students' raw attainment (see Section 7.3 and Section 11.2.4 for further details). The prospective element of this design prevents post hoc re-interpretation of the results and therefore provides more convincing evidence than the retrospective analyses that dominate educational effectiveness research. The most important consideration, however, was the proportion of variation that was explained by factors that are within and outside of schools' control. Since all extraneous sources of bias have ostensibly been removed by the value-added calculation, these variables should in theory be under schools' control. Several forms of regression modelling were used to evaluate whether this was the case (see next sub-section). The effect of schools' intake, teaching behaviours and policies was considered. A unique contribution of this thesis, however, was that the models also assessed the influence of

[28] The disadvantage variable refers to students proportion of students that are either eligible for free school meals or in care. This is the definition of disadvantage utilised by the Department for Education in the National Pupil Database.

examination-entry differences. That is to say, how much schools were punished for failing to adhere to the DfE's preferred curriculum (see Section 3.3). Whilst it is recognised that the Attainment 8 buckets and subject weightings were designed to incentivise schools' to provide all students with an academically-orientated program of study, it is argued here that if Progress 8 scores are intended to report upon the quality of schools' provisions, then the consequences of curricular deviations should not overwhelm the influence of instructional practices and policies. Otherwise Progress 8 would solely be a model of curricular adherence. This analysis will be the first to report upon the matter.

The third assessment identified the factors that account for the change in Progress 8 results over time, specifically those that explained the differences between schools' 2017 and 2018 performance ratings. Whilst this issues has been investigated before, past research has approached this in a technical/abstract way, where changes in value-added scores were evaluated based on the effect of deviations from the mean prevalence of factors over several years (see, for example, Raudenbush and Bryk, 1986). This analysis therefore provides more direct insight into the pragmatic value of Progress 8 ratings in real world situations. Once again, it was presumed that if Progress 8 provides a valid and reliable measure of school effectiveness, the most influential variants would be under schools' control.

11.2.3. The three parts of each analysis

Part 1: Simple linear models

In the first part of the analyses the relationship between each effectiveness factor and schools' Progress 8 ratings was evaluated, without controlling for the effect of other variables. These assessments followed the standard protocols. After plotting each relationship and concluding that all of the associations could be adequately described by linear functions, Pearson's *r* correlations were calculated. This information helped to identify the factors that could predict the highest percentage of schools' progress results.

It is important however not to confuse correlation with causation. A multitude of factors are thought to influence students' learning and the variance that each explains will overlap. This analysis did nothing to rule out alternative explanations for the correlations that were observed, which means that other factors may be partly or entirely responsible. This was still a useful first step in the identification of key influences however because the sample size prevented the effect of all variables from being modelled simultaneously. Agresti and Franklin (2014, pp. 636), for example, advise researchers that the maximum number of independent variables that are included in regression models should not exceed 1/10 of the number of cases in their sample. This is because 10 observations per variable are needed to ensure that the variation that is attributed to a factor has not occurred by chance. Whilst there is a little play in this figure, most statistical texts suggest a similar cut-off. Furthermore, although it is theoretically possible for extraneous variables to mask the percentage of variance that a factor explains, it is far more likely that controlling for additional influences would reduce the effect that is attributed to the independent variable. In other words, it is very unlikely that the factors that are ascribed low r-squared scores in the analysis are hiding much explanatory power. Even though these models have obvious limitations, they therefore allowed us to assess the potential effects of each variable and discarded the least relevant.

Part 2: Forward regression models

In the second part of the assessment, a multiple-regression model was constructed. This contained the 12 most predictive factors in each analysis.

Variables were selected for this model using forward selection. The procedure began with an empty model. Independent variables were then added into the regression equation starting with the variable that had the greatest association with the dependent variable. In other words, the measure that explained the highest percentage of the variation in Progress 8 ratings. The correlation between the remaining independent variables and the dependent variable was then reassessed, controlling for the variable that had already been entered into the model. This process was repeated until 12 effectiveness factors had been selected. The model was limited to this number because this was the maximum number of dependent variables that was justified by the sample size (see earlier discussion).

The resulting model provides more accurate information than the preceding simple-linear regression models. This is because the presence of statistical controls removes the overlap in the variance that each factor explains. It therefore gives a better indication of factor's causal impact (both collectively and individually).

The methodology however does have some notable weaknesses. Firstly, the models still rely upon correlational evidence. There remains, therefore, a risk that the factors will act as proxies for the sources of variation that have not been controlled. The models also ignore the temporal order of influences (i.e. which events occurred first) and their proximity to classroom interactions. The r-squared scores attributed to the variables are therefore biased in favour of those that were entered into the model first, as all of the overlap between factors is attributed to the first factor to account for the variance.

Part 3: Hierarchical models [29]

In the final stage of the analyses a hierarchical linear multiple-regression model was constructed. Variables were selected for this model using the forward-selection process discussed above. That is to say, that variables were added to the model, one by one, in accordance with their explanatory power[30]. This time, however, some classifications of variable were given preferential treatment. More specifically, the effect of intake differences was modelled before the influence of instructional practices, instructional practices were considered before schools' policies, and schools' policies before their examination entry practices.

To ensure that each of the aforementioned categories of variables were represented the 3 most influential variables from each group were included within the regression equation (after taking into account any factors that had already been entered into the model). This was not an arbitrary number, it was the maximum number of factors that the sample size permitted, divided by the number of categories that were included in the model. 9 factors were therefore modelled in the annual analyses

[29] Note that these models are referred to as being hierarchical because they give preferential treatment to certain classifications of variable. It is important to recall however that all of the data in the analyses was aggregated to the school level before it was entered into the regression models (see Section 8.3). These models are therefore not multi-level in the sense that most researchers would use the term. The title is retained to provide a clear means of distinguishing this specification of model from the previous.

[30] It should be noted, however, that the r-squared scores of factors are grouped during some aspects of the analysis so that the collective influence of each category of variable could be evaluated.

and 12 in the assessment of 2017-2018 change. Factors relating to schools' policies were not modelled in the annual analyses. This decision is justified elsewhere (see Section 11.2.6).

This form of model provides the most defensible account of school and factor effects, as the construction acknowledges the order in which variables will impact upon performance, namely that the differences between school intakes will usually predate any educational effects. It also reflects the pathways through which factors bring about their influence. In other words, the fact that school policies are primarily intended to influence stakeholders' behaviours. The r-squared scores reported for each variable however are still biased in favour of the variables that are entered first.

It is also important to acknowledge though that this ordering of factors has implications for the meaning that should be attributed to each factor. In this format the school-intake factors describe the effect that students' background, personal characteristics and task-related behaviours had upon schools' ratings. The instructional variables identify the impact that school tuition had over and above the effects that were attributable to differences in the school intakes. The school policy factors then ascertained whether the differences in institution's educational policies helped to explain any variation that was not accounted for by the aforementioned effects. Finally, the differences in schools' examination entries model the variation that can only be explained by differences in students' curricular.

The outputs from all specifications of model were then triangulated. This helped the researcher to construct a comprehensive picture of variable's influence that would negate the bias inherent in each representation. The reader is cautioned, therefore, not to place too great an emphasis upon individual associations, especially the r-squared scores of individual variables within the forward and hierarchical regression analyses as these will be heavily influenced by the order of variable entry. The directional effect of variables, the relative explanatory power of each category of variable and the overall percentage of variance that the correlates collectively explained however are more robust statistics that are likely to persist across the analyses.

Multicollinearity checks were also undertaken. These suggested that the association between the independent variables was insufficient to affect the substantive findings of the study. More specifically, the analyses revealed that close correlations (associations of r = 0.8 or higher) only occurred between variables that were intended to operationalise the same construct (i.e. the average number of GCSE qualifications per pupil including and excluding GCSE equivalent qualifications, the overall percentage of absence at each school and the percentage of persistent absentees, the percentage of Year 11 pupils that spoke English as and additional language and the percentage of Progress 8 entrants that spoke English as an additional language, the percentage of unstatemented SEN pupils and the overall percentage of SEN pupils). The one exception was that the schools that entered the highest proportion of their students for EBacc language qualifications also entered a higher percentage of their students for the English Baccalaureate. This is presumably because the former was the least often met criterion of the latter. When interpreting the results of the simple linear regression models it is therefore important to recognise that much of the predictive power that is attributed to the entry rate of EBacc language qualifications will be due to its association with the entry rate for the overall English Baccaluarate. This overlap should not impact upon the results of the multiple regression models, however, as the percentage of EBacc language entries was not modelled in these analyses.

11.2.4. The conceptual framework for modelling

i. The theoretical basis of our models

To enhance the utility of these models, their construction and the pool of operationalised effectiveness factors was based upon the findings of educational effectiveness research. More specifically, Creemers and Kyriakides' Dynamic Model of Educational Effectiveness (2008).

This was viewed as an appropriate foundation for the study because its predecessor, the Comprehensive Model (Creemers, 1994), was widely recognised as one of the most prominent and influential models in the field (Teddlie and Reynold's. 2000). The underlying constructs have therefore been subjected to extensive empirical testing (de Jong *et al.,* 2004; Driessen and Sleegers, 2000; Kyriakides *et al.*, 2000; Kyriakides, 2005a; Kyriakides and Tsangaridou, 2004; Reezigt *et al.*, 1999). The same is true of the Carroll Model (1963), the theory of learning upon which underpins both models.

This framework was useful for several reasons. First, it compiles a list of factors that are theoretically and empirically linked with students' raw attainment. It follows therefore that if Progress 8 is a valid and reliable indicator of school effectiveness, then these factors should help to explain both the difference in schools' performance ratings at specified moments in time, and the changes in schools' ratings over time. The analyses tests whether this is the case. Second, it describes the nature of these effects. Specifically, it specifies whether each factor promotes or hinders performance, whether this relationship is expected to be linear or non-linear and whether it acts directly upon students' learning or through other variables. This information informed the study's methodology and the interpretation of its results, including tiered structure of the hierarchical regression models. Finally, the model clearly states whether each factor can or cannot be influenced by schools. When evaluating whether Progress 8 effectively controls for non-school factors, this feature essentially maps out the variables that must be considered.

ii. Adaptations to the model

This analysis expanded upon the Dynamic Model, however, by evaluating the impact of additional variables. These fell into two categories. The first group refer to student intake characteristics that were not considered by Creemers and Kyriakidies. These influences were identified in other academic studies, suggested by practitioners, or used as proxies for influences that were difficult to operationalise. The second group of variables was used to identify differences in schools' examination entry practices. This is something that was not considered in the dynamic model as it is ordinarily used to evaluate students' learning in one subject-area at a time. A wide variety of qualifications can count towards schools' Progress 8 ratings however so it is reasonable to assume that these differences will impact upon schools' results.

With regards to the new intake variables (see Table 11.2.6a), those which report upon percentage of disadvantage students (a proxy for socio-economic status), and absence rates (which was treated as the inverse of time-on-task) were expected to have negative linear associations with students' learning. This is in line with the effect of the variables that they replaced. All classifications of special educational needs were also expected to have a detrimental effect upon student attainment levels as these individuals must overcome additional barriers in order to master the curricular, though students with Statements or Educational Health and Care plans should theoretically represent the most disadvantaged group. This does not imply that the group will always have the largest effect upon schools' ratings, however, as there are considerably fewer students in this category.

The percentage of non-mobile students and the percentage of students that spoke English as an additional language were assumed to have a positive impact upon schools' ratings. The former expectation was based around the logical assertion that any disruption to students' learning would detract from their academic progress. There are also potential knock-on effects for classmates if material needs to be repeated. The latter association has been established in past research (see, for example, Thomas *et al.*, 1997a). Whilst the underlying reason for this association remains unclear, several mechanisms have been suggested. These include the argument that students' initial language speaking proficiency may, in some cases, cap their Key Stage 2 attainment scores, meaning that their prior-knowledge and intellect are under-reported. This would mean that the students' Key Stage 4 performance would be compared with less-able students, making it seem as though the school had had a greater effect than it had. It is likewise possible that the communication skills of these students develop to a greater extent outside of school, that there are broader benefits to being bi-lingual, or given the overlap with ethnicity, that there are cultural differences in students' aspirations and work-ethic. All justifications, however, refer to extraneous influences that are outside of the schools' control.

It was further presumed that all examination entry variables would have a positive association with schools' Progress 8 scores and that the magnitude of these associations would be dictated by the alignment between the two sets of inclusion criteria. That is to say, that their impact was assumed to be influenced by the number of subject areas that were considered and how directly these refer to the Attainment 8 slots, especially those which can only include specified types of qualifications. The percentage of students' entering the English Baccalaureate, the average number of EBacc slots covered and the average number of Open slots covered were therefore expected to be amongst the most predictive factors. Whilst the influence of the entry rates for Maths and English were expect to exceed impact of the other subject areas. The only exception to this was that the number and percentage of students included in schools' calculations were expected to have weak but negative association with the magnitude of schools' Progress 8 scores as outlined in Gorard *et al.*, (2013).

iii. The dimensions of effectiveness that were considered

One of the unique features of the dynamic model is that all school-related variables are assessed from several perspectives. Five of these are emphasised in the original model; the *frequency* or quantity of specified actions, their *focus* or intention, the *stage* or timing of their implementation, the *quality* of actions/policies and the level of *differentiation* that is used. Collecting such extensive data from a sample of over 100 schools would, however, have been a colossal undertaking. This analysis therefore focused predominantly on the frequency dimension as the most research has been conducted in this area. It also assessed the quality of schools' policies and instructional practices as this complemented the aforementioned dimension. The remaining dimensions were given little consideration. School leaders were however asked to take into account the level of differentiation that took place when they reported upon the quality of policies and/or actions so that these considerations were not excluded entirely. Combining the dimensions of quality and differentiation in this way is valid and is something that was considered by the Creemers' as the model was being developed (Creemers and Kyriakides, 2008). Whilst the model was simplified it therefore took into account the most influential factors from educational effectiveness research. Though the failure to assess the focus and stage dimensions may mean that some influences are overlooked. This possibility is discussed when the results are interpreted.

11.2.5. Data collection

To balance the competing demand for high quality data and coverage of the underlying constructs two sources of information were utilised:

Data on school intakes, examination entries and attainment was collected from the National Pupil Database. This is an extensive database that is updated and maintained by the Department for Education. In theory in contains information on all state-funded schools and should therefore encompass the entirety of the research population. Whilst missing data is an issue, as discussed elsewhere, its coverage is undoubtedly better than we could have been achieved through other means. The data is also subjected to automated validation checks and goes through a two-stage revision process. The only shortfall is that the datasets were not designed specifically for this research project and consequently did not cover all of the intake characteristics that would have been evaluated under ideal conditions. Information of students' perseverance, subject motivation and thinking styles for example was not available. Since the most important differences in school intakes refer to differences in students' physical characteristics, background and prior-attainment, however, the omission was tolerated.

Data on schools' instructional practices and policies was collected using an electronic questionnaire. This was completed by a member of the school-leadership team between March 2018 and July 2018. In the vast majority of cases responses were received from the head-teacher themselves, though checks were performed to ensure that all questionnaires were completed by a suitable individual. A more direct assessment of the classroom-level variables may have been preferable, however, none of the alternatives considered were feasible given the intended sample size. The use of instruments, such as direct observation, teacher diaries and/or lesson plans would also be complicated by the fact that most of the students included in schools' Progress 8 calculations have taken different combinations of subjects and classes. Any attempts to evaluate teachers' instructional behaviour directly would therefore be fraught with difficulty. School-leaders questionnaires were therefore viewed as the best of the options available. It was nevertheless reasonable to expect that school-leaders would have been able to identify whether there had been substantial changes in their schools' provisions

11.2.6. The factors that were considered in each analyses

The following measures were taken from each school. These variables formed the pool of variables which was considered when the three types of formal effectiveness model were constructed.

Table 11.2.6a: The variables that were considered in the Shallow Regression Analyses

Category:	Variable:	Data source	Whether the variable was considered in set of analyses		
			2017 analysis:	2018 analysis:	Change analysis:
Student-intake variables	The overall percentage of absence across the school	NPD	Yes	Yes	Yes
	The percentage of persistent absentees at the school	NPD	Yes	Yes	Yes
	The percentage of Progress 8 entrants that were disadvantaged	NPD	Yes	Yes	Yes
	The percentage of Progress 8 entrants that spoke English as an additional language	NPD	Yes	Yes	Yes
	The percentage of Year 11 student that spoke English as an additional language	NPD	Yes	Yes	Yes
	The percentage of Progress 8 entrants that were female	NPD	Yes	Yes	Yes
	The percentage of Year 11 students with a Statement of Special Educational Need or an Educational Health and Care plan	NPD	Yes	Yes	Yes
	The percentage of Year 11 pupils with special educational needs but no Statement or Educational Health and Care plan	NPD	Yes	Yes	Yes
	The percentage of Year 11 pupils with special educational needs (with or without a Statement or Educational Health and Care plan)	NPD	Yes	Yes	Yes
	The percentage of Progress 8 entrants that were non-mobile	NPD	Yes	Yes	Yes
Instructional variables	The frequency of orientation tasks	Questionnaire	Yes	Yes	Yes
	The frequency of structuring tasks	Questionnaire	Yes	Yes	Yes
	The frequency of questioning	Questionnaire	Yes	Yes	Yes
	The frequency of teacher-modelling tasks	Questionnaire	Yes	Yes	Yes
	The frequency of application tasks	Questionnaire	Yes	Yes	Yes
	The frequency of on-task teacher-student interactions	Questionnaire	Yes	Yes	Yes
	The frequency of on-task student-student interactions	Questionnaire	Yes	Yes	Yes
	The frequency of classroom disruptions	Questionnaire	Yes	Yes	Yes
	The proportion of lesson time that was used for teaching	Questionnaire	No	No	Yes
	The frequency of classroom assessments	Questionnaire	Yes	Yes	Yes
	The quality of teachers' instructional behaviour	Questionnaire	No	No	Yes
	Teachers' coverage of the school curriculum	Questionnaire	No	No	Yes
School policies	School-level quantity of instruction	Questionnaire	No	No	Yes
	The alignment between school curriculum and assessed curriculum	Questionnaire	No	No	Yes
	The quality of the policies regulating teachers' instructional behaviours	Questionnaire	No	No	Yes
	The quality of the policies for evaluating the school teaching policies	Questionnaire	No	No	Yes
	Whether changes to the school teaching policies were based on evaluation data*	Questionnaire	No	No	Yes
	The quality of policies that regulate the school learning environment	Questionnaire	No	No	Yes
	The quality of policies for evaluating the school learning environment	Questionnaire	No	No	Yes
	Whether changes to the school learning environment were based on evaluation data*	Questionnaire	No	No	Yes
Examination entry	The number of students that were entered into the Progress 8 calculations (cohort size)	NPD	Yes	Yes	Yes

variables					
	The percentage of students that were entered into the Progress 8 calculations (coverage)	NPD	Yes	Yes	Yes
	The percentage of Year 11 cohort that entered the EBacc Maths subject area	NPD	Yes	Yes	Yes
	The percentage of Year 11 cohort that entered the EBacc English subject area	NPD	Yes	Yes	Yes
	The percentage of Year 11 cohort that entered the EBacc Science subject area	NPD	Yes	Yes	Yes
	The percentage of Year 11 cohort that entered the EBacc Humanities subject area	NPD	Yes	Yes	Yes
	The percentage of Year 11 cohort that entered the EBacc Language subject area	NPD	Yes	Yes	Yes
	The average number of EBacc bucket slots filled in Attainment 8 per pupil	NPD	Yes	Yes	Yes
	The average number of Open bucket slots filled in Attainment 8 per pupil	NPD	Yes	Yes	Yes
	The percentage of Year 11 with entries in all English Baccalaureate Subject Areas	NPD	Yes	Yes	Yes
	The average number of GCSEs per pupil (not including equivalencies)	NPD	Yes	Yes	Yes
	The average number of GCSEs per pupil (including equivalencies)	NPD	Yes	Yes	Yes

*Measure refers to changes implemented over the last 12 months. When changes did not take place school leaders specified whether the decision not to change was based upon evaluation data.

As previously specified these variables were intended to operationalise the frequency and quality dimensions of the Dynamic Model of Educational Effectiveness, as well as extraneous influences that may impact upon schools' ratings.

All 42 variables, however, were not considered in each of the three sets of analyses. Only 31 variables were assessed in 2017 and 2018 assessments. This ensured that school-leaders' questionnaires could be completed within an appropriate time frame.

As a result of these concessions, the annual models focused predominantly upon the influence of school intake, the frequency of effective teaching behaviours and schools' examination entry protocols. The effect of school-level teaching policies, policies for developing an effective learning environment and the mechanisms for evaluating of school policies only considered in the analysis of 2017-2018 changes.

No attempt was made to operationalise factors from the system level of the Dynamic Model. That is to say, that influences such as the national/regional education policies, the national/regional evaluation systems and the wider environment of schools were not assessed. This is appropriate as educational policy is reasonably centralised in England. Most of the aforementioned influences would therefore be constant across all schools. Moreover, the few effectiveness studies that have investigated the impact of middle-level bodies, such as Local Educational Authorities, have concluded that these types of institutions have little to no impact on students' progression (see, for example, Tymms *et al.*, 2008). This is often attributed to the fact that they are not ordinarily in a position to directly influence schools' practices (Creemers and Kyriakides, 2008). Neither of these omissions should therefore impact upon the results. The only factors that would have been evaluated under ideal circumstances are those that report upon the support that schools receive from external stakeholders and the expectations of the local community. These may differ from region to region and mean that some schools operate in more favourable conditions than others. A common claim, for example, is that high aspirations can create an achievement press that encourages high attainment (Valverde and Schmidt, 2000).

11.2.7. Measurement scales

All NPD datasets contained either scale-level data or dichotomous data that could be converted into this format. These variables could therefore be entered into the regression analysis without modifying the original scales. The only exception being suppressed data which was entered as the mid-value of the suppressed range.

Data from questionnaires was reported in one of three formats:

1. An ordinal scale identifying the frequency with which the actions took place
   (1 = almost never to 5 = very frequently)

2. An ordinal scale identifying whether the frequency/quality of the actions had changed since last academic year
   (In most cases these values were calculated by subtracting the schools' frequency/quality rating from 2018 from the 2017 rating. Factors that were only assessed in the change analysis were, however, scored on a 1-5 scale where 1= a large decease and 5 = a large increase.)

3. Dichotomous option boxes (yes/no)

For the purpose of this analysis however all scales were converted to, or treated as, interval-level. This is not an issue for the variables that relied upon dichotomous data. The validity of converting the ordinal data however depends upon legitimacy of treating all measurement increments as identical. The analysis also relies on school leaders making similar decisions as to what constitutes 'frequent' and 'very frequent'. Neither is assumed to align precisely. The conversion will therefore introduce some noise into the analysis.

11.3. Results

Prior to modelling, scatter-graphs were used to confirm that all relationships could be adequately represented by linear functions. These are too numerous to be present here, however, they are available in Appendix C.

Part 1: Individual regression analyses (with no control variables)

2017 Analysis:

Table 11.3a describes the association that each independent variable had with school performance.

Table 11.3a: The linear association that existed between each independent variable and schools' Progress 8 scores in 2017

Rank	Variable	Direction of association	R-squared
1	The percentage of Year 11 pupils entering all English Baccalaureate subject areas	Positive	0.330
2	The overall percentage of absence across the school	Negative	0.322
3	The percentage of persistent absentees across the school	Negative	0.302
4	Average number of EBacc slots filled per pupil in Attainment 8 measure	Positive	0.248
5	The percentage of pupils Year 11 pupils entering the English Baccalaureate language subject area	Positive	0.242
6	Frequency of classroom disruptions	Negative	0.226
7	Average number of open slots filled per pupil in Attainment 8 measure	Positive	0.214
8	The percentage of Progress 8 pupils that were disadvantaged	Negative	0.178
9	The average number of GCSE and equivalent entries per pupil	Positive	0.151
10	The percentage of Year 11 pupils entering the English Baccalaureate maths subject area	Positive	0.134
11	The percentage of Year 11 pupils entering the English Baccalaureate humanities subject area	Positive	0.131
12	The average number of GCSE entries per pupil (not including equivalencies)	Positive	0.126
13	The percentage of Year 11 pupils entering the English Baccalaureate English subject area	Positive	0.097
14	Frequency of teacher-modelling tasks	Positive	0.079
15	The percentage of non-mobile Progress 8 entrants	Positive	0.065
16	Frequency of orientation tasks	Positive	0.062
17	The percentage of Progress 8 pupils that spoke English as an additional language	Positive	0.051
18	The percentage of Year 11 pupils included in the Progress 8 measure	Negative	0.050
19	The percentage of Year 11 pupils that spoke English as an additional language	Positive	0.046
20	The percentage of Year 11 pupils entering the English Baccalaureate science subject area	Positive	0.045
21	Frequency of on-task student-student interactions during lessons	Positive	0.044
22	The percentage of Progress 8 entrants that were female	Positive	0.039
23	Frequency of on-task teacher-student interactions during lessons	Positive	0.037
24	Frequency of application tasks	Positive	0.027
25	The number of Year 11 pupils included in the Progress 8 measure	Positive	0.025
26	Frequency of questioning tasks	Positive	0.021
27	Frequency of classroom assessments	Positive	0.016
28	Frequency of structuring tasks	Positive	0.008
29	The percentage of Year 11 pupils that had a Statement of SEN or EHC plan	Negative	0.002
30	The percentage of Year 11 with SEN (with or without Statement/EHC plan)	Negative	0.002
31	The percentage of Year 11 pupils that had SEN but no Statement or EHC plan	Negative	0.001

* The shading of column 2 distinguishes between the 3 categories of variable assessed in this analysis; intake variables (darkest), examination entry variables (light grey) and classroom instructional practices (clear).

**The shading of column 3 signifies whether the direction the variables association with schools' Progress 8 ratings was consistent with our expectations (Clear = consistent, shaded = inconsistent). All anomalies are discussed in later sections.

Most of the effectiveness factors from the 2017 model were able to predict meaningful proportions of the variation in schools' Progress 8 results, though the effect of individual factors varied substantially (range in variance explained: 33.0% to 0.1%).

The average variance attributed to each category of variable

Table 11.3b: The average variance in Progress 8 scores attributed to each category of variable in 2017

Category of variable	Average r-squared of group	Average percentage explained
Student intake variables	0.101	10.1%
Instructional variables	0.058	5.8%
School Policies	Not assessed	Not assessed
Examination entry variables	0.149	14.9%

Schools' examination entry practices

The factors which explained the most variation described aspects of the schools' examination entry practices (see Table 11.3b). In other words, the types of qualifications that students entered. On average, these factors accounted for 14.9% of the difference in schools' scores (range: 33.0% to 2.5%). To help interpret their effect, these measures were collated into five sub-groups based on their conceptual similarity.

Sub-group 1: The percentage of Year 11 pupils entering all English Baccalaureate subject areas (1 variable).

Sub-group 2: The average number EBacc and Open Attainment 8 slots filled by the Year 11 pupils at each school (2 variables).

Sub-group 3: The average number of GCSEs that students entered, with or without the inclusion of equivalencies (2 variables).

Sub-group 4: The percentage of students that were entered for specific subject areas (5 variables).

Sub-group 5: The number and percentage of pupils included in the Progress 8 calculation.

The following observations were then made:

The percentage of Year 11 pupils entering all English Baccalaureate subject areas was the most influential sub-group. This factor had a positive relationship with performance that predicted 33.0% of the variation in schools' ratings.

The second most influential sub-category identified how many of the EBacc and Open Bucket slots had been filled. The effect was once again positive and explained, on average, 23.1% of the variance in school performance. The number of EBacc slots, however, had more predictive power (24.8% of variance explained) than the number of Open slots that students filled (21.4% of variance explained).

Information on the total number of qualifications that students entered was the next most informative sub-group. In both cases schools that entered students for a greater number of qualifications tended to receive higher Progress 8 scores. The average number of GCSE and equivalent qualifications, however, explained slightly more variation (15.1%) than the average number of GCSE qualifications excluding equivalencies (12.6%). Their average effect size was therefore 13.9%. [31]

The entry rates of students in specific subject areas were also meaningful predictors of success. All associations were positive and accounted for an average of 13.0% of the variation in schools' Progress 8 ratings. More specifically, the percentage of students entering EBacc Science explained 4.5%, the entry rate for English qualifications explained 9.7%, Humanities 13.1% and Maths 13.4%. The percentage of students from each school that entered an EBacc Language qualification however had considerably greater predictive power, explaining 24.2% of the difference in schools' Progress 8 results. Interestingly, this final variable therefore explained more variation than the number of open slots mentioned above.

Lastly, the number and percentage of Year 11 students that were included in the schools' Progress 8 calculations had the least impact upon schools' 2017 Progress 8 ratings. On average these factors explained just 3.8% of the variation in schools' ratings. The coverage of the cohort however had a negative relationship with schools' performance and proved to be the more effective predictor (5.0% of variance predicted). Cohort size had a positive relationship with school residuals and explained less of the variation in schools' ratings (2.5%). This discrepancy will be discussed further in the discussion section.

Differences in school intakes

Difference in school intakes were the second greatest predictor of schools' 2017 performance explaining, on average, 10.1% of the variation in Progress 8 ratings. The variation in this figure however was once again dramatic with factors explaining as much as 32.2% or as little as 0.1%.

Student absence was by far the most influential factor. Both of these measures correlated negatively with schools' ratings and explained an average of 31.2% their variance. In other words, the schools with high levels of absence performed worse on average than other schools. In this instance, however, the percentage of persistent absentees explained 2% more variation than the overall absence rate of the school, making it the better predictor.

The percentage of disadvantaged students was the next most influential variable. This had a strong negative relationship with performance meaning that schools with disadvantaged intakes were consistently given lower ratings. Differences in students' socio-economic resources were thus able to explain 17.8% of the variance in schools' ratings.

Student mobility was also a meaningful predictor of school performance. Our analysis shows that there was a positive correlation between the percentage of Progress 8 entrants from each school that have been educated at the same institution for two academic years and school performance. This relationship explained 6.5% of the variation in schools' ratings.

[31] The reader should also note that a very slight inverted-U relationship was found between the average number of GCSE and equivalent entries per school and schools' Progress 8 scores from 2017. The best fitting quadratic function however only accounted for an additional 0.9% of the variation in schools' scores.

The next best predictors of academic performance were the measures that evaluated the proportion of students that spoke English as an additional language. The percentage of Progress 8 entrants speaking English as an additional language explained 5.1% and the percentage of Year 11 students 4.6%, making the average effect size for this factor 4.9%. The relationship between this factor and school performance was a positive one, meaning that the schools with higher proportions of students that spoke English as an additional language outperformed other schools.

Another well-established correlate of academic success is gender. At secondary level the progress of female students tends to exceed that of males. This finding is reflected here. The correlation between the percentage Progress 8 entrants that were female and schools' ratings explained 3.9% of the variation in the 2017 scores.

The final factor evaluated was the proportion of the Key Stage 4 cohort that had special educational needs. Three related measures of this variable were taken; the percentage of students with special educational needs, the percentage of students with special educational needs and a Statement or Educational Health and Care plan, and the percentage of students with special educational needs without a Statement or Educational Health and Care plan. Schools with a high percentage of students with special educational needs performed slightly worse in the assessment, however, the three factors explained just 0.2%, 0.2% and 0.1% of the difference is schools' ratings respectively. The average effect of this category of variable was therefore 0.2% (1dp).

Instructional practices

Information on the use of 9 key instructional variables also helped to explain the differences in schools' performance (average variance explained by factors = 5.8%, range = 22.6% to 0.8%).

The frequency of classroom disruptions was the most effective predictor explaining 22.6% of the variance in schools' value-added results. This had a negative relationship with performance meaning that the schools with the poorest classroom behaviour tended to have lower Progress 8 ratings.

The frequency of the other instructional practices correlated positively with performance. The frequency of teacher-modelling tasks explained 7.9% of the results and orientation tasks 6.2%. Student-student interactions and teacher-student interactions explained comparable proportions of schools' scores, with the former being the slightly better predictor. These factors explained 4.4% and 3.7% of the variation respectively. Application tasks explained a lesser proportion of 2.7%, whilst questioning accounted for 2.1% of the results and classroom assessment 1.6%. The frequency of structuring tasks explained the least variation accounting for just 0.8% of the variation in schools' outcomes.

Most instructional variables therefore had comparable effect sizes, the only exception being the frequency of classroom disruptions which explained a considerably higher proportion of the variance in schools' results.

It is also worth restating that the scatter graphs of the aforementioned relationships showed no evidence of non-linear effects, i.e behaviours that initially enhance effectiveness but have a lesser or detrimental effect when overused.

2018 Analysis:

The results of the second analysis were comparable to the first (see Table 11.3c).

Table 11.3c: The linear association that existed between each independent variable and schools' Progress 8 scores in 2018

Rank	Variable	Direction of association	R-squared
1	Average number of open slots filled per pupil in Attainment 8 measure	Positive	0.389
2	Average number of EBacc slots filled per pupil in Attainment 8 measure	Positive	0.366
3	The percentage of persistent absentees across the school	Negative	0.315
4	The overall percentage of absence across the school	Negative	0.293
5	The percentage of pupils entering all English Baccalaureate subject areas	Positive	0.283
6	The percentage of Year 11 pupils entering the English Baccalaureate language subject area	Positive	0.251
7	The average number of GCSE entries per pupil (not including equivalencies)	Positive	0.244
8	The percentage of Year 11 pupils entering the English Baccalaureate maths subject area	Positive	0.221
9	The percentage of Year 11 pupils entering the English Baccalaureate English subject area	Positive	0.202
10	The average number of GCSE and equivalent entries per pupil	Positive	0.171
11	The percentage of Progress 8 pupils that were disadvantaged	Negative	0.166
12	The percentage of Year 11 pupils entering the English Baccalaureate science subject area	Positive	0.097
13	The percentage of Progress 8 pupils that spoke English as an additional language	Positive	0.068
14	The percentage of Progress 8 entrants that were female	Positive	0.068
15	The percentage of Year 11 pupils that spoke English as an additional language	Positive	0.057
16	Frequency of orientation tasks	Positive	0.056
17	Frequency of on-task teacher-student interactions during lessons	Positive	0.055
18	The percentage of Year 11 pupils that had a Statement of SEN or EHC plan	Negative	0.043
19	The percentage of Year 11 pupils entering the English Baccalaureate humanities qualification	Positive	0.040
20	The percentage of Year 11 pupils included in the Progress 8 measure	Negative	0.036
21	Frequency of classroom disruptions	Negative	0.032
22	Frequency of on-task student-student interactions during lessons	Positive	0.026
23	The percentage of non-mobile Progress 8 entrants	Positive	0.025
24	Frequency of teacher-modelling tasks	Positive	0.021
25	Frequency of structuring tasks	Positive	0.021
26	Frequency of application tasks	Positive	0.019
27	The number of Year 11 pupils included in the Progress 8 measure	Positive	0.007
28	The percentage of Year 11 with SEN (with or without Statement/EHC plan)	Negative	0.006
29	Frequency of questioning tasks	Positive	0.006
30	The percentage of Year 11 pupils that had SEN but no Statement or EHC plan	Negative	0.001
31	Frequency of classroom assessments	Positive	0.000

Most effectiveness factors explained a meaningful proportion of schools' value-added results. This time however the variables predicted between 38.9% and 0.0% (1dp) of the percentage of variation in schools' scores.

The average variance attributed to each category of variable

Table 11.3d: The average variance in Progress 8 scores attributed to each category of variable in 2018

Category of variable	Average r-squared of group	Average percentage explained
Intake variables	0.104	10.4%
Instructional practices	0.026	2.6%
School Policies	Not assessed	Not assessed
Examination entry variables	0.192	19.2%

Schools' examination entry practices

In 2018, the differences in schools' examination entries predicted an average of 19.2% of the variation in school outcomes (range: 38.9% to 0.7%). This made them the most influential group of variables in the analysis (see Table 11.3d).

This time, however, the average number of Open and EBacc slots filled was the most informative sub-group of variables. Both factors correlated positively with performance, meaning that the schools which complied with Progress 8 entry criteria tended to score better on average than schools that did not and these relationships explained 38.9% and 36.6% of the variance in schools' scores respectively. The average variance predicted by these factors was therefore 37.8%.

The percentage pupils entering all English Baccalaureate subject areas once again had a strong positive relationship with performance predicting 28.3% of outcomes.

The number of GCSEs entered (including and excluding equivalencies) was the next most informative sub-category. These variables had a positive association with performance that explained an average of 20.8% of the variation in schools' ratings. In 2018, however, the average number of GCSEs was more informative than the average number of GCSE and equivalent qualifications taken by pupils from each school. The individual measures explained 24.4% and 17.1% of the variation is Progress 8 scores respectively[32].

Knowledge of differences in the entry rate of each subject explained a notable portion of the results. These correlations were all positive but varied in magnitude. Differences in the proportion of students studying a modern foreign language proved to be the most important variable, explaining 25.1% of the differences in schools' annual scores. Whilst the entry rates in Maths, English and Science explained 22.1% 20.2% and 9.7% of the variation in ratings respectively. The least informative entry rate however was the percentage of students from each school entering the EBacc Humanities subject area, with differences in the entry rate of this subject explaining just 4.0% the variation. The average effect size of the sub-group was 16.2%.

The least influential variables in this class identified the size of each cohort and the coverage of the Progress 8 measure. These factors explained 0.7% and 3.6% of the variance in schools' performance respectively, making the average effect size of the sub-group 2.2%. The two variants, however, had opposing relationships with performance, with the number of Progress 8 entrants per school having a

[32] The reader should also note that slight inverted-U relationships were found between the average number of GCSE qualifications (including equivalencies) and schools' Progress 8 scores, and between the average number of GCSE qualifications (excluding equivalencies) and Progress 8 scores. The best fitting quadratic functions however only accounted for an additional 0.4% and 0.2% of the variation in schools' scores respectively.

positive association with schools' value-added ratings and the percentage of pupil included within the Progress 8 measure a negative one.

Differences in school intakes

Differences in school intakes had substantial predictive power. On average these variables accounted for 10.4% of the variation in schools' performance ratings (range: 31.5% to 0.1%).

The two measures of student absence, the percentage of persistent absentees at the school and the overall percentage of absence at the school, both had strong negative correlations with performance. These relationships were able to explain 31.5% and 29.3% of schools' results respectively, making the average effect size of factor 30.4%. This is almost twice the explanatory power of any of the other intake characteristics.

The percentage of disadvantaged pupils was the second most influential variable. This correlated negatively with performance and explained 16.6% of the variance in schools' scores.

The percentage of female students was also important. This factor correlated positively with school outcomes and explained 6.8% of the variation.

Two measures assessed the proportion of students that spoke English as an additional language; the percentage Progress 8 pupils that spoke English as an additional language and the percentage of Year 11 pupils that spoke English as an additional language. These explained 6.8% and 5.7% of the variation in schools' ratings respectively, making the average effect of this factor 6.3%. In both analyses, therefore, schools with a higher percentage of students that spoke English as a secondary language performed better, on average, than schools with a lower percentage of these students.

Higher than average ratings were also more common amongst schools with a high percentage of non-mobile students. That is to say, when the majority of the schools' Year 11 cohort had been educated at the same institution for at least two academic years. This factor accounted for 2.5% of the variation in schools' scores when the effect of other variables was not controlled.

The final and least influential intake bias evaluated was the effect of having students with special educational needs in a school cohort. All three related measures of this factor had a negative association with performance. In 2018, however, the percentage of students with a Statement or Educational Health and Care plan explained more variation (4.3%) than the overall percentage of students with special educational needs (0.6%) or those that had special educational needs but no plan (0.1%). The average percentage that these measures could predict was therefore 1.7%.

Whilst the most important distinction between school intakes was therefore the difference in student absence rates, this analysis indicates that the composition of school cohorts may have also have substantial effects upon schools' performance ratings. In particular the percentage of students from disadvantaged backgrounds.

Instructional practices

Differences in teaching practices had the least predictive power of any of the groups discussed thus far. The average variance explained by these factors was 2.6% with effect sizes ranging from 5.6% to 0.0% (1dp).

All nine instructional variables correlated positively with performance except for the frequency of classroom disruptions which was less prevalent in effective schools. Orientation tasks explained a higher proportion of the variance in schools' results than any other teaching practice, accounting for 5.6% of the variation in schools' results. The frequency of teacher-student interactions, student-student interactions and classroom disruptions were also influential explaining 5.5%, 2.6% and 3.2% of the scores respectively. The frequency of teacher-modelling and structuring tasks had a modest influence upon performance, each accounting for 2.1% of the variation in the dependent variable. As did the frequency of application tasks which predicted 1.9% of scores. The frequency of questioning and classroom assessment, however, only accounted for a negligible proportion of scores, helping to explain 0.6 and 0.0% (1dp) of schools performance ratings respectively.

Change Analysis:

This analysis investigated whether the change in schools' Progress 8 ratings between 2017 and 2018 could be explained by the changes in key effectiveness factors during the same time period. The results suggest that this is the case.

It was immediately apparent, however, that the factors in this analysis accounted for less than half of the variation that they explained in the annual analyses.

Table 11.3e: The linear association that existed between the change in each independent variable (2017-2018) and the change in schools' Progress 8 scores (2017-2018)

Rank	Variable	Direction	R-squared Score
1	Change in the average number of open slots filled per pupil in Attainment 8 measure	Positive	0.160
2	Change in the percentage of Year 11 pupils entering the English Baccalaureate English subject area	Positive	0.089
3	Change in the average number of GCSE and equivalent entries per pupil	Positive	0.073
4	Change in the quality of the policies for evaluating the school teaching policies	Positive	0.048
5	Change in the average number of EBacc slots filled per pupil in Attainment 8 measure	Positive	0.037
6	Change in proportion of lesson time that was used for teaching	Positive	0.031
7	Change in the percentage of persistent absentees across the school	Negative	0.027
8	Change in the frequency of structuring tasks	Positive	0.025
9	Change in the percentage of Year 11 with SEN (with or without Statement/EHC plan)	Negative	0.02
10	Change in the quality of teachers' instructional behaviour	Positive	0.019
11	Change in the frequency of classroom assessments	Positive	0.019
12	The alignment between the school curriculum and the assessed curriculum	Positive	0.019
13	Change in the percentage of Year 11 pupils entering the English Baccalaureate maths subject area	Positive	0.018
14	Change in the average number of GCSE entries per pupil (not including equivalencies)	Positive	0.014
15	Change in the percentage of non-mobile Progress 8 entrants	Negative	0.012
16	Change in the percentage of Year 11 pupils that had a Statement of SEN or EHC plan	Negative	0.012
17	Whether changes to the school teaching policies were based upon evaluation data	Positive	0.012
18	Change in the frequency of on-task student-student interactions during lessons	Positive	0.011
19	Change in the quality of the policies for evaluating the school learning environment	Positive	0.011
20	Change in the percentage of Progress 8 pupils that spoke English as an additional language	Positive	0.011
21	Change in the percentage of Year 11 pupils that had SEN but no Statement or EHC plan	Negative	0.011
22	Change in the frequency of teacher-modelling tasks	Positive	0.01
23	Change in the percentage of Year 11 pupils entering the English Baccalaureate science subject area	Positive	0.009
24	Change in the overall percentage of absence across the school	Negative	0.007
25	Change in the percentage of Progress 8 pupils that were disadvantaged	Negative	0.007
26	Change in the frequency of on-task teacher-student interactions during lessons	Positive	0.006
27	Whether change in the SLE policy were based upon evaluation data	Positive	0.006
28	Change in teachers' coverage of the school curriculum	Positive	0.005
29	Change in the percentage of Progress 8 entrants that were female	Negative	0.004
30	Change in the percentage of Year 11 pupils included in the Progress 8 measure	Positive	0.004
31	Change in the percentage of Year 11 pupils that spoke English as an additional language	Positive	0.004
32	Change in the frequency of orientation tasks	Positive	0.003
33	Change in the quality of the policies on the school learning environment	Positive	0.003
34	Change in the percentage of pupils entering all English Baccalaureate subject areas	Positive	0.002
35	Change in the frequency of classroom disruptions	Positive	0.002
36	Change in the percentage of pupils Year 11 pupils entering the English Baccalaureate language subject area	Positive	0.001
37	Change in the frequency of questioning tasks	Positive	0.001
38	Change in the frequency of application tasks	Negative	0.001
39	Change in the quality of the policies regulating teachers instructional behaviour	Positive	0.001
40	Change in the percentage of Year 11 pupils entering the English Baccalaureate humanities subject area	Negative	0.000
41	Quantity of instruction provided by the school policies	Negative	0.000
42	Change in the number of Year 11 pupils included in the Progress 8 measure	Positive	0.000

*Direction of association refers to the change in Progress 8 score associated with an increase in each variable.

The average variance attributed to each category of variable

Table 11.3f: The average variance in Progress 8 scores explained by each category of variable in the change analysis

Category of variable	Average r-squared of group	Average percentage explained
Changes in schools' intake	0.012	1.2%
Changes in schools' instructional practices	0.011	1.1%
Changes in schools' policies	0.012	1.2%
Changes in schools' exam entry patterns	0.034	3.4%

Changes in schools' examination entry practices

On average this category of variables explained 3.4% of the changes in Progress 8 scores (see Table 11.3f). This may seem like a small effect size, however, the variance ascribed to individual measures varied from 16.0% to 0.0% (1dp).

The most influential sub-group in the analysis was the change in schools' coverage of the EBacc and Open slots. Both of factors had positive correlations with the changes in schools' performance. Changes in the average number of Open slots filled however possessed far greater predictive power, explaining 16.0% of the variance in schools' Progress 8 ratings, in comparison to the 3.7% explained by the EBacc slots. The average effect size of this group of factors was therefore 9.9%.

The next most influential sub-group was the change in the average number of GCSE entries per pupil at each school. Two measures of this were taken, the difference in the number of GCSEs per pupil including equivalencies and the difference in the average number of pupils per school excluding equivalencies. These variables predicted 7.3% and 1.4% of change in Progress 8 scores respectively, making the mean variance explained by this subcategory 4.4 %. Both factors had a positive correlation with performance meaning that increases in the GCSE entry rates were associated with higher than average Progress 8 scores.

Changes in the entry rates of the individual subject-areas also correlated with changes in schools' Progress 8 scores. Specifically, differences in the percentage of pupils entering EBacc English from each school explained 8.9% of the variation in scores over time. Variations in the entry rate of maths and science explained lesser but still meaning proportions of 1.8% and 0.9% respectively, whilst changes in the percentage of pupils with EBacc language entries explained 0.1% of the variation in scores. The least influential subject area however was EBacc humanities which explained 0.0% (1dp) of the variance. The average percentage of variance explained by the change in the entry rate of individual subject areas was therefore 2.3%. It should be noted, though, that whilst the four subjects with the greatest associations with the changes in school performance had positive correlations, the increases in the entries for the humanities subject area were associated with decreases in Progress 8 scores.

Changes in the size of schools' Year 11 cohort and changes in schools' coverage of the Progress 8 measure both had a positive correlation with the changes in schools' performance ratings. Of the two variables, differences in percentage of Year 11 pupils entered into the Progress 8 calculation had the greater association with performance accounting for 0.4% of the variation in school scores, whilst the changes in the number of Progress 8 entrants per school explained 0.0% (1dp). The mean percentage of variance explained by these variables was therefore 0.2%.

The least predictive category of variable in this analysis however was the change in the percentage of school cohorts that entered all English Baccalaureate subject areas. This accounted for just 0.2% of the change in schools' scores. This is a notable departure from the relationship observed in the two annual analyses which is interpreted within the discussion section.

The results therefore showed that despite the decrease in r-squared scores, some of the changes in schools' examination entry procedures were still able to explain substantial proportions of the variation in schools' Progress 8 scores. In fact, the effect sizes of these variables appear even larger relative to the variance that was explained by other aspects of the schools' provisions (see below).

Changes in school intakes

Changes in school intakes helped to predict an average of 1.2% of the changes in Progress 8 scores, which makes them the third most informative group of variables in this analysis. Unlike the schools' examination entries, however, the effect attributed to each factor was similar (range in percentage of variance explained: 2.7% to 0.4%).

Changes in absence rates of each school were the most effective predictive factors, with the change in the percentage of persistent absentees at the school explaining 2.7% of the changes in schools' Progress 8 ratings, and the change in the overall percentage of absences accounting for 0.7%. The average effect size of the absence variables was therefore 1.7%. These variables had a negative relationship with performance meaning that schools with the increased rates of absence tended to receive lower Progress 8 scores than in their previous assessment.

Changes in the proportion of students with special educational needs (SEN) was the next most important indicator. The results showed that school cohorts that had increased levels of SEN students made less progress on average than their predecessors did during the preceding academic year. More specifically, the change in the overall percentage of SEN students in a schools' Year 11 cohort explained 2.0% of the variation in schools' scores. Whilst the change in the percentage of students with SEN and a Statement or EHC plan, and the change in the percentage of students with SEN without a Statement or EHC plan explained 1.2% and 1.1% of the changes respectively. On average these variables therefore accounted for 1.4% of the variation in schools' performance ratings.

The third most influential factor was students' mobility rates. 1.2% of the changes in schools' scores could be explained by acknowledging changes in the proportion of students that had been educated at their current school for at least two academic years. This association was negative, however, meaning that schools which had increased percentages of non-mobile pupils during the 2018 Progress 8 assessments tended to receive lower scores than they did in the previous academic year.

Differences in the percentage of students that spoke English as an additional language also predicted a small percentage of the variation in schools' scores. With changes in the percentage of Progress 8 entrants accounting for 1.1% of the variation and changes in the percentage of Year 11 pupils 0.4%. The average effect size of these factors was therefore 0.8%. Both factors had a positive correlation with the changes in schools' performance rating.

In this analysis a negative association was observed between changes in percentage of disadvantaged students that were included in schools' Progress 8 calculations and the deviation in their scores. In other words, schools which admitted a greater proportion of disadvantaged students in 2018 than they did in 2017 received lower Progress 8 ratings on average than in the preceding year. Conversely, schools with lower percentage of disadvantaged students tended to receive higher ratings than they did in 2017. This correlation, however, only explained 0.7% of the variation in schools' ratings.

Finally, the least predictive of the intake variables was the change in the percentage of female students. This factor explained 0.4% of the changes in schools' performance ratings. Interestingly, the direction of the association was inconsistent with the relationship found in the annual analyses. Specifically, when there was a higher percentage of girls in the 2018 cohort, schools' rating tended to decreased.

Changes to instructional practices

Changes in schools' instructional practice explained on average 1.1% of the change in schools' year-to-year performance ratings (range: 3.1% to 0.1%). On average these factors therefore had the least association with school performance.

The proportion of lesson time used for teaching had the greatest influence upon Progress 8 scores. The variance in this factor had a positive relationship with performance that explained 3.1% of the changes in schools' performance ratings. In other words, there was evidence to suggest that increasing the active learning time leads to small increases the school's Progress 8 ratings.

Changes in the quality of instructional behaviours and teachers' coverage of the school curriculum also predicted the change in school performance, both had positive correlations that explained 1.9% and 0.5% of the change in schools' scores respectively.

In terms of the frequency of specific behaviours, positive correlations were found between the use of all instructional behaviours and the improvements in Progress 8 scores. Changes in the frequency of structuring tasks and classroom assessments accounted for the most variance in schools' scores, explaining 2.5% and 1.9% of changes respectively. More regular on-task student-student interactions, teacher-modelling activities and teacher-student interactions proved beneficial and accounted for 1.1%, 1.0% and 0.6% of the deviations, whilst changes in the use of orientation tasks and questioning accounted for 0.3% and 0.1%. Oddly, classroom disruptions also occurred more often in improving schools. Though the association only predicted 0.2% of the variation in schools' outcomes.

The only exception was the weak association between decreases in the frequency of application tasks and increases in school performance. This explained 0.1% of the variation in Progress 8 change scores.

It would appear therefore that changes in schools' use of instructional practices did impact upon the schools' ratings. On, average, however, the individual correlations between the change in these factors and the change in schools' performance ratings were of a lesser magnitude than the effect sizes attributed to variations in school intake and examination entries.

Changes to school policies

All policy changes had a positive association with the school performance (average variance explained = 1.2%, range = 4.8% to 0.0% (1dp)). That is to say, that the schools which had improved upon their policies tended to experience more favourable changes in their ratings. The one exception to this was that schools' which increased the quantity of available instruction time performed worse overall than those that did not. This association however was very weak (0.0% of variance explained, 1dp).

Details of the individual relationships were described in Table 11.3e. These variables are now grouped to further interpret the results. These four sub-groups equate to the over-arching factors outlined within Dynamic Model of Educational Effectiveness (Creemers and Kyriakides, 2008).

Sub-group 1: Changes to the school teaching policies

Variable 1: Changes to the quantity of instruction policies

Variable 2: Changes in the alignment between the school curriculum and the assessed curriculum

Variable 3: Changes in the quality of the policies governing teachers' instructional behaviours.

Sub-group 2: Changes in the evaluation of the school teaching policies

Variable 1: Changes in the quality of the policies for evaluating the schools' teaching practices

Variable 2: Whether changes to the policies on the school teaching policies were informed by data

Sub-group 3: Changes to the policies on the school learning environment:

Variable 1: Changes in the quality of the policies governing the school learning environment

Sub-group 4: Changes in the evaluation of school learning environment:

Variable 1: Change in the quality of the policies for evaluating the school learning environment

Variable 2: Whether changes to the policies on the school learning environment were based upon data.

The outputs of regression modelling suggest that changes to the mechanisms for evaluating the school teaching policies had the greatest association with school performance. On average, these variables predicted 3.0% of the variation in schools' change scores. The quality of the evaluation procedures however accounted for a larger proportion of the changes (4.8%) than knowing whether th e changes were informed by data (1.2%).

The next most predictive macro-factor was the changes in the procedures for evaluating the school learning environment. On average these factors explained 0.9% of the variation in school performance, with changes to the policies for evaluating the environment explaining 1.1% of the disparity in schools' change scores and whether schools based any changes in policy upon evaluation data 0.6%.

Changes to the school teaching policies accounted for an average of 0.7% of the variation in school performance. Interestingly, though, most of the predictive capacity of this group stemmed from changes in alignment between the school curriculum and the assessed curriculum (1.9% of variance explained). The two remaining factors, changes in the quality of the policies on teachers' instructional practices and changes in the quantity of instruction time provided by the schools' policies, explained just 0.1% and 0.0% (1dp) respectively.

The least predictive group described the quality of the policies that regulate the school learning environment. This explained 0.3% of the variation in students' changes scores.

The results of this analysis therefore highlight that the factors which explain the highest proportion of the variance in schools' annual scores do not necessarily explain the highest proportion of the year-to-year changes. The substantive finding of the analysis were however unchanged, in that, the examination entry variables accounted for a greater percentage of the variation in Progress 8 scores than intake variables, and intake variables a greater percentage than schools' instructional practices. The school teaching policies had the second largest average effect size, explaining slightly more variation on average than intake or teaching practices.

Part 2: Forward-regression analyses

2017 Analysis

Table 11.3g identifies the 12 most influential variables within the 2017 analysis, as selected by forward regression modelling.

Table 11.3g: Forward-regression model of the 12 most influential variables in 2017 and their relationship with schools' Progress 8 scores

Rank	Variables	Change in R-squared	R-squared of model	B-value
1	The percentage of Year 11 students entering all English Baccalaureate subject areas	0.330	0.330	0.005
2	The overall percentage of absence across the school	0.126	0.456	-0.059
3	Frequency of classroom assessments	0.040	0.496	0.092
4	Average number of EBacc slots filled per pupil in Attainment 8 measure	0.030	0.526	0.613
5	The percentage of Progress 8 pupils that spoke English as an additional language	0.036	0.562	0.009
6	The percentage of Year 11 with SEN (with or without Statement/EHC plan)	0.019	0.582	0.012
7	The percentage of Progress 8 entrants that were female	0.021	0.603	0.003
8	The percentage of Progress 8 pupils that were disadvantaged	0.053	0.656	-0.014
9	Frequency of classroom disruptions	0.022	0.678	-0.074
10	Frequency of structuring tasks	0.009	0.687	-0.051
11	The average number of GCSE entries per pupil (not including equivalencies)	0.011	0.698	-0.157
12	The average number of GCSE and equivalent entries per pupil	0.020	0.718	0.129

* Beta values refer to the full 12 variable model.

Within the multiple-regression analyses (forward and hierarchical) the r-squared changes in column 3 identify the percentage of variance that each variable explained after taking into account the preceding variables, whilst column 4 keeps track of the variables combined effect.

It should be reemphasised, however, that although these statistics are accurate one needs to be cautious when interpreting the individual variable estimates. This is because the explanatory power of each factor will have been influenced by the order in which variables were entered into the model. It is therefore more defensible to observe that together the variables accounted for 71.8% of the variation in schools' value-added scores. And that the list of influential factors included 5 intake variables, 4 examination entry variables and 3 instructional practices, which explained 25.6%, 39.1% and 7.1% of the variation in school performance respectively. The effect of school policies was not considered in this analysis.

It is also important to recognise that the r-squared scores of examination entry practices are likely to have been exaggerated because an examination entry variable was entered into the model first, and that there is therefore a need to triangulate the results from all analyses to gain a comprehensive picture.

2018 Analysis

A comparable set of variables were identified in 2018 (see Table 11.3h).

Table 11.3h: Forward-regression model of the 12 most influential variables in 2018 and their relationship with schools' Progress 8 scores

Rank	Variables	Change in R-squared	R-squared of model	B-value
1	Average number of open slots filled per pupil in Attainment 8 measure	0.389	0.389	0.991
2	The percentage of persistent absentees across the school	0.066	0.455	-0.021
3	The percentage of Progress 8 entrants that were female	0.047	0.502	0.004
4	Average number of EBacc slots filled per pupil in Attainment 8 measure	0.036	0.538	0.990
5	The percentage of Progress 8 pupils that spoke English as an additional language	0.037	0.575	0.009
6	The percentage of Progress 8 pupils that were disadvantaged	0.067	0.642	-0.012
7	The percentage of Year 11 pupils included in the Progress 8 measure	0.009	0.651	-0.009
8	The percentage of Year 11 pupils that had SEN but no Statement or EHC plan	0.007	0.658	0.005
9	Frequency of questioning tasks	0.006	0.663	-0.070
10	Frequency of classroom assessments	0.012	0.675	0.071
11	The percentage of Year 11 pupils entering the English Baccalaureate humanities subject area	0.006	0.681	-0.002
12	The percentage of Year 11 pupils entering the English Baccalaureate science subject area	0.003	0.684	0.007

* Beta values refer to the full 12 variable model

Together these factors accounted for 68.4% of the variation in schools' performance ratings.

This time, there were 5 intake variables, 5 examination entry variables and 2 instructional practices that explained 22.4%, 44.2% and 1.7% of the variation in school performance respectively, when the effect of any preceding variables had been statistically controlled. Again, the influence of school policies was not considered.

Change Analysis

Table 11.3i: Forward-regression model of the 12 most influential variables in change analysis and their relationship with the 2017-2018 change in schools' Progress 8 scores

Rank	Variables	Change in R-squared	R-squared of model	B-value
1	Change in the average number of open slots filled per pupil in Attainment 8 measure	0.160	0.160	1.264
2	Change in the quality of policies for evaluating teaching policies	0.040	0.200	0.064
3	Change in the average number of GCSE and equivalent entries per pupil	0.036	0.236	0.225
4	Change in the average number of GCSE entries per pupil (not including equivalencies)	0.038	0.274	-0.195
5	Change in the percentage of persistent absentees across the school	0.035	0.309	-0.016
6	Change in the percentage of Year 11 pupils entering the English Baccalaureate English subject area	0.029	0.338	0.017
7	Change in the proportion of lesson time that was used for teaching	0.018	0.356	0.098
8	Change in the Average number of EBacc slots filled per pupil in Attainment 8 measure	0.018	0.374	0.380
9	Change in the quality of policies regulating instructional behaviours	0.013	0.387	-0.062
10	Change in the Frequency of classroom disruptions	0.013	0.400	0.034
11	Change in the percentage of non-mobile pupils	0.011	0.411	-0.019
12	Change in the percentage of Progress 8 entrants that were female	0.012	0.423	-0.004

* Beta values refer to the full 12 variable model

The 12 most influential variables from the change analysis were able to account for 42.3% of change in schools' ratings (see Table 11.3i).

5 of these were examination entry variables, 3 were intake factors, 2 were instructional practices and 2 were school policies. These groups accounted for 28.1%, 5.8%, 3.2% and 5.3% of the variation in school performance respectively when the effect of any preceding variables had been statistically controlled.

The overall pattern of results was therefore comparable across the three sets of analysis.

Part 3: Hierarchical linear regression models

2017 Analysis

Table 11.3j: Hierarchical linear regression model of the most influential variables in 2017 and their relationship with schools' Progress 8 scores

Rank	Variables	Change in R-squared	R-squared of model	B-value
1	The overall percentage of absence across the school	0.322	0.322	-0.126
2	The percentage of Progress 8 entrants that were female	0.047	0.369	0.004
3	The percentage of Progress 8 pupils that were disadvantaged	0.041	0.410	-0.000
4	Frequency of classroom disruptions	0.047	0.456	-0.051
5	Frequency of classroom assessments	0.033	0.489	0.102
6	Frequency of orientation tasks	0.004	0.494	0.018
7	Percentage Year 11 entering all English Baccalaureate subject areas	0.053	0.547	0.004
8	Average number of EBacc slots filled in Attainment 8	0.027	0.574	0.679
9	Number of pupils included in Progress 8 measure	0.023	0.597	0.001

* Beta values refer to the full 12 variable model

Hierarchical modelling of the 2017 data produced similar results (see Table 11.3j).

In this representation, 59.7% of the variation in schools' value-added scores was accounted for. The first tier of the model, i.e. the intake factors, was able to explain 41.0% of differences in schools' ratings. Adding a second tier, containing classroom behaviours, explained an additional 8.4%. The addition of the final tier of examination entry variables a further 10.4%.

It is stressed once more, though, that the reader should not place too great an emphasis on the estimates for individual variables, particularly when comparing the influence of factors that were allocated to the same group. This is because any overlap in the variance that these factors can explain would have been attributed to the variable that was entered into the model first. Whilst one can, for example, be reasonably confident that student absence was the most predictive intake variable, it is less certain that it is several times as influential as the percentage of female students per cohort or the percentages of disadvantaged students. This is because female and middle-class students may have better attendance than their peers.

2018 Analysis

Table 11.3k: Hierarchical linear regression model of the most influential variables in 2018 and their relationship with schools' Progress 8 scores

Rank	Variables	Change in R-squared	R-squared of model	B-value
1	The percentage of persistent absentees across the school	0.315	0.315	-0.033
2	The percentage of Progress 8 entrants that were female	0.074	0.389	0.005
3	The percentage of Progress 8 pupils that were disadvantaged	0.050	0.439	-0.004
4	Frequency of classroom assessments	0.017	0.457	0.083
5	Frequency of orientation tasks	0.009	0.466	0.048
6	Frequency of questioning	0.013	0.479	-0.086
7	Average number of open slots filled in Attainment 8	0.080	0.559	1.156
8	Average number of EBacc slots filled in Attainment 8	0.024	0.583	1.203
9	Percentage Yr11 entering Baccalaureate Humanities	0.021	0.603	-0.004

* Beta values refer to the full 12 variable model

The hierarchical model of the 2018 data was able to explain for 60.3% of the variation in schools' performance (see Table 11.3k).

43.9% of this was accounted for by intake tier of the model. The total variance explained then increased by 4.0% when the three classroom behaviours were added and by a further 12.5% when the examination entry variables were included.

Change Analysis

Table 11.3L: Hierarchical linear regression model of the most influential variables in the change analysis and their relationship with the 2017-2018 change in schools' Progress 8 scores

Rank	Variables	Change in R-squared	R-squared of model	B-value
1	Change in the percentage of persistent absentees	0.027	0.027	-0.019
2	Change in percentage of Year 11 pupils with SEN – With or without Statement/EHC plan	0.019	0.046	-0.006
3	Change in percentage of Year 11 pupils with SEN Statements or EHC plan	0.008	0.054	-0.006
4	Change in the proportion of lesson time that was used for teaching	0.030	0.084	0.102
5	Change in the frequency of structuring tasks	0.031	0.115	0.011
6	Change in the frequency of application tasks	0.010	0.125	-0.027
7	Change in quality of policies for evaluating teaching policies	0.014	0.139	0.039
8	Change in the quality of policies regulating instructional behaviours	0.015	0.155	-0.042
9	Whether changes in SLE policies were based on evaluation data	0.005	0. 159	0.027
10	Change in average number of Open Bucket slots filled in Attainment 8 per pupil	0.111	0.270	1.417
11	Change in average number of GCSEs per pupil – including equivalents	0.038	0.308	0.207
12	Change in average number of GCSEs per pupil – not including equivalents	0.053	0.361	-0.165

*The third variable is interchangeable with *the change in the percentage of Year 11 pupils with SEN but no Statement or Educational Health and Care plan*. Since the total percentage of SEN students has been accounted for these two variables express the same information from opposing perspectives. All of the data concerning this variable was therefore the same, except for the direction of association which was reversed.
** Beta values refer to the full 12 variable model

The final hierarchical model was able to account for 36.1% of the change in schools' Progress 8 scores between 2017 and 2018 (see Table 11.3L).

The first tier of the model, i.e. the intake factors, accounted for 5.4% of this. Adding the second- (classroom behaviours), third- (school policies) and fourth tiers (examination entry variables) increased the explainable variation by 7.1%, 3.4% and 20.2% respectively.

11.4. Discussion

This chapter was intended to establish whether differences in schools' performance ratings could be explained by correlates from school effectiveness research and thus whether Progress 8 provides a meaningful indicator of school success.

Could the variation in schools' performance ratings be predicted by established effectiveness factors?

With regards to the first of these objectives, the evidence is compelling. In all three sections of the analyses the operationalised factors predicted meaningful proportions of schools' results, both in terms of the variation that could be explained by individual variables (33.0% in 2017 and 38.9% in 2018 when the effect of other variables not statistically controlled) and their collective effect (59.7%-71.8% in the multiple-regression models with 2017 data, 60.3%-68.4% for the models with 2018 data). Furthermore, whilst the r-squared scores in the change analysis were smaller (16.0% and 36.1-42.3% respectively), this was to be expected as the values that the factors were predicting were much smaller. There will also have been larger quantities of construct irrelevant variance owing to the mechanisms discussed in Gorard (2010a). What is more, the directional effect of factors was consistent with their theoretical impact. Specifically, 96.8%, 75.0% and 88.9% of the interactions in the simple-regression models, forward-regression models and hierarchical analyses from 2017, 96.8%, 75.0% and 77.8% of the interactions from 2018, and 81.0%, 58.3% and 75.0% in interactions in the change analysis were consistent with the hypothesised effects. All of which supports the validity of Progress 8 assessments.

In terms of the consistency of these findings, the effect attributed to individual variables was reasonably stable across the analyses. There was, for example, an r=0.830 correlation between the variance that each factor accounted for in the 2017 and 2018 simple linear-regression models, and many of the factors identified in the multiple-regression and hierarchical model represent similar aspects of schools' provisions. The results make it clear however that the variables that account for the differences in schools' performance at specified moments in time are not necessarily the best predictors of the change in schools' scores over time. As evidenced by the low to moderate correlations between the r-squared scores of variables in the 2017 and changes analyses (r=0.206) and the 2018 and change analyses (r=0.516). With hindsight this is understandable as the stability of a variable does not detract from its importance.

It is therefore concluded that both the within year differences in schools' Progress 8 ratings and the change in schools' ratings over time can be predicted by the type of factors that account for the differences in schools' raw-attainment.

Does Progress 8 provide a fair method of evaluating schools' contribution?

Evidence relating to the second research objective was, however, more concerning. In the majority of analyses examination entry variables were able to explain the highest proportion of the variation in schools' performance, both at specified moments in time and over time (see Table 11.4a). Intake factors were the second most predictive category, though there was evidence to suggest that these variables may assume greater importance when the structure of the underlying data is acknowledged. Therefore, whilst differences in schools' instructional practices and policies had a meaningful relationship with schools' performance ratings, this may be dwarfed by factors that are outside of schools control.

Table 11.4a: The variance in Progress 8 scores explained by each category of variable in the simple, forward and hierarchical models of school effectiveness

	2017:	2018:	2017-2018 Change
Part 1: Simple linear regression models	Intake: 10.1% (4.8%**) Instruct: 5.8% Policy: N/A Entries: 14.9%	Intake: 10.4% (5.4%**) Instruct: 2.6% Policy: N/A Entries: 19.2%	Intake: 1.2% (1.0%**) Instruct: 1.1% Policy: 1.2% Entries: 3.4%
Part 2: Forward-regression model	Intake: 25.6% (13.0%**) Instruct: 7.1% Policy: N/A Entries: 39.1%	Intake: 22.4% (15.8%**) Instruct: 1.7% Policy: N/A Entries: 44.2%	Intake: 5.8% (2.3%**) Instruct: 3.2% Policy: 5.3% Entries: 28.1%
Part 3: Hierarchical-regression model	Intake: 41.0% (24.6%**) Instruct: 8.4% Policy: N/A Entries: 10.4%	Intake: 43.9% (25.9%**) Instruct: 4.0% Policy: N/A Entries: 12.5%	Intake: 5.4% (2.8%**) Instruct: 7.1% Policy: 3.4% Entries: 20.2%

*It is important to recognise that these percentages refer to different statistics. The simple linear regression row refers to the average variance that could be explained by each category of variable when the effect of extraneous influences was ignored. The latter two, to the overall percentage of variance that the variables in each class referred to after the influence of preceding factors had been accounted for. The rank-order and magnitude of effects reported within each section are therefore not intended to be identical. The three perspectives, however, report upon related matters, so patterns within the results are meaningful.
**The percentages in brackets report upon the contribution of the intake variables if the two absence variables are excluded from the calculation. In the case of the 2017 and 2018 hierarchical models, for example, this means the percentage of variation explained by the two remaining intake variables (the percentage of female and disadvantaged Progress 8 entrants) if they are entered into the model first.

Although there are reasons to suspect that these figures may give an exaggerated impression of the bias within Progress 8 assessments (see later discussions), the results are troubling. Especially, when one considers that the majority of variables included in the intake and examination-entry categories should not, it is argued here, be considered as genuine school effects.

A more in-depth discussion of the interactions that occurred within each category of variable will now be provided. Followed by a discussion of the methodological weaknesses in the research design and the extent to which they may have impacted upon the results.

i. Examination entry variables

Difference in schools' examination entry practices were closely associated with their performance ratings.

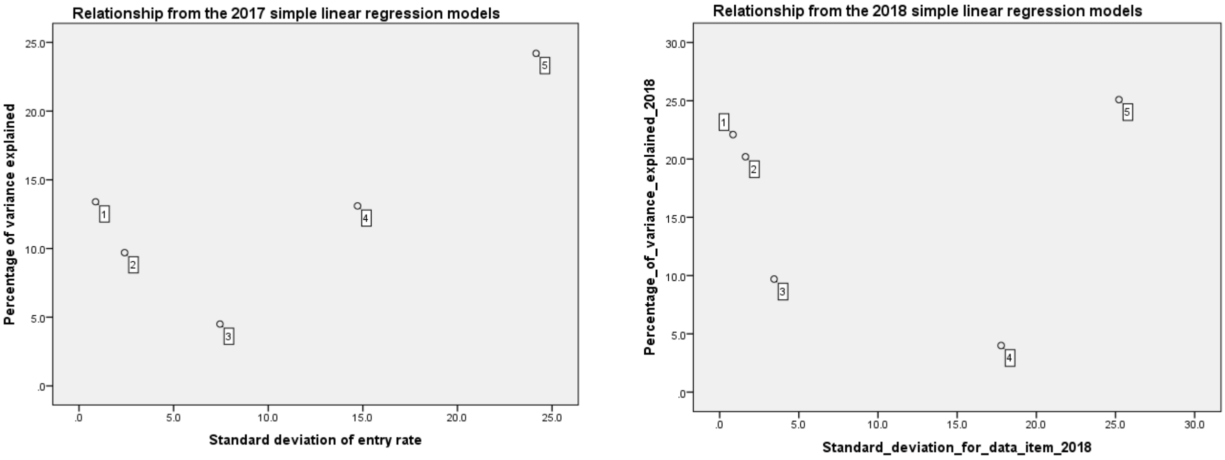
For the most part these interactions were consistent with our expectations. 31/36 (86.1%) of the results from the simple linear-regression models were in-line with relationships outlined in Section 7.3 (or Section 11.2.4 in the case of new factors). As were the 11/14 (78.6%) of the relationships within the forward-regression models and 6/9 (66.7%) within the hierarchical model. It would appear therefore that the greater the proportion of students that filled the Attainment 8 buckets, the more favourable a schools' rating were likely to be. This is, of course, a logical association that one would have expected to find.

Under-appreciated associations

There were however two instances where the interactions between variables were more subtle than previously appreciated.

Firstly, there was evidence to suggest that the variation in subject entry rates may have a non-linear association with school performance (see Figure 11.4a). That is to say, that entries into the least and most entered subject areas may have had the closest association with schools' progression ratings because these are the areas where the differences are most overt.

Figure 11.4a: Scatter graphs of the relationship between the standard deviation of subject entry rates and the variation in Progress 8 scores that the variables explained in 2017 and 2018



*Where;
1. The percentage of student entering the EBacc maths subject area.
2. The percentage of student entering the EBacc English subject area.
3. The percentage of student entering the EBacc science subject area.
4. The percentage of student entering the EBacc humanities subject area.
5. The percentage of student entering the EBacc language subject area.

Secondly, whilst factors such as the average number of EBacc and Open slots filled by students and the average number of GCSEs entered by school cohort explained substantial portions of both the variation in schools' results at specified moments in time and the change in schools' results over time, the same could not be said for the percentage of students' to enter all English Baccalaureate subject areas. This variable predicted more variation than any other factor in the 2017 simple linear regression analyses (33.0%) and a large proportion of differences the following year (28.3%). Yet in the change analysis it accounted for just 0.2% of changes in schools' ratings. The best explanation for this is that the measure provided a close but imperfect proxy for coverage of the Attainment 8 slots. Since both are threshold measures that only recognise certain increases in examination entries, it is therefore possible for the average coverage of the Attainment 8 slots to increase without this being reflected in the percentage of pupils that entered the full English Baccalaureate, or vice versa.

Unexpected associations

Across the 9 analyses, 11 variables had unanticipated directional effects.

The most common inconsistency was for the size of schools' cohorts to have a positive correlation with the schools' performance rating. This happened in 4/4 (100%) of the analyses which included the variable. That is to say within the 3 simple regression models and the hierarchical regression model of 2017 data. There two possible explanations for this. Either having a low number of Progress 8 entrants did not advantage smaller schools in the way which Gorard hypothesised, which would not be a radical conclusion given that the primary effect of having a small number of entrants is an increase in the range and instability of schools' ratings rather than the directional bias (see Gorard *et al.*, 2013), or the effect was present but overwhelmed by the influence of other factors. Neither situation would be surprising given that in all analyses these relationships accounted for less than 2.5% of the variation in schools' performance. In this instance, however, it is argued argue that the foremost explanation is more likely due to the consistency of the result across datasets. Furthermore, the percentage of Year 11 pupils entered into schools' Progress 8 calculations exhibited the expected directional effect in 3/4 (75%) of analyses that considered the matter (the three simple regression models and the forward-regression model of 2018 data). One could therefore postulate that it may be the type of pupils that tend to be excluded from schools' ratings that biases schools scores rather than school size. Gorard et al., (2013), however, did not attempt to distinguish between the two effects.

The entry-rates for the EBacc humanity subject area also had an inconsistent association with school performance. Within the 2017 and 2018 simple regression models, the variable adhered to the pre-established expectations (see Section 11.2.4) and had low-moderate positive linear association with school performance (r-squared = 13.1% and 4.0% respectively) but in the forward regression model of the 2018 data, the hierarchical regression model of the 2018 data and the simple-linear regression model of 2017-18 changes the correlation was very weak and negative (r-squared = 0.0% to 2.6%). The three discrepancies, however, could be explained by the magnitude of the effect and the use of statistical controls. In the first instance, for example, the relationship accounted for 0.0% (1dp) of the change in schools' scores. Whilst it is technically true that the direction of the association conflicted with our expectations, for all intents and purposes one can read into this that the variable had no discernible impact and the tiny association that existed most likely occurred due to chance. In the latter two analyses the effect of the variable was evaluated after taking into account the impact that the average number of EBacc and Open slots entries had upon schools' scores. It is therefore logical that once schools' coverage of the Attainment 8 slots had been accounted for, the percentage of students' entering EBacc Humanities qualifications would lose its explanatory power. What is more, if some students' studied for qualification that did not count towards their progress score then this would explain the direction of the relationship.

Finally, the average number of GCSE entries (excluding equivalencies) exhibited a negative association with school performance on three occasions; in the 2017 forward-regression model, the forward-regression model of 2017-2018 changes and the hierarchical model of the 2017-18 changes. In each instance, however, either the average number of GCSEs entered by students (including equivalencies) and/or multiple proxies for coverage of the Attainment 8 slots had already been taken into account. The unexpected correlations are therefore presumed to indicate that at a certain point the benefit of entering students for additional qualifications tapers off. This type of relationship was therefore predicted and consistent with the inverted-U relationships that were found between the average number of GCSE and equivalent qualifications and Progress 8 ratings in 2017, and the inverted-U relationship found between the 2018 progress scores and the average number of GCSE including and excluding equivalents (see Footnotes 19 and 20).

All of the anomalous results therefore had plausible explanations.

The consistency of the results across datasets

In terms of the consistency of factor's effect sizes, a high level of association was once again found between the percentage of the variation in Progress 8 scores that each examination entry factors could explain in 2017 and 2018, when the effect of extraneous variables was not accounted for (r=0.779). The same variables, most notably, the measures of students overall entry rates (e.g. the average number of EBacc/Open slots filled) also tended to emerge as the most influential examination entry variables in the multiple-regression models. This implies that there was a reasonably consistent gradation of effects within the annual analyses.

Despite some familiar variables appearing within the forward- and hierarchical-models of change, however, there was evidence to suggest that the factors that account for the greatest proportion of the within-year variance in schools' scores are not necessarily the same factors that explain the changes in schools' ratings over time. Including the fact that the correlation between the effect sizes recorded in the simple linear regression models of the 2017 and 2017-18 change data, and the 2018 and 2017-18 change data, had low to moderate levels of association (r=0.157 and r=0.540) respectively. This is presumably because some of the most influential factors are stable.

Interpretations and implications

Whilst differences in students' examination entry patterns are framed by schools' curriculum decisions these influences were not considered to be indicative of genuine school effects. This is based upon the belief that there is no direct link between these variables and the quality of the schools' instructional provisions.

This statement, however, is open for debate. In this thesis all learning was valued equally. That is to say, that a school was considered to be effective whether it enhanced students' progress in academic or vocational areas. Others however have argued that certain types of knowledge should be prioritised. The DfE, for example, designed the weightings of Progress 8 so as to promote learning in particular subject areas and types of qualification. From this perspective one might look upon a school that specialises in maths instruction as providing more useful instruction than one that specialised in sport, music or art. Which interpretation one accepts is of course an ideological rather than a methodological decision. Under either definition however the variance explained by these factors does not refer to the characteristics of the schools' teaching policies or practices. It is therefore argued that Progress 8's ability to report upon the quality of schools' provisions is therefore contingent upon these variables having a low to moderate effect. Taken at face value, however, the results of this analysis suggest that this may not be the case and that the schools' annual performance ratings may be overwhelmed by these kinds of influences.

Alternative interpretations of the associations

It is important to recognise, though, that there may be other reasons for the magnitude of these associations. The predominant concern is that the relationship been Progress 8 ratings, school quality and students' examination entries might have been reciprocal. That is to say, that in addition to students'/schools' curricular decisions having implications for the schools' performance ratings, differences in the quality of schools' tutorage might also have impacted upon students'/schools' curricular decisions. This would have occurred if the pupils that made greater academic progress, relative to students with comparable prior-attainment, were more likely to enter (or be entered) for a higher number of qualifications in a more rounded selection of subjects. If this was the case, the aforementioned interpretation of the data may have slightly or grossly overstated the casual impact of the examination entry variables.

Steps were taken to minimise this risk. Specifically, the examination entry variables were entered into the final tier of hierarchical regression models, meaning that the associations report upon the percentage of variance that could be explained by these factors after the specified differences in school intakes, teaching practices and policies had been taken into account.

After looking at the results of the 9 analyses collectively (see Table 11.4a) though, one has to wonder if these precautions were sufficient. Within the 2017 and 2018 model, the results followed a logical pattern. That is to say, that within the simple and forward regression analyses, examination entry variables accounted for largest portion of the variance in schools' performance (14.9% and 39.1% in 2017, 19.2% and 44.2% in 2018, respectively). The most important category of variable then changed during the hierarchical analyses as intake and instructional variables were given preferential treatment. In fact, whenever the entry of the intake and instructional variables preceded the consideration of the examination entry differences, the intake factors accounted for almost four times the variance that exam entries explained. Thus within the 2017 and 2018 hierarchical models examination entry differences explain just 10.4% and 12.5% of the variance respectively after the effect of the other variables had been taken into account. Within the 2017-18 change analyses, however, a different pattern emerged. In this instance, intake factors started off explaining a far lower percentage of the variance in schools' results (an average of 1.2% in simple models and 5.8% collectively in forward-regression model). In fact, the percentage of variance was so low that it barely exceeded to predictive power of instructional practices. This difference in starting point meant that when the intake and instructional variables were given preferential treatment and entered into the 2017-18 hierarchical model early, the ratio of effect sizes attributed to intake and examination entry variables barely changed. Now, it could genuinely be the case that intake factors are relatively stable and therefore have little impact upon the change in schools' ratings over time. As we shall discuss shortly however there is a plausible reason for suspecting that the low r-squared scores attributed to the intake factors during the change analysis are due, at least part, to a methodological shortfall in the operationalisation of the two absence variables. If that is the case then the distribution described could instead be attributed to the examination-entry variables' capacity to mop up any variation that is not explained by the preceding factors. This uncertainty will need to be taken into account when the results are interpreted.

ii. Intake variables

Differences between school intakes explained substantial proportions of the variation in schools' Progress 8 results. In fact, this was the second most predictive category of variable in 5 of the 9 analyses performed. What is more, 28/30 (93.3%) of the relationships modelled in the simple linear-regression analyses, 9/13 (69.2%) of the interaction in the forward-regression models and 9/9 (100%) of the interactions in the hierarchical models were consistent with expectations, which supports the notion that these relationships are genuine rather than coincidental associations.

In fact, the data suggests that intake factors may explain even higher proportions of variance when the structure of the underlying data is acknowledged. This is because regression analyses are likely to attribute the most variation to variables that are entered earlier in the model. The forward-regression models are therefore likely to have exaggerated the effect of examination entry variables, as an examination entry variable was entered first in each instance. In the hierarchical model, however, intake factors were entered early-on to acknowledge that these differences are more proximate to students' learning and often will pre-date the other effects. When this was the case the total percentage of variance explained by such factor jumped from around 25% to more than 40%. It should also be noted that whilst it is possible to argue the same point in reverse and thereby claim that the latter figure overstates the influence of intake factors, this stance is supported by school effectiveness theory and research.

The effect of intake differences was also highly consistent year-to-year. In terms of the percentage of variance explained by each variable in the simple linear regression models, for example, there was a near perfect correlation between proportion of deviation that these considerations accounted for in 2017 and 2018 (r=0.980). A measure of student absence, the percentage of female students and the percentage of disadvantaged students also appeared in all of the annual multiple-regression models and the majority of the change analyses. Though once again, the modelling highlighted that with the exception of student absence rates, the factors that are responsible for the within-year differences in students' progress scores are not necessarily the same as those that are responsible for year on year changes. As evidenced by the positive but modest (r=0.229 and r=0.259) correlation between the r-squared scores reported in the 2017 and Change Analysis simple linear regression models, and the 2018 and Change Analysis simple linear regression models respectively.

There were predictably however a few instances where variables interaction with Progress 8 ratings did not conform to expectations (six relationships across the nine analyses). Two of these related to the percentage of female students per cohort, two to the percentage of non-mobile students, one to the percentage of SEN students (with or without a Statement/EHC plan) and one to the percentage of SEN students (without a Statement or EHC plan). All however were isolated incidents (unique to one dataset, i.e. the 2017, 2018 or 2017-18 data) that conflicted with the overall pattern or results for the associated variable (six of nine analyses, for example, found a positive association between the percentages of female students and Progress 8 rating). The associations were also weak (0.4-1.9% of variance explained) in relation to the correctly predicted relationships (up to 7.4% explained by each variable), and defied our efforts to devise a logical explanation. It is therefore argued that the events were most likely due to non-causal, chance-based associations. Though it is possible that the meaning of the 4/6 variables that were assessed late within forward-regression models had become distorted to the point that their implications were difficult to track.

Overall, the results from this section are very concerning as all of the intake factors that were collated in this category could be considered as extraneous influences that are predominantly out of schools' control. The evidence presented in this section therefore suggests that Progress 8 provides a biased measure of school performance that will punishes schools with disadvantaged intakes. The categorisation of the attendance variables, however, is debatable. Whilst it is argued here that attendance levels are ultimately mediated by students and their parents, school policies and teachers' behaviours may also play their part (Creemers and Kyriakides, 2008). For this reason some researchers would consider it unjust to have treated these influences as non-school factors that need to be controlled. At the same time it would be very unfair to uncritically assume that all of these differences were attributable to differences in schools' provisions. Table 11.4a therefore reports the effect that student intake variables had when the attendance variables were included and excluded. Whilst this reduces the estimate of bias significantly, it does not have any substantive impact upon the study's conclusions.

Alternative explanations of the data

There was however a further methodological concern. Whilst the two measures of student absence rates were influential variables in all of the analyses, the percentage of variance that they accounted for was substantially lower within the three analyses of change. This statement, of course, could be applied to most of the variables in our analysis due to the fact that these models were attempting to p redict smaller variations in schools' performance. In this instance, though, the effect was more dramatic and there is a strong argument for believing it artificial. Specifically, this was the only variable in the analyses that had to be modelled at school-, rather than cohort-level. That is to say that the overall percentage of absence in 2017, for example, reported upon the rate on non-attendance across all school year-groups (7-11). It therefore stands to reason that this would introduce more noise into the analyses and decreased the percentage of variance that the variables accounted for. The problem would be exacerbated during the change analyses, however, as any inaccuracies would be larger in relation the measurement scale. Difference in student attainment level may thus have had a more substantial impact upon the changes to schools' Progress 8 ratings than the results imply.

iii. Instructional behaviours

Instructional behaviours accounted for modest proportions of the variation in schools' performance (see Table 11.4b). In fact, they were the least predictive category of variable in all analyses, except the final hierarchical analysis of 2017-2018 changes.

Though the differences in factor's effects was not clear cut, the results suggest that the classroom learning environment, orientation tasks and teacher-modelling had the greatest impact upon schools' annual performance ratings. These behaviours accounted for an average of 7.0%, 5.9% and 5.0% of the variation in schools' annual Progress 8 scores respectively[33]. Whereas applications tasks, structuring, questioning and classroom assessments accounted for averages of 2.3%, 1.5%, 1.4% and 0.8%. Teachers' ability to manage instructional time also had a close association with schools' ratings, but this factor was only evaluated within one set on analyses[34].

The variables that best explained the variance at specified moments in time, however, we not necessarily the most effective predictors of year-to-year changes.

[33] It should be acknowledged there was a particularly close association between classroom disruptions and Progress 8 scores in 2017 (r =- 0.475). This is presumed to have been a one-off chance occurrence. If this is the case, this figure will have exaggerated the effect attributed to the classroom learning environment. Even if this figure is excluded from the cited average, however, the factor had a notable effect upon schools' annual Progress 8 scores (3.9%).

Table 11.4b: The variation in Progress 8 scores that was explained by each instructional behaviour in the simple, forward and hierarchical models of school effectiveness

	Part 1: Simple regression models			Part 2: Forward-regression models			Part 3: Hierarchical-regression models		
Variable	2017	2018	Change	2017	2018	Change	2017	2018	Change
1. Frequency of orientation tasks	6.2%	5.6%	0.3%				0.4%	0.9 %	
2. Frequency of structuring	0.8%	2.1%	2.5%	0.9%					3.1%
3. Frequency of questioning	2.1%	0.6%	0.1%		0.6%			1.3 %	
4. Frequency of teacher-modelling	7.9%	2.1%	1.0%						
5. Frequency of application tasks	2.7%	1.9%	0.1%						1.0%
6. Frequency of on-task teacher-student interactions	3.7%	5.5%	0.6%						
7. Frequency of on-task student-student interactions	4.4%	2.6%	1.1%						
8. Frequency of classroom disruptions	22.6%	3.2%	0.2%	2.2%		1.3%	4.7%		
9. Proportion of lesson time used for teaching			3.1%			1.8%			3.0%
10. Frequency of classroom assessments	1.6%	0.0%	1.9%	4.0%	1.2%		3.3%	1.7 %	
11. Quality of teachers' instructional behaviours			1.9%						
12. Teachers' coverage of the school curriculum			0.5%						

*Percentages in Part 1 analyses refer to the percentage of variance explained when extraneous variables are not controlled. The percentage in the Part 2 and 3 analyses refer the additional variance that the factor accounted for after controlling for the preceding variables from that model.

**Shaded cells indicate that the direction of the relationship was unanticipated.

These results are discouraging as all of the aforementioned influences would be considered as genuine school effects. Were one to take these results at face value, it would then follow that the differences in schools' instructional practices accounted for less variation than the non-school influences discussed thus far. Progress 8 figures would therefore not only be influenced by external sources of bias but most likely overwhelmed by them. Three alternative interpretations, however, are possible. These are discussed below.

Alternative explanation 1:

The first explanation of these results accepts these interactions as they are reported. That is to say that the stated differences in schools' teaching practices had a predictable but very limited effect.

The dynamic model specifies, however, that educational factors can be viewed from 5 different perspectives; the frequency of actions, their focus, timing, quality and the amount of differentiation that took place. For the most part, this analysis concentrated on only one of these dimensions (the frequency dimension). More comprehensive modelling of the relationships may therefore have increased the percentage of variance that these factors were able to explain. The same could also be said for the use of multi-level modelling, the consideration of same-level interactions, non-linear relationships and/or clustering effects (i.e. grouping of variables that lead to particularly effective outcomes) which are deemed plausible but were not considered by the current methodology. Such an interpretation would be consistent with Creemers and Kyriakides' (2008) modelling of classroom-level effectiveness factors (see Section 7.4.3, part 2).

Alternative explanation 2:

A similar situation might also have arisen if school leaders were not able to report upon classroom-level behaviours with enough precision. This would be understandable as some of the questions within the research questionnaire were very specific and are likely to have taxed the knowledge that leaders had about their teachers' practices. Particularly when one recognises that a representative of the schools' management team is unlikely to have been present to observe all of students' instruction.

The analysis therefore relied upon the assumption that most, if not all leaders, would engage in some form of formative evaluation and that this would provide the necessary insight. Many schools also provide professional development activities or promote particular practices. These undertakings would make it easier to report upon the characteristics of classroom instruction. Furthermore, one of the reasons for including a 'change analysis' was the supposition that school leaders may find it easier to report upon changes in school practice than the absolute prevalence of particular activities, especially if these changes were deliberately brought about.

It should be noted, though, that the demand placed upon questionnaire respondents was even greater in this instance because the assessment took place at school level. Leaders were therefore expected to summarise the behaviours of any individuals who provided instruction to the Year 11 cohort. This would not have been an easy task if teachers were given professional autonomy. Additional construct irrelevant variance may also have entered into the analysis because the questionnaires were returned over a five month period. Some leaders may therefore have had more information to act upon.

The research design of this analysis may thus have under-estimated the difficultly of leaders' task. Were this the case then the percentage of variance explained by classroom instructional variables would be underreported due to an excess of error in leaders' appraisals. This would also account from the eight instructional behaviours having comparable effect sizes.

This interpretation is supported by Creemers and Kyriakides' (2008) empirical assessments of their measurement framework. Specifically, the results reported in Section 7.4.3 (part 1) of this thesis, which demonstrate that all research instruments provide an imperfect appraisal of educational effectiveness factors, and that several different forms of data collection methods should ideally be drawn up to assess each variable. Practical restrictions, however, made this impossible in the current analysis (see Section 11.2.5).

Alternative explanation 3:

The final explanation is the most concerning, at least in terms of the methodological implications.

A recognised weakness of regression analysis is that all factors will correlate to a certain extent, whether there is a causal relationship between them or not. Factor's effect sizes might not therefore be representative of their true influence. This limitation must be considered when interpreting the results of any regression based analysis, but is particularly applicable here because of the modest and relatively indistinct associations reported in this sub-section. After applying statistical controls, the standard practice within educational effectiveness researchers would be to use significance tests or another form of probably statistic to help judge whether the observed relationships were meaningful. That is to say, whether they are likely to have occurred by chance. As discussed in Section 6.3, however, this is not what significance tests report, especially when one has utilised data from a non-random sample. The approach was therefore considered unhelpful in making this distinction. Given the comparability of the stated instructional effects, how then can we determine which differences are meaningful?

Whilst it is not possible to distinguish between the four explanations (the original and three alternative explanations) with certainty, one can make an informed judgement about the generalisability of the results by evaluating two aspects of their consistency; how consistently the findings adhered to the relationships established within school effectiveness research, and the consistency of the results across datasets and model specifications.

As Table 11.4b makes clear, the majority of the interactions acted in directions that were consistent with our expectations. More specifically, 28/30 (93.3%) of the associations from the simple linear-regression models were correctly anticipated, 4/7 (57.1%) of the relationships within the forward-regression models and 7/9 (77/.8%) of correlations in the hierarchical regression models. What is more, all of the unanticipated associations had small effect sizes (0.2 % to 1.3% variance explained) and relationships that defied logical explanation. It is reasonable to conclude therefore that these anomalies were most likely non-causal chance associations. This evidence provides support for the results but hints that construct irrelevant variance may have some impact.

The annual effect sizes reported for each factor also correlate with the mean effect sizes from Kyriakides *et al.'s* (2013) meta-analysis of classroom-level effectiveness factors. This evidence is discussed in detail in Section 7.4.1. To more precise, there is an r = 0.679 correlation between the r-squared scores of each factor in the 2017 simple linear regression analysis and those in Kyriakides' study, and an r = 0.320 association between r-squared scores from the 2018 simple linear regression models and the same figures. Whilst this alignment is not perfect, it suggests that results of this study are generalizable, and that causal mechanisms do therefore underpin the observed relationships. That being said, the effect sizes reported in Kyriakides *et al.*, (2013) study were noticeably larger, which lends further support to alternative interpretations 1 and 2. It is also important to acknowledge that in order to compare these results, it was necessary to average the effect sizes of teacher-student interactions, student-student interactions and classroom-disruptions, to attain an average effect size for the classroom learning environment variables. This significance of this will be discussed shortly.

The evidence on the second matter, however, is more troubling. Whilst there was a strong positive correlation between the percentage of variance that each over-arching classroom factor explained in the three sets of simple linear regression analyses (r = 0.744 between the r-squared scores of over-arching factors in the 2017 and 2018, r = 0.502 between the r-squared scores of the over-arching factors in the 2017 and change analyses, and r = 0.874 correlation between the r-squared scores of the over-arching factors in the 2018 and change analyses), this association drops considerably when the association between the individual factors was considered (r = 0.264 association between the r-squared scores of variables in the 2017 and 2018 simple linear regression analyses, r = -0.398 between the r-squared scores of variables in the 2017 and change analyses, and r = -0.334 between the r-squared scores of variables in the 2018 and change analyses.) That is to say, when the three elements of the classroom learning environment are operationalised as separate variables, as they were during our analyses. This illustrates that the results of the three simple linear regression analyses were loosely aligned, but that the effect attributed to some variables varied substantially. Furthermore, whilst the directional effect of the most influential classroom behaviours was unaffected by the type of statistical modelling that was used, the effect ascribed to the least influential variables was often inconsistent.

The frequency of questioning, for example, had a positive association with performance in the simple linear regression models and a negative association within the forward and hierarchical models. Both observations re-emphasise that one should be cautious of interpreting individual parameter estimates, especially factors those with very low r-squared scores.

The available evidence therefore suggests that cited effect sizes do reflect the impact that instructional behaviours have upon schools' value-added scores. The measures, however, are likely to have been imperfect and our interpretation of the data will need to reflect this.

iv. School policies

School policy variables were only evaluated in the three analyses of the 2017-18 changes. The results of these assessments were mixed. The outputs of the simple and forward-regression models indicate that these variables had low to moderate levels of association with the year-to-year changes in schools' performance ratings, which were comparable to those of intake variables and surpassed the relationship with schools' practices. Logical patterns therefore emerged within the data that are suggestive of causal associations. Firstly, the mechanisms that govern schools' policies were shown to be more influential than the content of the policies themselves. This finding is in-line with Creemers' and Kyriakides (2008) argument that policy changes and improvement efforts must always consider schools' strengths and weakness, as addressing shortfalls in schools' provisions is likely to bring about greater improvement than further developing effective areas. The practice also helps to ensure that there is continuity in schools' improvement efforts. Likewise, it makes sense that the policies and evaluation mechanisms that regulate what happens in classrooms would be more influential than those that govern behaviour outside of classrooms, as classrooms are the locus of the education experience (Scheerens, 1992). And whilst it was assumed that it would be important for schools to base their policy changes upon evaluation data, it makes sense that these factors would account for small portions of the differences between schools, as they were operationalised as a dichotomous variables that described the schools' actions in a typical or general case. This simplistic operationalisation may have reduced the variance that these factors were able to explain.

On the other hand, when the underlying structure of the data was acknowledged (see hierarchical model of 2017-18 changes), the association between school policies and Progress 8 ratings dropped. At which point the variables accounted for less variation than any of the categories discussed thus far. It would seem therefore that these variables account for little variation that the differences in schools' intake and instructional practices cannot. Taken at face value this suggests that school policies have a meaningful impact upon school performance but one that is predominately indirect and surpassed by other influences.

There are however several reasons to be cautious when interpreting these results.

For starters, in order to reduce the length of the questionnaire the effect of school policies was only evaluated in the simple-, forward- and hierarchical analyses of the 2017-2018 changes. Whilst it was presumed that the variables had a comparable effect upon schools' annual performance ratings, this is only speculation.

Secondly, although the majority of the interactions were in line with expectations, the percentage of anticipated associations was lower than in any category of variable discussed thus far. Just 7/8 (87.5%) of the simple-linear associations, 1/2 (50%) of the associations from the forward-regression model and 2/3 (66.7%) of the hierarchical- interactions were in-line with expectations. However, the anomalies all accounted from less than 0.1% of the variation in schools' Progress 8 ratings and lacked a logical explanation. It was therefore presumed that these were non-causal associations that occurred by chance. The finding that increasing scheduled instructional time does not necessarily lead to improvements in educational effectiveness is a slight exception, though, as this outcome is arguably in line with the findings of early schools effectiveness research (Creemers and Kyriakides, 2008).

Finally, although the reported associations were logical and in line with the theoretical arguments outlined within the Dynamic Model, the rank order of variables' effects was inconsistent with past research. In fact, there was a r = -0.830 correlation between the percentages of variance that the policies explained within this analysis and the effect sizes reported within Creemers and Kyriakides (2008) meta-analysis of school-level effectiveness factors. In the aforementioned study, for example, the effect of policies exceeded the influence of schools' evaluation mechanisms. This disagreement is less alarming, however, when one recognises that the mean effect sizes that Creemers and Kyriakdies' reported for each policy were very similar to one another, and the standard deviation of effect sizes across and within studies was high (see Section 7.4.2 for further details). The analysis also evaluated the factors influence upon schools' Progress 8 scores at specified moments in time, rather than their effect upon the year-to-year changes in schools' ratings.

In light of these concerns, one cannot rule out the possibility that the three alternative interpretations from the previous sub-section might have exerted some influence upon the measures of school policies. If this was the case, then the analyses would have under-reported the influence of school-level decision making. The risk was presumed to be lower in this instance though, given the larger effect sizes reported and the fact school policies are far more stable and easy to report upon than the collective behaviour of the Year 11 teaching staff.

11.5. Conclusion

This section set out to establish whether variation in school-level Progress 8 ratings is indicative of genuine differences in school effectiveness.

Whilst all of the regression models agree that correlates from educational effectiveness research are able to account for the majority of this deviation, the evidence collected in this section suggest that external factors such as differences in schools' intake and examination practices have significant sway over schools' scores. In fact, if the regression outputs are accepted at face value then these effects may explain more than twice the variation that is attributable to differences in school quality. Such an outcome would mean that Progress 8 is not only invalid but profoundly unfair.

Throughout the analysis, however, the reader's attention has been drawn to several reasons for suspecting that the effect attributed to these biases might have been exaggerated. These included the use of cohort-level data, the categorisation of student absence as factor that is predominantly outside of schools control, the precision of the instruments used to collect data of schools' policies and instructional practices, the coverage of the underlying effectiveness constructs, the simplistic modelling of complex and interactive effects and the possibility that examination entry variables might act as proxies for school effects.

Even the most favourable interpretation of the data, which interprets student absence as being entirely under schools' control, assumes that the questionnaire was only able to report a fraction of the effect that school-related factors were actually responsible for, completely ignores any effect attributed to schools' examination entry practices and attributes any association between intake factor and performance to the peer effect, however, must still conclude that the differences in the composition of school intakes still accounted for between 13.0% and 25.9% of the variation in schools' annual Progress 8 ratings and between 2.3% to 2.8% of the change in schools' progress scores over time.[35]

The one finding that cannot therefore be disputed is that Progress 8 provides a biased measure of Type B effects that favours schools' with advantaged intakes.

[35] These figures refer to the lowest and highest percentage of variance explained by intake factors within the multiple regression models, when the two absence variables were excluded from the calculation. See Table 11.4a for further details.

12. Detailed Regression Analysis

12.1 Chapter Introduction

The last chapter used regression analysis to examine the validity of school-level Progress 8 assessments. Specifically, it tested whether the differences in schools' annual performance ratings and the change in schools' ratings over time could be explained by correlates from educational effectiveness research. A notable shortfall of the evaluation, however, was its scope. More attention was paided to differences in schools' instructional practices than to deviations in policy and several dimensions of effectiveness were neglected. This may ultimately have given a false impression of the variance could be accounted for by school-related factors. This chapter therefore takes a closer look at the performance of 9 schools by operationalising a far wider range of effectiveness variables and studying their impact upon schools' performance ratings. The increased scrutiny however came at a price, namely that the achieved sample size was modest and only able to support basic forms of statistical analysis. The interpretation of data therefore relies upon more speculative methods.

12.2 Method Section

12.2.1. Research sample

As previously stated, all state-funded mainstream schools in England are obligated to take part in Progress 8 assessments. In 2018, after excluding pupil referral units and schools that did not educated students from the age of 11 to 16, the population then referred to 2991 schools (EduBase, 2018). The researcher visited all eligible institutions within a 50 mile radius of Durham. Nine of these were selected to take part in a case study. All were from the North England and all took part on a voluntary basis.

More specifically, the sample contained one converter academy, two sponsor-led academies, three community schools and three mainstream foundation schools. These schools were slightly smaller than most, having a mean cohort size of 110.2 students in 2018 (sd. = 54.5). Though the percentage of students included within the measure was very close to the national average (95.5%). The other characteristics of the school cohorts were fairly typical, with the mean percentages of disadvantaged, female and non-mobile students being 28.7%, 49.9% and 94.2% respectively. While 1.4% of students had Special Educational Needs (SEN) and a Statement or Educational Health Care plan (EHC), and 14.8% had SEN but no Statement or EHC plan. The main compositional distinction was therefore that students that spoke English as an additional language were under-represented, with the mean proportion of EAL students per Year 11 cohort being just 3.1% (the national average in 2018 was 16.3%). In most respects the sampled schools were therefore representative of the average state-funded school.

The most pertinent distinction, however, was that high achieving schools (those with high Attainment 8 scores) and 'effective' schools (those with high Progress 8 scores) were under-represented within the sample. The mean average Attainment 8 score of the sampled schools was therefore 42.311, whilst the mean average Attainment 8 score within the population was 47.334. Similarly the mean Progress 8 score was -0.359 (sd = 0.532) within the sample, and 0.013 (sd=0.449) within the population. The assessed range of scores was also skewed (range -1.58 to 0.22). However, as a reasonable range of scores was assessed, the effect of these deviations was assumed to be minimal. It should also help that the primary intent was to report upon the explicability of schools' ratings rather than to make inferences to the overall population of state schools.

Should the reader be interested Section 11.2.1 provides more in-depth of information on the population of state-funded mainstream schools. It is stressed, however, that this analysis was intended to act as a case study that provides detailed insight into the factors than impacted upon the performance of the sampled schools. It is not, therefore claimed that the findings can be inferred to other contexts uncritically.

12.2.2. Recoding of missing, suppressed and incompatible data items

To enable the inclusion of all 9 respondents, concession were made.

Firstly, 2.8% of the data required for the analyses was missing. In all cases this was because the respondent had failed to answer a question in their questionnaire (3.1%, 1.5% and 3.2% of questions in 2017, 2018 and 2017-18 change analyses respectively).

Similarly, 1.2% of questionnaire responses were coded as N/A (1.0% of items in the 2017 analysis, 1.3% of items in the 2018 analysis and 1.3% of items in the 2017-18 change analysis). These were responses that, although technically valid, indicate that the question did not apply to their school. For example, a respondent may have selected this answer when a question enquired about the specificity of a policy that did not exist.

Both types of response were coded as the mean for the variable. This allowed the respondent to be retained within the sample whilst preventing their response from interfering with the analysis of the specified factor.

Finally, 8 of the variables that were sourced from the National Pupil Database contained suppressed responses (2 variables from the 2017 analysis, 3 variables from the 2018 analysis and 3 variables from the change analysis). This is done to protect the identity of individuals when between 1 and 5 students per school possess a particular characteristics. These redactions influenced 0.7% of the reported data items (0.8%, 0.5% and 0.9% in the 2017, 2018 and change analysis respectively). To address the problem all schools with suppressed data were assumed to have 3 students within the category. This is the mid-point of the possible range and therefore limits the potential for error.

In most cases the effect of the inaccuracies should therefore have been minimal. A major exception to this was, however, the variable that reported upon percentage of Progress 8 entrants that spoke English as an additional language. 6 of the 9 data items that were used to calculate these percentages were suppressed within each analysis. To remedy this, the percentage of Year 11 students that spoke English as an additional language was also recorded. This is a comparable, though slightly less specific measure that did not contain suppressed data. Since the two populations (Progress 8 entrants and Year 11 students) are not identical and there was no way of establishing which was the more accurate statistic, both measures were retained within the analysis.

12.2.3. Research design

The methodology of this section was comparable to the first element of the Shallow Regression Analyses.

In order to establish whether the variation in schools' Progress 8 scores could be explained by the kinds of factors that school performance is normally attributed to, three sets of simple linear-regression analyses were performed. These reported upon the relationship that existed between established effectiveness factors and schools' ratings from 2017, schools' ratings from 2018 and the change in schools' ratings between the 2017 and 2018 assessments.

The results were interpreted based upon the direction of the associations, their magnitude and whether the relationships were consistent with the interactions theorised within academic research. That is to say, the impact that the variables have upon students' raw attainment (see Section 7.3 and 11.2.4 for further details). The most important consideration, however, was the proportion of variation that could be explained by factors that are within and outside of schools' control. Theoretically, if Progress 8 provides a valid and unbiased measure of school effectiveness then high proportions of the results should be explained by the former group, whilst the latter should have a limited impact. The analysis assesses whether this was the case.

There was, however, a weakness in this research methodology that must be acknowledged. During regression analyses the number of independent variables that can be operationalised is contingent upon the size of one's sample. Most statisticians therefore recommend that researchers maintain a ratio of at least 10 observations for every independent variable included within regression models (Agresti and Franklin, 2014). This helps to ensure that the variation that is ascribed to a particular factor has not occurred by chance. Though there is enough flexibility in this figure to legitimise the current approach, extraneous influences upon the relationships could not be taken into account. When interpreting the results it is therefore important to recall that the reported effect sizes refer to the percentage of variance that the factors could explain, not the percentage that they are causally responsible for. The results still provide useful information, however, as it is unlikely that variables which are ascribed low r-squared scores are hiding much variation. Duplicating the analysis across two academic years (2017 and 2018) therefore helped to rule out coincidental associations, whilst the consideration of 2017-2018 changes established the time-order of events.

It should also be noted that the same shortfall prevented the analysis from evaluating the combined effect of variables. In other words to summate the influence of school-related and non-school factors. When interpreting the data one is therefore forced to presume that the areas which accounted for the highest percentage of the variation in Progress 8 scores on average, explained the most variation collectively. Whether this is actually the case will depend on the level of multicollinearity between variables. The results of the preceding analyses however suggest that this assumption is valid (see Chapter 11).

12.2.4 The selection of independent variables

To help ensure that the most important effectiveness factors were considered, the selection of independent variables was based upon findings of educational effectiveness research. Specifically, the analysis utilised effectiveness factors from the Dynamic Model of Educational Effectiveness (Creemers and Kyriakidies, 2008). Additional variables were also considered, if there was a logical reason for their inclusion, including several intake factors and examination entry differences that were likely to impact upon schools' progression scores, the consistency of classroom practices and the time dedicated to particular subject areas.

All supplementary intake variables (measures relating to disadvantage or absence) were expected to have a negative and linear association with attainment, except for the percentage of non-mobile students per cohort and the percentage of students' speaking English as an additional language, which were expected to have a positive and linear correlation. All examination entry variables were assumed to have a positive association with school performance, with the exception of the number and percentage of Year 11 students' included within schools' calculations. Likewise, the more time that was dedicated to a particular subject area and/or the more consistent teachers' practices the more favourable schools' results were expected to be. These expectations were discussed in detail within previous chapters of this thesis. See Section 7.3 for more information of variables included within the Dynamic Model and Section 11.2.4 for further information on the stated additions.

The main difference between this analysis and the previous one, however, is that several dimension of each factor were considered. This enabled the analysis to explore the impact of qualitative differences such as the specificity of actions, their purpose, timing, quality and the level of implementation support available.

The variables that were considered in each analysis:

Intake variables

The following intake variables were sourced from the National Pupil Database performance tables.

Table 12.2.4a: Intake variables included in the Detailed Regression Analyses

N	Variable Name	2017 Analysis	2018 Analysis	Change Analysis
1	Overall percentage of absence at the school	Yes	Yes	Yes
2	Percentage of persistent absentees at the school	Yes	Yes	Yes
3	Percentage of Progress 8 entrants that were disadvantaged	Yes	Yes	Yes
4	Percentage of Progress 8 entrants that spoke English as an additional language	Yes	Yes	Yes
5	Percentage of Year 11 pupils that spoke English as an additional language	Yes	Yes	Yes
6	Percentage of girls in the Progress 8 measure	Yes	Yes	Yes
7	Percentage of Year 11 pupils that had SEN and a Statement or EHC plan	Yes	Yes	Yes
8	Percentage of Year 11 with SEN but no Statement or EHC plan	Yes	Yes	Yes
9	Percentage of Year 11 pupils with SEN (with or without Statement/EHC plan)	Yes	Yes	Yes
10	Percentage of Progress 8 entrants that were non-mobile	Yes	Yes	Yes

Instructional practices

Details of schools' instructional practices were collected using a school-leader questionnaire that was completed between March 2018 and July 2018 (see Appendix D).

Variables 11-22 assess the frequency the specified behaviours, 23-38 report upon their focus and 39-57 describe the timing of actions. Variables 58-72 examine the quality of teachers' instructional behaviour, 73-81 report the level of differentiation that took place, Variables 82-85 the consistency of behaviours and Variable 86 teachers' coverage of the school curriculum (see Table 12.2.4b).

Table 12.2.4b: Instructional variables included in the Detailed Regression Analyses

N	Dimension	Variable Name	2017 Analysis	2018 Analysis	Change Analysis
11	Frequency	Frequency of orientation tasks	Yes	Yes	Yes
12		Frequency of structuring tasks	Yes	Yes	Yes
13		Frequency of questioning	Yes	Yes	Yes
14		Frequency of open-ended questions	Yes	Yes	Yes
15		Frequency of teacher-modelling tasks	Yes	Yes	Yes
16		Frequency of application tasks	Yes	Yes	Yes
17		Frequency of teacher-student interactions	Yes	Yes	Yes
18		Frequency of student-student interactions	Yes	Yes	Yes
19		Proportion of lesson time that was used for teaching	No	No	Yes
20		Frequency of classroom disruptions	Yes	Yes	Yes
21		How frequently teacher responded to classroom disruptions	Yes	Yes	Yes
22		Frequency of classroom assessments	Yes	Yes	Yes
23	Focus	Whether orientation tasks referred to a series, the whole or part of lessons	Yes	Yes	Yes
24		Number of objectives behind each orientation tasks	No	No	Yes
25		Whether structuring tasks referred to a series, the whole or part of lessons	Yes	Yes	Yes
26		Number of objectives behind each structuring tasks	No	No	Yes
27		Whether questioning referred to a series, the whole or part of lessons	Yes	Yes	Yes
28		Number of objectives behind each questioning tasks	No	No	Yes
29		Number of circumstances that teacher-modelling tasks could be applied to	No	No	Yes
30		Number of times teachers introduced more than one strategy for solving a problem	No	No	Yes
31		Whether application tasks referred to a series, the whole or part of lessons	Yes	Yes	Yes
32		Number of objectives behind application tasks	No	No	Yes
33		Proportion of teacher-student interactions that were task-related	No	No	Yes
34		Proportion of student-student interactions that were task-related	No	No	Yes
35		Proportion of classroom disruptions that were due to previously unresolved issues	No	No	Yes
36		Extent that teachers attempted to address the issue behind disruptions	No	No	Yes
37		Change in the range of assessment methods	No	No	Yes
38		Number of objectives behind each classroom assessment task	No	No	Yes
39	Stage	Parts of the lesson in which orientation tasks consistently took place*	Yes	Yes	Yes
40		Parts of the year in which orientation tasks consistently took place*	Yes	Yes	Yes
41		The extent to which teachers orientation tasks take on board students' perspective	No	No	Yes
42		Parts of the lesson in which structuring tasks consistently took place consistently*	Yes	Yes	Yes
43		Parts of the year in which structuring tasks consistently took place*	Yes	Yes	Yes
44		Parts of the lesson in which questioning tasks consistently took place consistently*	Yes	Yes	Yes
45		Parts of the year in which questioning tasks consistently took place*	Yes	Yes	Yes
46		The proportion of teacher-modelling tasks which introduced strategies after the problem	No	No	Yes
47		Parts of the lesson in which application tasks consistently took place*	Yes	Yes	Yes
48		Parts of the year in which application tasks consistently took place*	Yes	Yes	Yes
49		Parts of the lesson in which teacher-student interactions consistently took place*	Yes	Yes	Yes

50		Parts of the year in which teacher-student interactions consistently took place*	Yes	Yes	Yes
51		Parts of the lesson in which student-student interactions consistently took place*	Yes	Yes	Yes
52		Parts of the year in which student-student interactions consistently took place*	Yes	Yes	Yes
53		Parts of the lesson in which classroom disruptions consistently took place*	Yes	Yes	Yes
54		Parts of the year in which classroom disruptions consistently took place*	Yes	Yes	Yes
55		Parts of the lesson in which classroom assessments consistently took place*	Yes	Yes	Yes
56		Parts of the year in which classroom assessment tasks consistently took place*	Yes	Yes	Yes
57		Speed which classroom assessments were analysed, reported and acted upon	No	No	Yes
58	Quality	Clarity of orientation tasks	No	No	Yes
59		Influence of orientation tasks on students' learning	No	No	Yes
60		Clarity of structuring tasks	No	No	Yes
61		Influence that structuring tasks had on students' learning	No	No	Yes
62		Extent to which lessons and schemes of work were structured so that the easier tasks preceded the difficult ones	No	No	Yes
63		Clarity of questioning	No	No	Yes
64		Appropriateness of question difficulty	No	No	Yes
65		Extent that teachers sustained their interaction with the original respondent during questioning by rephrasing and giving clues	No	No	Yes
66		Clarity with which problem-solving strategies were introduced	No	No	Yes
67		Extent that application tasks expanded on the material that was taught in the lessons	No	No	Yes
68		Extent that teachers' interventions were able to establish the desired form of interaction	No	No	Yes
69		Extent that teachers interventions solved the underlying issues behind classroom disruptions	No	No	Yes
70		Extent that classroom assessments measured what they were intended to measure)	No	No	Yes
71		Amount of constructive feedback that was given to students after classroom assessments	No	No	Yes
72		Influence of assessments on students' learning	No	No	Yes
73	Differentiation	Teachers' ability to adapt orientation tasks to meet students' individual needs	No	No	Yes
74		Teachers' ability to adapt structuring tasks to meet students' individual needs	No	No	Yes
75		Teachers' ability to adapt questioning tasks to meet students' individual needs	No	No	Yes
76		Teachers' ability to adapt teacher-modelling tasks to meet students' individual needs	No	No	Yes
77		Teachers' ability to adapt application tasks to meet students' individual needs	No	No	Yes
78		Teachers' ability to adapt their strategies for establishing on-task behaviour to meet students' individual needs	No	No	Yes
79		Teachers' ability to adapt their strategies for dealing with classroom disruptions to meet students' individual needs	No	No	Yes
80		Teachers' ability to adapt the allocation of lesson time around students' individual needs	No	No	Yes
81		Teachers' ability to adapt classroom assessments and feedback to meet students' individual needs	No	No	Yes
82	Consistency	Consistency in the proportion of lesson time that was used for teaching	No	Yes	Yes
83		Consistency in teachers' coverage of the school curriculum	No	Yes	Yes
84		Consistency in the quality of teachers' instruction	No	Yes	Yes
85		Consistency of teaching style(s) used by teachers	No	Yes	Yes
86	N/A	Teachers' coverage of the school curriculum	No	No	Yes

*Lessons and the school year were dissected into three segments; their beginning, middle and end. These variables were then scored from 0-3 based on the number of segment in which the behaviour was performed.

Under ideal circumstances a more direct method of assessment, such as teacher diaries or lesson observations, would have been preferable. These techniques get closer to the educational process and are therefore are more likely to detect instructional effects. Unfortunately it was not possible to secure the requisite level of access to teachers or their classrooms. A direct evaluation of instructional behaviours would also have been complicated by the fact that most students included within schools' Progress 8 calculations would have studied different combinations of subjects. As school leaders are arguably the ultimate authorities on their school, however, it is reasonable to assume that they will be aware of major shifts in pedagogical practice.

School policies

The characteristics of schools' policies were also evaluated using the school-leader questionnaire. It is important to recognise, though, that the term 'school policy' is used to refer to any formal or informal communication that helps to standardise the schools' approach. The production of documentation is assumed to have little to no effect on students' performance unless accompanied by other communicative efforts. This definition was made clear to school-leaders.

Variables 87-94 were used to evaluate the scope of schools' policies (officially classified as the 'frequency dimension' of the policies.), Variables 95-106 their focus and Variables 107-122 the duration of their implementation. Variable 123-150 report upon the quality of the policies and Variables 151-158 upon the level of differentiation that they permitted. The final set of variables report upon the instructional time dedicated to specific areas (see Table 12.2.4d).

Table 12.2.4c: School policy variables included in the Detailed Regression Analyses

N	Dimension	Variable Name	2017 Analysis	2018 Analysis	Change Analysis
87	Frequency	Coverage of the quantity of instruction policies (4 policy areas)	Yes	Yes	Yes
88		Coverage of the policies for providing students with learning opportunities (9 policies areas)	Yes	Yes	Yes
89		Coverage of the schools' instructional behaviour policies (8 policy areas)	Yes	Yes	Yes
90		Coverage of the policies for creating an effective school learning environment (5 policy areas)	Yes	Yes	Yes
91		Frequency with which the school collected data on the school teaching policies	No	No	Yes
92		Number of sources of information that the evaluations of the school teaching policies drew upon	No	No	Yes
93		Frequency with which the school collected data on the school learning environment (SLE)	No	No	Yes
94		Number of sources of information that the evaluations of the policies on the school learning environment drew upon	No	No	Yes
95	Focus	The extent to which the quantity of instruction policies dictated teachers' and students' actions	Yes	Yes	Yes
96		Number of objectives that were pursued by the quantity of instruction policies	Yes	Yes	Yes
97		The extent to which the policies on the provision of learning opportunities dictated teachers' and students' actions	Yes	Yes	Yes
98		Number of objectives that were pursued by the learning opportunity policies	Yes	Yes	Yes
99		The extent to which the policies on teachers' instructional behaviour dictated teachers' and students' actions	Yes	Yes	Yes
100		: The number of objectives pursued by the policies on	Yes	Yes	Yes

		teachers' instructional behaviour			
101		The extent to which the SLE policies dictated teachers' and students' actions	Yes	Yes	Yes
102		The number of objectives pursued by the SLE policies	Yes	Yes	Yes
103		The number of aspects of the school teaching policies that were evaluated (6 policy areas total)	Yes	Yes	Yes
104		The level of feedback generated by the evaluations of the school teaching policies	Yes	Yes	Yes
105		The number of aspects of the SLE policies that were evaluated (6 areas total)	Yes	Yes	Yes
106		The level of feedback generated by the evaluations of the SLE policies	Yes	Yes	Yes
107	Stage	Number of years that the quantity of instruction policies had been implemented	Yes	Yes	Yes
108		Average number of years between modifications of the quantity of instruction policies*	Yes	Yes	Yes
109		Whether changes to the quantity of instruction policies were based upon evaluation data	Yes	Yes	Yes
110		Number of years that the learning opportunity policies had been implemented	Yes	Yes	Yes
111		Average number of years between modifications of the learning opportunity policies*	Yes	Yes	Yes
112		Whether changes to the learning opportunity policies were based upon evaluation data	Yes	Yes	Yes
113		Number of years that the instructional behaviour policy had been implemented	Yes	Yes	Yes
114		Average number of years between modifications of the instructional behaviour policy*	Yes	Yes	Yes
115		Whether changes to the instructional behaviour policies were based upon evaluation data	Yes	Yes	Yes
116		Number of years that the SLE policies had been implemented	Yes	Yes	Yes
117		Average number of years between modifications of the SLE policies*	Yes	Yes	Yes
118		Whether changes to the SLE policies were based upon evaluation data	Yes	Yes	Yes
119		Frequency with which the school evaluated the school teaching policies	No	No	Yes
120		Whether there was a formalised procedure for evaluating the mechanisms that were used to assess the school teaching policies	Yes	Yes	Yes
121		Frequency with which the school evaluated the SLE policies	No	No	Yes
122		Whether there was a formalised procedure for evaluating the mechanisms that were used to assess the SLE policies	Yes	Yes	Yes
123	Quality	Clarity of the quantity of instruction policies	No	No	Yes
124		Alignment between the quantity of instruction policies and the academic literature	No	No	Yes
125		Level of support provided to teachers and/or students to implement the quantity of instruction policies	No	No	Yes
126		Influence that the quantity of instruction policies had on teachers’ and students’ behaviour	No	No	Yes
127		Clarity of the learning opportunity policies	No	No	Yes
128		Alignment between the learning opportunity policies and the academic literature	No	No	Yes
129		Level of support provided to teachers and/or students to implement the learning opportunity policies	No	No	Yes
130		Influence that the learning opportunity policies had on teachers’ and students’ behaviour	No	No	Yes
131		Clarity of instructional behaviour policy	No	No	Yes
132		Alignment between the instructional behaviour and the academic literature	No	No	Yes
133		Level of support provided to teachers and/or students to implement the instructional behaviour policy	No	No	Yes
134		Influence that the instructional behaviour policy had on teachers’ and students’ behaviour	No	No	Yes
135		Clarity of the SLE policies	No	No	Yes

136		Alignment between the SLE policies and the academic literature	No	No	Yes
137		Level of support provided to teachers and/or students to implement the SLE policies	No	No	Yes
138		Influence that the SLE policies had on teachers' and students' behaviour	No	No	Yes
139		Reliability of the mechanisms used to evaluate the school teaching policies	Yes	Yes	Yes
140		Proportion of evaluation data that was used to inform decisions about school teaching policies	Yes	Yes	Yes
141		Extent to which evaluations of the school teaching policies assessed the factors that they were intended to assess (face validity)	Yes	Yes	Yes
142		Strength of the relationship between the evaluations of the school teaching policies and students' learning	Yes	Yes	Yes
143		Extent to which the benefits of monitoring the school teaching policy outweighed the drawbacks	Yes	Yes	Yes
144		Reliability of the mechanisms used to evaluate the SLE policies	Yes	Yes	Yes
145		Proportion of evaluation data that was used to inform decisions about SLE policies	Yes	Yes	Yes
146		Extent to which evaluations of the SLE policies assessed the factors that they were intended to assess (face validity)	Yes	Yes	Yes
147		Strength of the relationship between the evaluations of the SLE policies and students' learning	Yes	Yes	Yes
148		Extent to which the benefits of monitoring the SLE policy outweighed the drawbacks	Yes	Yes	Yes
149		Amount of instruction time that was provided to students by the school policies (quantity of instruction polices)	No	No	Yes
150		The alignment between the school curriculum and the content assessed at KS4 (learning opportunity policy)	No	No	Yes
151	Differentiation	Level of differentiation in the quantity of instruction policies	No	No	Yes
152		Level of differentiation in the learning opportunity policies	No	No	Yes
153		Extent to which teachers were encouraged to differentiate the learning opportunities that they offer to students	No	No	Yes
154		Level of differentiation in the instructional behaviour policies	No	No	Yes
155		Extent to which teachers were encouraged to differentiate their use of the 8 instructional behaviours	No	No	Yes
156		Level of differentiation in the SLE policies	No	No	Yes
157		Emphasis that was placed on evaluating the under-performing aspects the schools' teaching provisions	No	No	Yes
158		Emphasis placed on evaluating the underperforming aspects of the SLE	No	No	Yes
159	Allocation of instructional time	Instruction time dedicated to Mathematics	No	No	Yes
160		Instruction time dedicated to English Language and English Literature	No	No	Yes
161		Instruction time dedicated to other EBacc subjects	No	No	Yes
162		Instruction time dedicated to Non-EBacc GCSEs and Non-GCSEs	No	No	Yes
163		Instruction time dedicated to Level 3 qualifications	No	No	Yes

* Measures referred to changes that were made over the last 12 months. When no changes had taken place school leaders were asked to specify whether this decision was based upon evaluation data.

Examination entry variables

The final category of variables identified differences in schools' examination entry practices. These are listed below.

Table 12.2.4d: Examination entry variables included in the Detailed Regression Analyses

N	Variable Name	2017 Analysis	2018 Analysis	Change Analysis
164	Number of pupils included in Progress 8 measure	Yes	Yes	Yes
165	Percentage of Year 11 pupils entered into Progress 8	Yes	Yes	Yes
166	Percentage of Year 11 entering Baccalaureate Maths	Yes	Yes	Yes
167	Percentage of Year 11 entering Baccalaureate English	Yes	Yes	Yes
168	Percentage of Year 11 entering Baccalaureate Science	Yes	Yes	Yes
169	Percentage of Year 11 entering Baccalaureate Humanities	Yes	Yes	Yes
170	Percentage of Year 11 entering Baccalaureate Language	Yes	Yes	Yes
171	Average number of EBacc slots filled in Attainment 8	Yes	Yes	Yes
172	Average number of open slots filled in Attainment 8	Yes	Yes	Yes
173	Percentage of Year 11 entering all English Baccalaureate subject areas	Yes	Yes	Yes
174	Average number of GCSE entries per pupil (not including equivalencies)	Yes	Yes	Yes
175	Average number of GCSE entries per pupil (including equivalencies)	Yes	Yes	Yes
176	Number of students entered for Level 3 qualifications (AS Levels)	No	No	Yes

All variables were sourced from publically available NPD data.

Contextual variables

Context or 'system level' interactions were not considered during the assessment. This is because all state-funded schools are governed by the comparable policies and evaluation procedures. Mid-level organisational bodies, such as Local Educational Authorities have also been shown to have a limited impact upon students' performance (Tymms et al., 2008). The only contextual-level influences from the Dynamic Model that would have been operationalised under ideal circumstances are therefore the influence of local stakeholders and support from the local community. The collection of this data was however considered unfeasible.

12.2.5. Measurement scales

As per the Shallow Regression Analysis, all NPD data sets contained ratio-level data, whilst questionnaire responses were report on ordinal or dichotomous scales. All data, however, was treated as ratio-level to allow the analysis to take place. This may have added some construct irrelevant variance into the assessment of schools' instructional practices and policies. See Appendix D for further information.

12.3 Results

2017 Analysis

The association between each independent variable and school performance is described in Appendix E.

Part 1: The average variance explained by each category of variable:

On average the operationalised variables explained 16.6% of the variation in schools' performance ratings. However, the strength of these relationships varied dramatically with some variables accounting for as much as 76.4% of the deviation in schools' results and other predicting less than 0.1%.

Table 12.3a: The average variation in Progress 8 scores explained by each classification of variable in the 2017 Detailed Regression Analysis

Category:	Average linear r-squared of group (3dp):	Average percentage of variance explained (1dp):
School intake variables	0.217	21.7%
Instructional behaviours	0.190	19.0%
School policy variables	0.074	7.4%
Examination entry variables	0.367	36.7%

Examination Entry Variables:

Examination entry variables were the most effective predictors of school performance (see Table 12.3a). These factors accounted for an average of 36.7% of the variation in schools' ratings (range 65.6% to 0.2%). To help interpret the results, these variables were grouped into the same five sub-groups that were utilised during the shallow regression analyse. Namely, the percentage of students entering all English Baccalaureate subject areas (1 variable), the average number of EBacc and Open slots filled by the pupils from each school (2 variables), the average number of GCSEs that students entered including and excluding equivalencies (2 variables), the percentage of students that were entered for specific subject areas (5 variables) and the number and percentage of Year 11 pupils included within schools' Progress 8 calculations (2 variables).

The most influential sub-group was the percentage of students that entered all English Baccalaureate subject areas. This variable had a strong positive association with performance, meaning that schools with the higher proportions of students entering the English Baccalaureate tended to outperform schools with lower entry rates. This relationship accounted for 62.8% of the variation in schools' ratings.

The average number of GCSEs entered by students was the second most informative group. These variables had a positive association with school performance that explained an average of 39.1% of the variation in schools' ratings. Though, the statistic was more informative when GCSE equivalent qualifications were included (variance explained = 39.8%) rather than excluded (variance explained = 38.4%).

The next most influential factors were the average number of EBacc and Open slots filled in schools' Attainment 8 calculations. The average number of EBacc slots filled by the students at each school accounted for 31.6% of the variation in schools' results and coverage of the open slots explained 45.9%. This made the average effect size for this sub-category 38.8%. Both associations were positive meaning that schools with the best coverage performed better on average than the schools with lesser coverage.

The number and percentage of Year 11 pupils entered into schools' Progress 8 calculations also had a sizeable association with school performance. These variables accounted for an average of 32.9% of the results. However, whilst the size of schools had a strong positive association with schools results that accounted for 65.6% of the variation, the percentage of Year 11 pupils had a negative association than accounted for 0.2% of deviation in schools' ratings.

The least informative sub-group of examination-entry variables was therefore the entry rates for individual subject areas (average variance explained = 31.3%). Even so, all accounted for meaningful proportions of school performance. The most predictive indicator was the percentage of students with entries in the EBacc language subject areas (variance explained = 56.8%), followed by the percentage with entries in the humanities (38.4%), English (38.0%) and maths (15.5%). The least predictive subject entry rate was the sciences (7.9%).

Intake Variables:

Differences in school intakes were the second most informative category of variable. On average these indicators were able to account for 21.7% of the variation in schools' performance (range 54.2% to 0.0% (1dp)).

The most influential factor was student absence (average variance explained = 43.1%). Both the percentage of absence at each school and the percentage of persistent absentees had strong negative associations with school performance, which suggests that the more instruction students missed the more detrimental the effect upon their learning (variance explained 32.0% and 54.2% respectively).

The three SEN variables were also informative (average variation = 27.1%). Interestingly, each measure had a disparate relationship with school performance. The percentage Year 11 pupils with SEN and a Statement or EHC plan had a negative association with school performance (variance explained = 35.2%), whilst the percentages of Year 11 with SEN but no Statement or EHC plan and the overall percentage of students with SEN (with or without a Statement/EHC plan) had positive linear associations with schools' results (variance explained = 27.3% and 18.7% respectively).

Finally, the percentage of Progress 8 entrants that spoke English as an additional language and the percentage of Year 11 student that spoke English as an additional language explained meaningful percentages of the variation in school performance (26.6% and 21.6% respectively), making the average effect size attributed to this sub-group 24.1%.

The remaining intake factors were less predictive. The percentage of female students per cohort and the percentage of non-mobile students per cohort, for example, had negative associations with school performance that accounted for 1.5% and 0.0% (1dp) of the results respectively. The percentage of disadvantage students had a positive association that explained 0.0% (1dp).

Instructional Practices:

Instructional practices were the next most influential classification of variable. On average these measures explained 19.0% of the variation in schools' performance (range = 76.6% to 0.0%).

Table 12.3b: The relationship between classroom instructional practices and schools' 2017 Progress 8 scores

Rank	Instructional behaviour	Variables considered	Linear association	Average r-squared
1	Classroom learning environment	1. Frequency of teacher-student interactions	Positive (r^2=0.694)	0.321
		2. Stages of lesson in which the teacher-student interactions consistently took place	Negative (r^2=0.020)	
		3. Stages of academic year in which teacher-student interactions consistently took place	Positive (r^2=0.084)	
		4. Frequency of student-student interactions	Positive (r^2=0.709)	
		5. Stages of lesson in which student-student interactions consistently took place	Positive (r^2=0.000)	
		6. Stages of academic year in which student-student interactions consistently took place	Positive (r^2=0.084)	
		7. Frequency of classroom disruptions	Negative (r^2=0.381)	
		8. How frequently teachers responded to classroom disruptions	Negative (r^2=0.764)	
		9. Stages of lesson in which classroom disruptions consistently took place activity takes place	Negative (r^2=0.086)	
		10. Stages of academic year in which classroom disruptions consistently took place	Negative (r^2=0.384)	
2	Questioning	1. Frequency of questioning	Positive (r^2=0.321)	0.181
		2. Frequency of open-ended questions	Positive (r^2=0.358)	
		3. Whether questioning typically refer to a series, whole or part of the lessons	Negative (r^2=0.001)	
		4. Stages of lesson in which questioning tasks consistently took place	Negative (r^2=0.112)	
		5. Stages of academic year in which questioning tasks consistently took place	Negative (r^2=0.113)	
3	Orientation	1. Frequency of orientation tasks	Negative (r^2=0.005)	0.143
		2. Whether orientation tasks typically referred to a series, whole or part of the lesson	Positive (r^2=0.367)	

		3. Stages of lesson in which orientation tasks consistently took place	Negative (r^2=0.170)	
		4. Stages of the academic year in which orientation tasks consistently took place	Positive (r^2=0.029)	
4	Teacher-modelling	1. Frequency of teacher-modelling tasks	Positive (r^2=0.140)	0.140
5	Structuring	1. Frequency of structuring tasks	Positive (r^2=0.042)	0.126
		2. Whether structuring tasks typically referred to a series, whole or part of the lesson	Positive (r^2=0.267)	
		3. Stages of lesson in which structuring tasks consistently took place	Negative (r^2=0.170)	
		4. Stages of academic year in which structuring tasks consistently took place	Positive (r^2=0.026)	
6	Application	1. Frequency of application tasks	Positive (r^2=0.166)	0.102
		2. Whether application tasks typically referred to a series, whole or part of the lessons	Positive (r^2=0.117)	
		3. Stages of lesson during which application tasks consistently took place	Negative (r^2=0.042)	
		4. Stages of academic year in which application tasks consistently took place	Positive (r^2=0.084)	
7	Classroom assessment	1. Frequency of classroom assessment tasks	Positive (r^2=0.076)	0.051
		2. Stages of lesson in which classroom assessments consistently took place	Positive (r^2=0.062)	
		3. Stages of academic year in which classroom assessments consistently took place	Positive (r^2=0.016)	

*All of the variables that evaluate classroom disruptions and teachers' ability to deal with them also impact upon the percentage of lesson time that is utilised productively. The average percentage of variance explained by teacher' ability to manage lesson time is not reported within the 2017 analysis however as no unique variables were considered.

The three aspects of the classroom learning environment were the most effective indicators of school performance (average variance explained = 0.321). Of the three, classroom disruptions accounted for the highest percentage of performance (average variance explained = 40.4%), followed by teacher-student interactions (average variance explained = 26.6%) and student-student interactions (average variance explained = 26.4%).

The remaining instructional behaviours accounted for comparable proportions of variance. The various aspects of questioning explained an average of 18.1% of the variation in schools' scores, the orientation variables 14.3%, Teacher-modelling 14.0% and structuring 12.6%. The application and classroom assessment variables accounted for an average of 10.2% and 5.1% of the variation in schools' scores respectively.

The direction of these associations is reported in the table above.

School Policies:

On average, school policy variables had the least association with school performance. The mean effect size of these variables was 7.4%, and the range 37.4% to 0.0%.

Table 12.3c: The relationship between school policy factors and schools' 2017 Progress 8 scores

Rank	Policy Area	Variables considered	Linear association	Average R-squared
1	Evaluation of the school teaching policies	1. Aspects of the school teaching policies that were evaluated (6 policy areas)	Positive (r^2=0.011)	0.108
		2. Level of feedback generated by the evaluation of school teaching policies	Positive (r^2=0.001)	
		3. Whether there was a formalised procedure for evaluating the mechanisms that are used to assess the school teaching policies	Positive (r^2=0.017)	
		4. Reliability of mechanisms that were used to evaluate the school teaching policies	Positive (r^2=0.374)	
		5. Proportion of evaluation data that was used formatively; teaching evaluations.	Positive (r^2=0.173)	
		6. Face validity of the mechanisms that used to evaluate the school teaching policies	Positive (r^2=0.108)	
		7. Influence that the evaluations of school teaching policies had upon students' learning	Positive (r^2=0.051)	
		8. Extent to which the benefits of evaluating the school teaching policies outweighed the drawbacks	Positive (r^2=0.127)	
2	School teaching policies	1. Coverage of quantity of instruction policy areas (4 policy areas)	Positive (r^2=0.121)	0.086
		2. Coverage of learning opportunity policies (9 policy areas)	Positive (r^2=0.001)	
		3. How many of the 8 effective teaching behaviours were covered by the school teaching policies (8 policy areas)	Negative (r^2=0.001)	
		4. Extent that the quantity of instruction policies dictated teachers' and students' actions	Negative (r^2=0.169)	
		5. Number of objectives pursued by quantity of instruction policies	Negative (r^2=0.195)	
		6. Extent that the policies on the provision of learning opportunities dictated teachers' and students' actions	Negative (r^2=0.146)	
		7. Number of objectives pursued by the policies on the provision of learning opportunities	Negative (r^2=0.195)	
		8. Extent that the policies on teachers' instructional behaviour dictated teachers' and students' actions	Positive (r^2=0.045)	
		9. Number of objectives pursued by the policies on teachers' instructional behaviours.	Negative (r^2=0.077)	

		10. Number of years the current quantity of instruction policies had been implemented	Negative (r^2=0.004)	
		11. Average number of years between modifications of the quantity of instruction policies	Negative (r^2=0.047)	
		12. Whether modifications to the quantity of instruction policies were based upon data from systematic evaluations	Negative (r^2=0.000)	
		13. Number of years that the current learning opportunity policy had been implemented	Negative (r^2=0.098)	
		14. Average number of years between modifications of the school's learning opportunity policies.	Negative (r^2=0.020)	
		15. Whether modifications to the learning opportunity policies were based upon data from systematic evaluations	Positive (r^2=0.032)	
		16. Number of years that the current policies on teachers' instructional behaviours had been implemented	Positive (r^2=0.334)	
		17. Average number of years between modifications of the policies on teachers' instructional behaviours	Positive (r^2=0.025)	
		18. Whether modifications to the instructional behaviour policies were based upon data from systematic evaluations	Positive (r^2=0.032)	
3	SLE policies	1. Coverage of the SLE policies (5 policy areas)	Negative (r^2=0.005)	0.068
		2. Extent that the SLE policies dictated teachers' and students' actions	Positive (r^2=0.060)	
		3. Number of objectives pursued by SLE policies	Negative (r^2=0.202)	
		4. Number of years that the current SLE policies had been implemented?	Positive (r^2=0.087)	
		5. Average number of years between modifications of SLE policies	Positive (r^2=0.024)	
		6. Were modifications in the SLE policies were based upon data from systematic evaluations	Positive (r^2=0.032)	
4	Evaluation of the SLE policies	1. Aspects of the SLE policies that were evaluated (6 aspects total)	Positive (r^2=0.000)	0.018
		2. Level of feedback generated by the evaluations of the SLE	Positive (r^2=0.002)	
		3. Whether there was a formalised procedure for evaluating the mechanisms that were used to assess the SLE policies	Positive (r^2=0.017)	
		4. Reliability of mechanisms that were used to evaluate the SLE	Positive (r^2=0.053)	
		5. Proportion of evaluation data that was used formatively; SLE evaluations.	Negative (r^2=0.001)	
		6. Face validity of the mechanisms that were used to evaluate the SLE policies	Negative (r^2=0.001)	
		7. Influence that the evaluations of the SLE polices had upon students'	Negative	

		learning	(r^2=0.056)	
		8. Extent to which the benefits of evaluating the SLE policies outweighed the drawbacks	Negative (r^2=0.011)	

Variables concerned with the evaluation of the school teaching policy explained the most variation (average variance explained = 10.8%), followed by those associated with the school teaching policies (average variance explained = 8.6%) and the policies that regulated the school learning environment (average variance explained = 6.8%). The variables that reported upon the procedures for evaluating the school learning environment explained the least (average variance explained = 1.8%).

The three sub-divisions of the school teaching policies accounted for similar proportions of variance, with the quantity of instruction variables accounting for an average of 8.9% of the variance in schools' performance, the instructional behaviour policies 8.6% and the learning opportunity policies 8.2% (see Table 12.3c).

Part 2: The average variance explained by each dimension of effectiveness

Instructional Practice:

Within this analysis three dimensions of instructional practices were evaluated; the frequency, focus and timing of effective behaviours.

Table 12.3d: The average variance in Progress 8 scores explained by each dimension of instructional practices in the 2017 Detailed Regression Analysis

Dimension:	Average linear r-squared of group (3dp):	Average percentage of variance explained (1dp):
Frequency	0.332	33.2%
Focus	0.188	18.8%
Stage	0.093	9.3%
Quality	Not assessed	Not assessed
Differentiation	Not assessed	Not assessed

The results demonstrate that, on average, the deviations in schools' result were best explained by the prevalence of effective behaviours, followed by their focus and finally their timing (see Table 12.3d).

School Policies:

Similarly, four dimensions of school policies were considered.

Table 12.3e: The average variance in Progress 8 scores explained by each dimension of school policies in the 2017 Detailed Regression Analysis

Dimension:	Average linear r-squared of group (3dp):	Average percentage of variance explained (1dp):
Frequency	0.032	3.2%
Focus	0.092	9.2%
Stage	0.055	5.5%
Quality	0.095	9.5%
Differentiation	Not assessed	Not assessed

As can be seen from the table above, on average, variables which evaluated the quality of schools' policies accounted for the highest percentage of the variation in schools' performance, followed by variables that assessed their focus (specificity and purpose) and variables that reported upon the duration of policies implementation and/or the nature of any modifications. On average the frequency dimension, which at this level reports upon the quantity of policies that a school has introduced and that number of topics that they cover, accounted from the lowest proportion of schools' results.

2018 Analysis

The association between each independent variable and school performance is described in Appendix E.

Part 1: The average variance explained by each category of variable:

In 2018, effectiveness variables explained an average of 17.5% of the variation in schools' performance ratings. The strength of these relationships varied dramatically, however, with some relationships accounting for as much as 76.4% of the deviation in schools' results and other predicting less than 0.1%.

Table 12.3f: The average variance in Progress 8 scores explained by each category of variable in the 2018 Detailed Regression Analysis

Category:	Average linear r-squared of group (3dp):	Average percentage of variance explained (1dp):
School intake variables	0.305	30.5%
Instructional variables	0.155	15.5%
School policy variables	0.105	10.5%
Examination entry variables	0.362	36.2%

Examination Entry Variables:

Examination entry variables were the most effective predictors of schools' ratings (average variance explained by variables = 36.2%, range 76.1% to 2.6%). Each had positive association with school performance, meaning that higher entry rates were associated with higher progress scores. The only exception being the percentage of Year11 students included within schools' calculations.

The breakdown of these figures suggests that the average number of EBacc and open slots filled was the most influential sub-group (average variance explained = 51.8%). These factors had disparate relationships however that explained 27.5% and 76.1% of the variation respectively.

The entry rate for individual subjects areas was also important (average variance explained = 42.5%), with the percentage of students entering GCSE maths, English, language, science and humanities accounting for 69.5%, 61.5%, 37.4%, 33.3% and 10.9% of the variation in schools' results.

The average number of GCSE entries including and excluding equivalent qualifications, explained 57.4% and 21.2% respectively (average variance explained = 39.2%).

Whilst the percentage of students entering all English Baccalaureate subject areas accounted for 25.7% of deviations.

The least informative sub-group was the number and percentage of Year 11 pupils included within schools' Progress 8 calculation (average variance explained = 6.1%). These factors accounted for 9.5% and 2.6% of the variation in schools' scores.

Intake Variables:

School intake variables were almost as informative (see Table 12.3f). On average these variables explained 30.5% of the variation in schools' performance ratings. Once again, however, there were substantial differences in variables explanatory power, with some factors accounting for large portions of the variation in schools' scores and other explaining lesser percentages (range = 74.1% to 0.0%).

The best predictors were the two measures of student absence, the overall percentage of absence at the school and the percentage of persistent absentees (average variance explained = 67.4%). These factors had strong negative associations with school performance that explained 74.1% and 60.7% of the variation in schools' scores respectively.

The percentage of disadvantage students per cohort had a negative association with school performance. This accounted 49.0% variance in schools' ratings.

Whilst the three SEN variables, the overall percentage of SEN students, the percentage of students with SEN and a Statement or EHC plan, and the percentage of students with SEN but no Statement or EHC plan had strong negative relationships with performance that accounted for 39.1%, 36.9% and 25.5% of the variation in Progress 8 results respectively (average variance explained = 33.8%).

Schools with a higher percentage of female students had a tendency to receive more favourable ratings (variance explained = 6.9%).

The percentages of Year 11 and Progress 8 entrants speaking English as an additional language explained 7.5% and 4.8% of the variance in schools' scores respectively (average variance explained =

6.2%). Schools with a higher proportion of these students received less favourable ratings in both instances.

Finally, the percentage of non-mobile students had a negative relationship with performance that accounted for 0.0% (1dp) of the deviation in results.

Instructional Practices:

Instructional practices accounted for modest proportions of schools' performance ratings (mean percentage of variance explained by variables = 15.5%, range 55.3% to 0.0% (1dp)).

Table 12.3g: The relationship between instructional factors and schools' 2018 Progress 8 scores

Rank	Instructional behaviour	Variables considered	Linear association	Average R-squared
1	Teacher-modelling	1. Frequency of teacher-modelling tasks	Positive (r^2=0.216)	0.216
2	Classroom learning environment	1. Frequency of teacher-student interactions	Positive (r^2=0.217)	0.175
		2. Stages of lesson in which the teacher-student interactions consistently took place	Positive (r^2=0.105)	
		3. Stages of academic year in which teacher-student interactions consistently took place	Positive (r^2=0.057)	
		4. Frequency of student-student interactions	Positive (r^2=0.553)	
		5. Stages of lesson in which student-student interactions consistently took place	Positive (r^2=0.023)	
		6. Stages of academic year in which student-student interactions consistently took place	Positive (r^2=0.057)	
		7. Frequency of classroom disruptions	Negative (r^2=0.334)	
		8. How frequently teachers responded to classroom disruptions	Negative (r^2=0.072)	
		9. Stages of lesson in which classroom disruptions consistently took place activity takes place	Negative (r^2=0.003)	
		10. Stages of academic year in which classroom disruptions consistently took place	Negative (r^2=0.319)	
3	Questioning	1. Frequency of questioning	Positive (r^2=0.297)	0.172
		2. Frequency of open-ended questions	Positive (r^2=0.348)	
		3. Whether questioning typically referred to a series, whole or part of the lessons	Negative (r^2=0.079)	

		4. Stages of lesson in which questioning tasks consistently took place	Negative (r^2=0.035)	
		5. Stages of academic year in which questioning tasks consistently took place	Negative (r^2=0.101)	
4	Classroom assessment	1. Frequency of classroom assessment tasks	Positive (r^2=0.184)	0.113
		2. Stages of lesson in which classroom assessments consistently took place	Positive (r^2=0.022)	
		3. Stages of academic year in which classroom assessments consistently took place	Negative (r^2=0.132)	
5	Structuring	1. Frequency of structuring tasks	Positive (r^2=0.079)	0.101
		2. Whether structuring tasks typically referred to a series, whole or part of the lesson	Positive (r^2=0.208)	
		3. Stages of lesson in which structuring tasks consistently took place	Negative (r^2=0.115)	
		4. Stages of academic year in which structuring tasks consistently took place	Positive (r^2=0.002)	
6	Application	1. Frequency of application tasks	Positive (r^2=0.226)	0.083
		2. Whether application tasks typically referred to a series, whole or part of the lessons	Negative (r^2=0.025)	
		3. Stages of lesson during which application tasks consistently took place	Positive (r^2=0.023)	
		4. Stages of academic year in which application tasks consistently took place	Positive (r^2=0.057)	
7	Orientation	1. Frequency of orientation tasks	Positive (r^2=0.000)	0.033
		2. Whether orientation tasks typically referred to a series, whole or part of the lesson	Positive (r^2=0.015)	
		3. Stages of lesson in which orientation tasks consistently took place	Negative (r^2=0.115)	
		4. Stages of the academic year in which orientation tasks consistently took place	Negative (r^2=0.002)	

The teacher-modelling variable had the closest association with schools' scores (see Table 12.3g). This explained 21.6% of the variation in schools' ratings.

The school learning environment variables had the next closest association (average variance explained = 17.5%). All three aspects of this factor were influential, with student-student interactions, classroom disruptions and teacher-student interactions accounting for an average of 21.1%, 18.5% and 12.6% of the variation respectively.

Aspects of questioning, classroom assessment and structuring accounted for an average of 17.2%, 11.3% and 10.1% of the variation respectively, whist application and orientation variables accounted for an average of 8.3 and 3.3%.

School Policies:

As a group, differences in schools' policies had the least association with school performance. On average these variables accounted for 10.5% of the variation is Progress 8 ratings (range 72.3% to 0.0%).

Table 12.3h: The relationship between school policy factors and schools' 2018 Progress 8 scores

Rank	Policy area	Variables considered	Linear association	Average R-squared
1	Evaluation of the school teaching policies	1. Aspects of the school teaching policies that were evaluated (6 aspects total)	Negative (r^2=0.021)	0.224
		2. Level of feedback generated by the evaluation of school teaching policies	Negative (r^2=0.295)	
		3. Whether there was a formalised procedure for evaluating the mechanisms that were used to assess the school teaching policies	N/A	
		4. Reliability of mechanisms that were used to evaluate the school teaching policies	Positive (r^2=0.209)	
		5. Proportion of evaluation data that was used formatively; teaching evaluations.	Positive (r^2=0.723)	
		6. Face-validity of the mechanisms that were used to evaluate the school teaching policies	Positive (r^2=0.003)	
		7. Influence that the evaluations of the school teaching policies had upon students' learning	N/A	
		8. Extent to which the benefits of evaluating the school teaching policies outweighed the drawbacks	Positive (r^2=0.544)	
2	School teaching policies	1. Coverage of quantity of instruction policy areas (4 policies areas)	Positive (r^2=0.070)	0.101
		2. Coverage of learning opportunity policies (9 policy areas)	Negative (r^2=0.039)	
		3. How many of 8 teaching behaviours are covered by the school teaching policies (8 policy areas)	Negative (r^2=0.002)	
		4. Extent that the quantity of instruction policies dictated teachers' and students' actions	Negative (r^2=0.318)	
		5. Number of objectives pursued by quantity of instruction policies	Positive (r^2=0.005)	
		6. Extent that the policies on the provision of learning opportunities dictated teachers' and students' actions	Negative (r^2=0.318)	
		7. Number of objectives pursued by the policies on the provision of learning	Positive	

		opportunities	(r^2=0.005)	
		8. Extent that the policies on teachers' instructional behaviour dictated teachers' and students' actions	Negative (r^2=0.012)	
		9. Number of objectives pursued by the policies on teachers' instructional behaviours.	Negative (r^2=0.034)	
		10. Number of years that the current quantity of instruction policies had been implemented	Negative (r^2=0.269)	
		11. Average number of years between modifications of the quantity of instruction policies	Negative (r^2=0.095)	
		12. Whether modifications to the quantity of instruction policies were based upon data from systematic evaluations	Positive (r^2=0.001)	
		13. Number of years that the current learning opportunity policies had been implemented	Negative (r^2=0.357)	
		14. Average number of years between modifications of the learning opportunity policies	Negative (r^2=0.006)	
		15. Whether modifications to the learning opportunity policies were based upon data from systematic evaluations	Positive (r^2=0.034)	
		16. How long the current instructional behaviour policies had been implemented	Positive (r^2=0.090)	
		17. Average number of years between modifications of the instructional behaviours policies	Positive (r^2=0.162)	
		18. Whether modifications to the instructional behaviour policies were based upon data from systematic evaluations	N/A	
3	Evaluation of the SLE policies	1. Aspects of the SLE policies that were evaluated (6 aspects total)	Positive (r^2=0.119)	0.049
		2. Level of feedback generated by the evaluations of the SLE	Negative (r^2=0.256)	
		3. Whether there was a formalised procedure for evaluating the mechanisms that were used to assess the SLE policies	N/A	
		4. Reliability of mechanisms that were used to evaluate the SLE	Positive (r^2=0.001)	
		5. Proportion of evaluation data that was used formatively; SLE evaluations.	Positive (r^2=0.000)	
		6. Face-validity of the mechanisms that used to evaluate the SLE policies	Negative (r^2=0.002)	
		7. Influence that the evaluation of the SLE polices had upon students learning	Negative (r^2=0.002)	
		8. Extent to which the benefits of evaluating the SLE policies outweighed the drawbacks	Positive (r^2=0.010)	
4	SLE policies	1. Coverage of the SLE (5 policy areas)	Negative (r^2=0.025)	0.031
		2. Extent that the SLE policies dictated teachers' and students' actions	Negative (r^2=0.001)	

		3. Number of objectives pursued by SLE policies	Negative (r^2=0.034)	
		4. How long the current SLE policies had been implemented	Positive (r^2=0.073)	
		5. Average number of years between modifications of the SLE policies	Positive (r^2=0.054)	
		6. Whether modifications to the SLE policies were based upon data from systematic evaluations	N/A	

Variables concerned with the evaluation of the school teaching policy explained the most variation (average variance explained = 22.4%, see Table 12.3h), followed by those of the school teaching policies (average variance explained = 10.1%) and variables associated with procedures for evaluating the school learning environment (average variance explained = 4.9%). The variables associated with the SLE policies accounted for the least variation (average variance explained = 3.1%).

The three sub-divisions of the school teaching policies accounted for similar proportions of variance. This time, however, the quantity of instruction policies explained an average of 13.4% of the variance in schools' scores, whilst the mean variance explained by learning opportunity variables and instructional behaviour variables was 12.7% and 5.0% respectively.

Part 2: The average variance explained by each dimension of effectiveness

In 2018, the same dimensions emerged as being the most predictive.

Instructional Practices:

Table 12.3i: The average variance in Progress 8 scores explained by each dimension of instructional practices in the 2018 Detailed Regression Analysis

Dimension:	Average linear r-squared of group (3dp):	Average percentage of variance explained (1dp):
Frequency	0.231	23.1%
Focus	0.082	8.2%
Stage	0.073	7.3%
Quality	0.351	35.1%
Differentiation	Not assessed	Not assessed

The frequency of instructional behaviours had a close relationship with the school performance, whilst the focus and timing of instructional practices had moderate associations. The most predictive dimension, however, was the quality of instructional behaviours. On average these variables accounted for 35.1% of deviation in schools' results (see Table 12.3i).

In this analysis, however, an additional classification of variable was assessed, the consistency of instructional practices across the school. These four variables had a relatively strong association with school performance (average variance explained = 35.1%), though as discussed in the next section, the relationship between consistency and performance was not consistent with the pre-specified expectations.

School Policies:

Table 12.3j: The average variance in Progress 8 scores explained by each dimension of school policies in the 2018 Detailed Regression Analysis

Dimension:	Average linear r-squared of group (3dp):	Average percentage of variance explained (1dp):
Frequency	0.034	3.4%
Focus	0.118	11.8%
Stage	0.082	8.2%
Quality	0.149	14.9%
Differentiation	Not assessed	Not assessed

With regards to school policies, the quality of policies had the closest association with schools' performance ratings, followed by their focus (specificity and purpose) and the timing of their implementation. The least predictive dimension was the frequency dimension which described the quantity of policies that the school had in place and the areas that they covered (see Table 12.3j).

Change Analysis

The association between each independent variable and school performance is described in Appendix E.

Part 1: The average variance attributed to each category of variable:

Changes in the status of key effectiveness variables were able to predict an average of 17.9% of the variation in schools' performance over time.

Table 12.3k: The average variance in Progress 8 scores explained by each classification of variable in the Detailed Regression Change Analysis

Category:	Average linear r-squared of group (3dp):	Average percentage of variance explained (1dp):
School intake variables	0.179	17.9%
Instructional variables	0.167	16.7%
School policy variables	0.183	18.3%
Examination entry variables	0.224	22.4%

Examination Entry Variables:

Overall, examination entry variables were the most effective predictors of school performance (see Table 12.3k). Changes in these indicators explained an average of 22.4% of the change in schools' Progress 8 ratings (range = 46.6% to 2.1%).

Changes in the entry rates for EBacc and open slots were the most influential sub-category of variable. Both factors had positive correlations with performance that accounted for an average of 36.2% of the variation in Progress 8 changes. Deviations in the coverage of the EBacc slots, however, explained a higher proportion of the variation (46.5%) than changes in the average number of Open slots filled (25.9%).

Changes in the percentage of Year 11 pupils with entries in all English Baccalaureate qualification areas explained a similar level of variation (36.0%), with increases in entry rates tending to occur more frequently within improving schools.

The third most influential sub-group was the number and percentage of Year 11 students to be included within schools' Progress 8 calculations (average variance explained = 32.4%). Increases in both variables had positive associations with school performance. Changes in number of entrants were however the more influential factor (43.7% variance explained by former, 21.0% by the latter).

Changes in the entry rates for individual subject areas had positive relationships with school performance that were able to explain substantial portions of the differences in schools' ratings (average variance explained = 21.4%). Specifically, the entry rate for EBacc language accounted for 40.0% of the variation, whilst the entry rates for EBacc humanities, maths, science and English accounted for 20.6%, 20.3%, 17.1% and 8.9% respectively.

Changes in the average number of GCSE qualifications had a modest positive association with the changes in schools' performance ratings. With changes in the number of GCSE and equivalent qualifications accounting for 6.2% of the variation in schools' results and the number of GCSE excluding equivalencies 2.1%. The average predictive power of this sub-group was therefore 4.2%.

Finally, increases in the number of Level 3 entries (AS levels) were associated with higher Progress 8 scores. The predictive capacity of this factor was, however, very limited and accounted for only 3.4% of the results.

Intake Variables:

Changes in school intake had a moderate association with school performance (average variation explained by changes in the variables = 17.9%; range 61.6% to 0.1%).

In this analysis, changes in percentage of female Progress 8 entrants from each school explained the highest proportion of the changes in schools' performance (61.6%). More specifically, cohorts that had a higher percentage of girls in their 2018 cohort tended to improve more than cohorts with a higher proportion of boys.

Student absence rates were also influential (average variance explained = 24.7%), with the change in the percentage of persistent absentees accounting for 26.9% of the differences in schools' results and change in the overall rate of non-attendance 22.4%. Both factors had a negative association with the change in schools' ratings meaning that institutions tended to receive less favourable appraisal if the average rate of absence increased.

The three special educational needs variables were informative (average variance explained = 16.5%), though their effect varied substantially, with the change in the overall percentage of SEN accounting for 33.1% of the cross-year change, changes in the percentage of SEN pupils without a Statement or EHC plan explaining 15.7% and the change in the percentage of students with an SEN Statement or EHC plan 0.8%. The direction of association however was negative in all cases.

Changes in the rate of student mobility accounted for 15.7% of the variance in schools' ratings, with increases in the percentage of non-mobile students being more common amongst improving schools.

The remaining factors had less predictive capacity.

Changes in the percentage of Year 11 pupils that spoke English as an additional language, for example, had a positive association with school performance that accounted for 1.6% of the variation in schools' ratings, whilst the percentage of Progress 8 students that spoke English as an additional language had a negative association that explained 0.7%. This made the average effect size for the EAL variables 1.2%.

Changes in the percentage of disadvantage students, however, were the least predictive intake factor. In fact, the slight negative association between this variable and performance accounted for only 0.1% of the variation in schools' scores.

Instructional Practices:

Changes in schools' instructional practices had the least association with school performance. On average, changes in these variables were only able to accounted 16.7% of the variance in schools' scores (range = 67.8% to 0.0%).

Table 12.3L: The relationship between the changes in schools' instructional practices and the 2017-2018 change in schools' Progress 8 scores

Rank	Instructional behaviour	Variables considered	Linear association	Average r-squared score
1	Classroom assessment	1. Change in the frequency of classroom assessment tasks	Positive (r^2=0.323)	0.299
		2. Change in the range of assessment methods used by teachers during classroom assessments	Positive (r^2=0.224)	
		3. Change in the number of objectives behind classroom assessment task	Positive (r^2=0.444)	
		4. Change in the stages of lessons in which classroom assessments took place	Positive (r^2=0.472)	
		5. Change in the stages of academic year that classroom assessments took place	Negative (r^2=0.421)	
		6. Change in the speed with which classroom assessments are analysed,	Positive	

		reported and acted upon.	$(r^2=0.584)$	
		7. Change in the face-validity of classroom assessments	Positive $(r^2=0.238)$	
		8. Change in the amount of constructive feedback that was given to students during/after classroom assessments	Positive $(r^2=0.047)$	
		9. Change in the influence of classroom assessments on students' learning	Positive $(r^2=0.238)$	
		10. Change in teachers' ability to adapt classroom assessments and feedback around students' individual needs	Positive $(r^2=0.000)$	
2	Teacher-modelling	1. Change in the frequency of teacher-modelling tasks	Positive $(r^2=0.112)$	0.231
		2. Change in the number of circumstances that problem-solving strategies could be applied to	Positive $(r^2=0.076)$	
		3. Change in the number of times that teachers introduced more than one strategy for solving a problem	Positive $(r^2=0.098)$	
		4. Change in the proportion of teacher-modelling tasks which introduced problem-solving strategy after the problem	Positive $(r^2=0.002)$	
		5. Change in the clarity with which problem-solving strategies were introduced	Positive $(r^2=0.548)$	
		6. Change in teachers' ability to adapt teacher-modelling tasks to meet students' individual needs	Positive $(r^2=0.548)$	
3	Structuring	1. Change in the frequency of structuring tasks	Negative $(r^2=0.135)$	0.200
		2. Change in whether structuring tasks typically referred to a series, the whole, or part of the lesson	N/A	
		3. Change in the number of objectives behind structuring tasks	Positive $(r^2=0.155)$	
		4. Change in the stages of lessons in which structuring tasks took place	Negative $(r^2=0.259)$	
		5. Change in the stages of academic year that structuring tasks took place	N/A	
		6. Change in the clarity of structuring tasks	Positive $(r^2=0.147)$	
		7. Change in the influence that structuring tasks had on students' learning	Positive $(r^2=0.548)$	
		8. Change in the extent to which lessons and schemes of work were structured so that the easier tasks preceded the difficult ones	Positive $(r^2=0.548)$	
		9. Change in teachers' ability to adapt structuring tasks to meet students' individual needs	Positive $(r^2=0.004)$	
4	Application	1. Change in the frequency of application tasks	Positive $(r^2=0.468)$	0.164
		2. Change in whether applications tasks referred to a series, the whole or part of the lesson	Positive $(r^2=0.017)$	

		3. Change in the number of objectives behind application tasks	Positive (r^2=0.155)	
		4. Change in the stages of lessons in which application took place	N/A	
		5. Change in the stages of academic year that application tasks took place	N/A	
		6. Change in the extent to which application tasks expanded upon the material that was taught in the lessons	Positive (r^2=0.238)	
		7. Change in teachers' ability to adapt application tasks to meet students' individual needs	Positive (r^2=0.272)	
5	Classroom learning environment	1. Change in the frequency of teacher-student interactions	Positive (r^2=0.333)	0.156
		2. Change in the proportion of teacher-student interactions that were task-related	Negative (r^2=0.004)	
		3. Change in the stages of the lesson in which teacher-student interactions took place	Positive (r^2=0.015)	
		4. Change in the stages of academic year that teacher-student interactions took place	N/A	
		5. Change in the extent to which teachers' interventions were able to establish the desired form of interaction (on-task behaviour)	Positive (r^2=0.548)	
		6. Change in teachers' ability to adapt their strategies for establishing on-task behaviour to individual students' needs	Positive (r^2=0.200)	
		7. Change in the frequency of student-student interactions	Positive (r^2=0.076)	
		8. Change in the proportion of student-student interactions that were task-related	Negative (r^2=0.004)	
		9. Change in the stages of lessons in which student-student interactions took place	Positive (r^2=0.073)	
		10. Change in the stages of academic year that student-student interactions consistently took place	N/A	
		11. Change in the extent to which teachers' interventions were able to establish the desired form of interaction (on-task behaviour)	Positive (r^2=0.548)	
		12. Change in teachers' ability to adapt their strategies for establishing on-task behaviour to individual students' needs	Positive (r^2=0.200)	
		13. Change in the frequency of classroom disruptions	Negative (r^2=0.099)	
		14. Change in the frequently with which teachers responded to classroom disruptions	Positive (r^2=0.678)	
		15. Change in the proportion of disruptions that were due to previously unresolved issues	Positive (r^2=0.022)	
		16. Change in the extent to which teachers attempted to address the issue behind disruptions	Positive (r^2=0.067)	
		17. Change in the stages of lessons in which classroom disruptions took place	Positive (r^2=0.000)	

		18. Change in the stages of academic year that classroom disruptions took place	N/A	
		19. Change in the extent to which teachers' interventions solved the issues underlying classroom disruptions	Positive (r^2=0.253)	
		20. Change teachers' ability to adapt their strategies for dealing with classroom disruptions to individual students' needs	Positive (r^2=0.001)	
6	Questioning	1. Change in the frequency of questioning	Negative (r^2=0.139)	0.135
		2. Change in the frequency of open-ended questions	Negative (r^2=0.001)	
		3. Change in whether questioning tasks referred to a series, the whole or part of the lesson	Positive (r^2=0.017)	
		4. Change in the number of objectives behind questioning tasks	Negative (r^2=0.021)	
		5. Change in the stages of lessons in which questioning tasks took place	Negative (r^2=0.576)	
		6. Change in the stages of academic year that questioning tasks took place	Negative (r^2=0.531)	
		7. Change in the clarity of questioning	Positive (r^2=0.000)	
		8. Change in the appropriateness of question difficulty	Positive (r^2=0.044)	
		9. Change in the extent to which teachers sustained their interaction with the original respondent during questioning by rephrasing queries and giving clues	Positive (r^2=0.007)	
		10. Change in teachers' ability to adapt questioning tasks to meet students' individual needs	Positive (r^2=0.018)	
7	Management of time	1. Change in the proportion of lesson time that was used for teaching	Positive (r^2=0.019)	0.117
		2. Change in teachers' ability to adapt the allocation of lesson time around students' individual needs	Positive (r^2=0.030)	
		3. Change in the frequency of classroom disruptions	Negative (r^2=0.099)	
		4. Change in the frequently with which teachers responded to classroom disruptions	Positive (r^2=0.678)	
		5. Change in the proportion of disruptions that were due to previously unresolved issues	Positive (r^2=0.022)	
		6. Change in the extent to which teachers attempted to address the issue behind disruptions	Positive (r^2=0.067	
		7. Change in the stages of lessons in which classroom disruptions took place	Positive (r^2=0.000)	
		8. Change in the stages of academic year that classroom disruptions took place	N/A	
		9. Change in the extent to which teachers' interventions solved the issues underlying classroom disruptions	Positive (r^2=0.253)	

		10. Change teachers' ability to adapt their strategies for dealing with classroom disruptions to individual students' needs	Positive (r^2=0.001)	
8	Orientation	1. Change in the frequency of orientation tasks	Positive (r^2=0.015)	0.070
		2. Change in whether orientation tasks typically referred to a series, the whole, or part of the lesson	Negative (r^2=0.091)	
		3. Change in the number of objectives behind orientation tasks	Positive (r^2=0.076)	
		4. Change in the stages of lessons in which orientation tasks took place	Negative (r^2=0.259)	
		5. Change in the stages of academic year that orientation tasks took place	Negative (r^2=0.038)	
		6. Change in the extent to which teachers' orientation tasks consistently took on board students' perspective	Positive (r^2=0.002)	
		7. Change in the clarity of orientation tasks	Positive (r^2=0.004)	
		8. Change in the influence that orientation tasks had on students' learning	Positive (r^2=0.147)	
		9. Change in teachers' ability to adapt orientation tasks to meet students' individual needs	Positive (r^2=0.002)	

Changes in the classroom assessment variables were the most predictive, explaining 29.9% of the variation in school performance ratings, followed by the teacher-modelling (average variance = 23.1%), structuring (20.0%) and applications variables (16.4%). The classroom learning environment variables were also important considerations (average variance explained = 15.6%), with teacher-student, student-student and classroom disruptions exhibiting average effect sizes of 18.3%, 15.0% and 14.0% respectively. Management of time variables had a modest association with the change in schools' performance ratings (average variance explained = 11.7%). The orientation variables, however, had the lowest mean effect (7.0%).

Changes in teachers' coverage of the school curriculum were also influential and accounted for 15.4% of the variation in schools' results.

The directional effect of all variables is specified within Table 12.3L.

School Policies:

School policies were the second most influential category of variables (average percentage of variance accounted for by variables = 18.3%, range 64.4% to 0.0%).

Table 12.3m: The relationship between the changes in school policy factors and the 2017-2018 change in schools' Progress 8 scores

Rank	Factor	Variables included within factor	Linear association	Linear r-squared
1	School teaching policies	1. Change in the coverage of the schools' policies on quantity of instruction (4 policy areas)	Negative (r^2=0.105)	0.225
		2. Change in the coverage of the schools' learning opportunity policies (9 areas assessed)	Negative (r^2=0.055)	
		3. Change in the coverage of the schools' teaching behaviour policies (8 areas assessed)	Positive (r^2=0.019)	
		4. Change in the extent to which the quantity of instruction policies dictated teachers' and students' actions	Negative (r^2=0.004)	
		5. Change in the number of objectives that were pursued by the quantity of instruction policies.	Positive (r^2=0.332)	
		6. Change in the extent to which the policies on the provision of learning opportunities dictated teachers' and students' actions	Positive (r^2=0.127)	
		7. Change in the number of objectives pursued by the policies for the provision of learning opportunities	Positive (r^2=0.322)	
		8. Change in the extent to which the policies on teachers' instructional behaviour dictated teachers' and students' actions.	Negative (r^2=0.000)	
		9. Change in the number of objectives pursued by the policies on teachers' instructional behaviour	Positive (r^2=0.553)	
		10. Number of years that the quantity of instruction policy has been implemented	Negative (r^2=0.606)	
		11. The average number of years between modifications of the quantity of instruction policies	Negative (r^2=0.033)	
		12. Change in the whether the modifications to the quantity of instruction policies were based upon evaluation data (formative use of evaluation data)	Positive (r^2=0.254)	
		13. Number of years that the policies for providing learning opportunities have been implemented	Negative (r^2=0.579)	
		14. The average number of years between modifications of the policies for providing learning opportunities	Positive (r^2=0.003)	
		15. Change in whether the modifications to the policies for providing learning opportunities were based upon evaluation data (formative use of evaluation data)	Positive (r^2=0.219)	
		16. Number of years that the policies of teachers' instructional behaviours have been implemented	Negative (r^2=0.062)	
		17. The average number of years between modifications of the policies on teachers' instructional behaviours	Positive (r^2=0.142)	
		18. Change in whether the modifications to the policies on teachers' instructional behaviours were based upon evaluation data (formative use of evaluation data)	Positive (r^2=0.254)	
		19. Change in the clarity of quantity of instruction policies	Positive	

			(r^2=0.337)	
		20. Change in the alignment between the quantity of instruction policies and the academic literature	Positive (r^2=0.635	
		21. Change in the level of support provided to teachers and/or students to implement the quantity of instruction policies	Positive (r^2=0.443)	
		22. Change in the level of influence that the quantity of instruction policies had on teachers' and students' behaviour	Positive (r^2=0.644)	
		23. Change in the clarity of the policies for providing learning opportunities	Positive (r^2=0.215)	
		24. Change in the alignment between the policies on the provision of learning opportunities and the academic literature	Positive (r^2=0.415)	
		25. Change in the level of support provided to teachers and/or students to implement the policies on the provision of learning opportunities	Positive (r^2=0.312)	
		26. Change in the level of influence that the policies on the provision of learning opportunities had on teachers' and students' behaviour	Positive (r^2=0.317)	
		27. Change in the clarity of the policies on teachers' instructional behaviours	Negative (r^2=0.174)	
		28. Change in the alignment between the policies on teachers' instructional behaviours and the academic literature	Positive (r^2=0.019)	
		29. Change in the level of support provided to teachers and/or students to implement the policies on teachers' instructional behaviours	Positive (r^2=0.015)	
		30. Change in the level of influence that the policies on the policies on teachers' instructional behaviours had on teachers' and students' behaviour	Positive (r^2=0.044)	
		31. Change in the amount of instruction time that was provided to students by the school policies	Positive (r^2=0.050)	
		32. Change in the alignment between the school curriculum and the content assessed at KS4	Positive (r^2=0.023)	
		33. Change in the level of differentiation in the quantity of instruction policies	Positive (r^2=0.183)	
		34. Change in the level of differentiation in the policies that govern the provision of students' learning opportunities	Positive (r^2=0.112)	
		35. Change in the extent to which teachers were encouraged to differentiate the learning opportunities that they offer to students	Positive (r^2=0.286)	
		36. Change in the level of differentiation in the policies governing teachers' instructional behaviours	Positive (r^2=0.503)	
		37. Change in the extent to which teachers were encouraged to differentiate their use of the 8 instructional behaviours	Positive (r^2=0.185)	
		38. Change in the instruction time dedicated to Mathematics	Positive (r^2=0.139)	
		39. Change in the instruction time dedicated to English Language and English Literature	Positive (r^2=0.139)	
		40. Change in the instruction time dedicated to other EBacc subjects	Positive	

			(r^2=0.192)	
		41. Change in the instruction time dedicated to Non-EBacc GCSEs and Non-GCSEs	Positive (r^2=0.254)	
		42. Change in the instruction time dedicated to Level 3 qualifications	Positive (r^2=0.137)	
2	SLE policies	1. Change in the coverage of the SLE polices (5 areas assessed)	Positive (r^2=0.002)	0.210
		2. Change in the extent to which the SLE policies dictated teachers' and students' actions.	Positive (r^2=0.012)	
		3. Change in the number of objectives pursued by the SLE policies.	Positive (r^2=0.465)	
		4. Number of years that the SLE policies has been implemented (Duplicate of 2018 variable)	Negative (r^2=0.000)	
		5. The average number of years between modifications of the SLE policies (Duplicate of 2018 variable)	Positive (r^2=0.247)	
		6. Change in whether the modifications to the SLE policies were based upon evaluation data (formative use of evaluation data)	Positive (r^2=0.254)	
		7. Change in the clarity of the SLE policies	Positive (r^2=0.152)	
		8. Change in the alignment between the SLE policies and the academic literature	Positive (r^2=0.367)	
		9. Change in the level of support provided to teachers and/or students to implement the SLE policies	Positive (r^2=0.017)	
		10. Change in the level of influence that the SLE policies had on teachers' and students' behaviour	Positive (r^2=0.291)	
		11. Change in the level of differentiation in the SLE policies	Positive (r^2=0.503)	
3	Evaluation of the school teaching policies	1. Change in frequency with which the school collects data on the school teaching policies	Positive (r^2=0.166)	0.136
		2. Change in number of sources of information that the evaluations of the school teaching policies drew upon	Negative (r^2=0.000)	
		3. Change in the number of aspects of the school teaching policies that were evaluated. (6 policy areas)	Negative (r^2=0.001)	
		4. Change in the level of feedback generated by the evaluation of the school teaching policies	Positive (r^2=0.296)	
		5. Change in the frequency with which the school evaluated the school teaching policies	Positive (r^2=0.172)	
		6. Change in whether there was a formalised procedure for evaluating the mechanisms that were used to assess the school teaching policies	Positive (r^2=0.089)	
		7. Change in the reliability of the mechanisms/processes that evaluated the school teaching policies	Positive (r^2=0.284)	
		8. Change in whether evaluation data was used to inform decisions about school teaching practice	Positive (r^2=0.000)	

		9. Change in the extent to which the evaluations of the school teaching policies assessed the factors that they were intended to assess (face validity)	Positive (r^2=0.361)	
		10. Change in the strength of the relationship between the evaluations of the school teaching policies and students' learning	Positive (r^2=0.027)	
		11. Change in the extent to which the benefits of monitoring the school teaching policy outweighed the drawbacks	Positive (r^2=0.014)	
		12. Change in the emphasis that was placed on evaluating the under-performing aspects the schools' instructional provisions	Positive (r^2=0.221)	
4	Evaluation of the SLE policies	1. Change in frequency with which the school collects data on the SLE	Positive (r^2=0.004)	0.059
		2. Change in number of sources of information that the evaluations of the SLE policies drew upon	Positive (r^2=0.004)	
		3. Change in the number of aspects of the SLE policies that were evaluated. (6 areas total)	Positive (r^2=0.007)	
		4. Change in the level of feedback generated by the evaluation of the SLE policies	Positive (r^2=0.126)	
		5. Change in the frequency with which the school evaluates the SLE	Positive (r^2=0.172)	
		6. Change in whether there was a formalised procedure for evaluating the mechanisms that were used to assess the school's policies for creating an effective learning environment	Positive (r^2=0.089)	
		7. Change in the reliability of the mechanisms/processes that evaluate the SLE	Negative (r^2=0.032)	
		8. Change in whether evaluation data that was used to inform decisions about the SLE	Negative (r^2=0.024)	
		9. Change in the strength of the relationship between the evaluations of SLE and students' learning	N/A	
		10. Change in the extent to which the benefits of monitoring the SLE outweighed the drawbacks	N/A	
		11. Change in the extent to which evaluations of the SLE assessed the factors they were intended to assess	Negative (r^2=0.024)	
		12. Change in the emphasis that was placed on evaluating the underperforming aspects of the SLE	Positive (r^2=0.221)	

Changes in the school teaching policies had the closest association with school performance (average r-squared = 22.5%), with changes in the quantity of instruction, learning opportunity and instructional behaviour policies having average effect sizes of 26.4%[36], 23.0% and 16.4% respectively. The school learning environment policies explained a comparable but lesser proportion of variance (average = 21.0%), whilst variables associated with the evaluation of school teaching policies and the evaluation of the school learning environment accounted for an averages of 13.6% and 5.9%.

[36] It should be noted that this average includes the effect of 5 variables that are not explicitly outlined in the dynamic model, variables 38-42 in the table above. These were added to acknowledge that Progress 8 evaluates students' learning across multiple subject areas. These variables did not influence the ranks order of the groups' effect. Average effect size of group with variables excluded = 0.302.

The direction of the individual relationships varied and is reported within Table 12.3m.

Part 2: The average variance explained by each dimension of effectiveness

There were also consistencies in the dimensions that accounted for the highest proportions of variation.

Instructional Practices:

Table 12.3n: The average variance in Progress 8 scores explained by each dimension of instructional practices during the Detailed Regression Analysis of 2017-18 changes

Dimension:	Average linear r-squared of group (3dp):	Average percentage of variance explained (1dp):
Frequency	0.200	20.0%
Focus	0.092	9.2%
Stage	0.170	17.0%
Quality	0.237	23.7%
Differentiation	0.119	11.9%

On average changes in the quality of instructional practices explained the highest proportions of variance, followed by changes in the frequency of effective behaviours. Changes in the timing of effective behaviours and the level of differentiation accounted for more modest proportions, whilst changes in the focus of instructional variables had the lowest mean effect (see Table 12.3n).

This analysis, however, included two additional classifications of variables, the consistency of instructional practices across the school and teachers' coverage of the school curriculum. These variables had a relatively strong association with school performance (average variance explained by the four consistency variables = 20.0%, variance explained by teachers' coverage of the school curriculum = 15.1%), though as discussed in the next section, the relationship between consistency and performance was not consistent with expectations.

School Policies:

Table 12.3o: The average variance in Progress 8 scores explained by each dimension of school polices during the Detailed Regression Analysis of 2017-18 changes

Dimension:	Average linear r-squared of group (3dp):	Average percentage of variance explained (1dp):
Frequency	0.044	4.4%
Focus	0.187	18.7%
Stage	0.148	14.8%
Quality	0.187	18.7%
Differentiation	0.221	22.1%

Likewise, variables which reported changes in the quality and focus of school policies predicted substantial proportions of the variation in schools' performance, whilst the stage dimension variables accounted for a lesser but comparable amount. The mean effect attributed to the frequency variables was much lower than the other classifications of variables.

The differentiation variables, which were only evaluated within the change analysis, were however able to account for the more variation than any other classification of variable (see Table 12.3o).

12.4 Discussion

This chapter set out to duplicate the findings of the Shallow Regression Analysis, whilst expanding upon the assessment of instructional practices and school policy variables. In particular, the intent was to establish whether school effectiveness correlates were able to account for meaningful proportions of the variation in schools' Progress 8 results and the relative influence of factors that were inside and outside of schools' control.

Could the variation in schools' performance ratings be predicted by established effectiveness factors?

The results confirm that key effectiveness factors from academic research were able to account for meaningful proportions of the variation in schools' performance ratings, both in terms of the differences in schools' scores at specified moments in time (range of variance explained by variables in the 2017 models = 76.4% to 0.0 (1dp), range of variance explained by variables in the 2018 models = 76.1% to 0.0) and the deviation in schools' results over time (range of variance explained by variables in the 2017-18 change analyses = 67.84% to 0.0). What is more, the directional effect of these variables was consistent with their theoretical impact, specifically, 63.4% of interactions from the 2017 models, 65.2% of the assessable directional effects from the 2018 models and 76.6% of the assessable directional effects from the 2017-18 change models were in line with expectations[37]). The effect attributed to individual variables was also moderately consistent across the analyses (correlation between the r-squared scores associated with each variable in the 2017 and 2018 analyses = 0.475), which suggests that Progress 8 results are influenced by the kinds of factors that school effectiveness is normally attributed to.

The results make it clear, however, that the variables that account for the largest percentage of the deviation in schools' annual performance ratings are not necessarily the best predictors of the change in schools' scores over time. This is apparent from the weak to non-existent association between the r-squared scores of variables in the 2017 and change analyse (r=0.232) and those from the 2018 and change analyses (r= 0.026).

[37] Variables that exhibited no variation were excluded from the 2018 and Change Analysis figures as it was impossible to state whether their directional effect was in line with expectations.

Does Progress 8 provide a fair method of evaluating schools' contribution?

Evidence relating to the second research objective was more concerning.

Within the two annual analyses, examination entry variables were able to explain the highest proportions of the variation in schools' value-added results, followed by intake factors, instructional variables and then deviations in school polices.

In the change analysis, a comparable rank order of the average effect sizes was evident. The main distinction being that the effects attributed to the school policy variables increased markedly in relation to the percentage of variance explained by other categories of variable (see Table 12.4a).

Table 12.4a: The average variance in Progress 8 scores explained by each category of variable in the 2017, 2018 and 2017-2018 Detailed Regression Analyses

2017 models:	2018 models:	Models of 2017-18 change:
Intake: 21.7% (16.4%*)	Intake: 30.5% (21.2%*)	Intake: 17.9% (16.2%*)
Instruct: 19.0%	Instruct: 15.5%	Instruct: 16.7%
Policies: 7.4%	Policies: 10.5%	Policies: 18.3%
Entries: 36.7%	Entries: 36.2%	Entries: 22.4%

* These figures specify the average variance that could be explained by intake variables when the influence of the two attendance variables are excluded from the calculation.

Whilst there are reasons to suspect that these figures may give an exaggerated impression of the bias within Progress 8 assessments (see later discussions), the results are troubling as they imply that the predominant determinants of schools' value-added results are outside of schools' control.

One should be aware, however, that the effect attributed to specific variables varied substantially across the analyses, both in terms of their direction of their influence and the percentage of variance that they explain. Whilst a portion of this deviation may be genuine, the utilisation of such a small sample and the multi-collinearity between independent variables will have contributed to this instability. It is therefore necessary to interpret the results collectively to rule out this influence of chance associations. It is also significant, that these anomalies occurred more frequently in groups that had a weaker connection with schools' performance ratings. This is because subtle relationships are more likely to be overwhelmed by extraneous influences and does not therefore provide a basis for challenging the validity of Progress 8.

A more in-depth discussion of the interactions that occurred within each category of variable is now provided, followed by a critique of the research methodology and how particular shortfalls might have impacted upon the results.

Examination Entry Variables:

Table 12.4b: The average variance in Progress 8 scores explained by the examination entry sub-groups in the 2017, 2018 and 2017-2018 Detailed Regression Analyses

Sub-group of variables	2017 Analysis	Rank	2018 Analysis	Rank	2017-18 Average	Rank	2017-18 Change	Rank
Coverage of the EBacc and Open Slots	38.8%	3	51.8%	1	45.3%	1	36.2%	1
The percentage of students entering all English Baccalaureate subject areas	62.8%	1	25.7%	4	44.3%	2	36.0%	2
Entry rate for individual GCSE subject areas	31.3%	5	42.5%	2	36.9%	3	21.4%	4
Average number of GCSEs entered	39.1%	2	39.2%	3	39.2%	4	4.2%	5
Number and percentage of Year 11 students included in schools' calculations	32.9%	4	6.1%	5	19.5%	5	32.4%	3

* Average scores refer to the mean two annual scores reported in table.

Overall, examination entry variables had the closest association with school performance.

The most informative sub-groups appear to have contained variables that had a broad focus and related directly to the Attainment 8 entry criteria, that is, the average number of EBacc and open slots filled by the students from each school and the percentage of students entering all English Baccalaureate subject areas (see Table 12.4b). The next closest associations related to variables that satisfied one these criterions, i.e. the entry rate for specific GCSE subject areas and average number of GCSE entered by students. Whilst variables with indirect relationships, such as the number and percentage of Year 11 students included within schools' Progress 8 calculation or the entry rate for Level 3 qualifications (only evaluated in change analysis) accounted for the smallest percentage of variation in schools' results. This is logical and in-line with the pre-specified expectations.

The vast majority of relationships also exhibited the anticipated directional effects (91.9% of the variables from the three analyses; 11/12 examination entry variables from the 2017 models, 12/12 from the 2018 models and 11/13 variables from the change analysis). Which suggests that the more Attainment 8 slots that school cohorts filled the higher their schools' score was likely to be. The major exception to both statements being the number and percentage of Year 11 pupils entered into schools' calculations. These variables were expected to have a weak negative association with school performance, yet in three out of six observations had moderate-strong correlations with schools' Progress 8 results (see results of 2017 and change analyses). What is more, the implied effect was positive in some models and negative in others. Due to the inconsistency in of these associations (50/50 split in directional effect and whether variables appear to have had a small or larger effect), it is presumed that these were relatively inconsequential variables that correlated with school performance by chance during some of the analyses.

Taken at face value the results therefore imply that schools' examination entry practices played a key role in the determination of schools' Progress 8 results that was consistent with the their logical impact. Such an outcome has mixed implications for Progress 8. Whilst the explicability of these associations supports the conclusion that schools' value-added results are indicative of changes within the institutions, it is argued here that schools' curricular decisions should not be one of the main determinants of their ratings. In fact, regardless of whether learning in particular areas is more prised, it is asserted that if the measure is intended to report upon the quality of schools' provisions then these differences should have at most a moderate effect upon schools' value-added ratings (see Section 11.4 for a full discussion of assertion). An uncritical interpretation of these results therefore suggests that these kinds of factors had too great an influence and that Progress 8 predominately reported whether schools' were adhering to the desired curriculum.

As stated above, however, there are reasons to suspect that this form of correlational analyses may give an exaggerated impression of the influence that these factors exert. The predominant concern is that a reciprocal relationship may have existed between school effectiveness and students' examination ratings. That is to say, that in additional to schools' examination entries having implications for their progress score, it might also be the case that differences in school effectiveness (i.e. students' progress) impact upon students' examination entries, even after differences in prior-attainment have been taken into account. If this form of reciprocal relationship existed the analysis would have ascribed any association between the two factors to the examination entry variables and thus overstate their effect. The inability to distinguish cause from effect is a recognised weakness of correlational analyses that could only be unravelled by additional research. This subsequent uncertainty must therefore be acknowledged when interpreting the results. It is also important to recall that this analysis did not account for the influence of extraneous variables. If the reader is interested in the causal influence of the variables, it is therefore important to consult Chapter 13 where the evidence from all empirical sections is collated.

Intake Variables:

Intake variables were the second most informative group, though their predictive capacity was surpassed by the school policy variables within the change analysis.

Table 12.4c: The average variance in Progress 8 scores explained by school intake factors in the 2017, 2018 and 2017-2018 Detailed Regression Analyses

Sub-group of variables	2017 Analysis	Rank	2018 Analysis	Rank	2017-18 Average	Rank	2017-2018 Change	Rank
Absence	43.1%	1	67.4%	1	55.3%	1	24.7%	2
SEN	27.1%	2	33.8%	3	30.5%	2	16.5%	3
EAL	24.1%	3	6.9%	4	15.5%	4	1.2%	5
Gender	1.5%	4	6.2%	5	3.9%	5	61.6%	1
Disadvantage	0.0%	5	49.0%	2	24.5%	3	0.1%	6
Non-mobility	0.0%	6	0.0%	6	0.0%	6	15.7%	4

* Average scores refer to the mean two annual scores reported in table.

Two sub-groups in particular had large and relatively stable effects that persisted across the three analyses, the percentage of absence variables (overall absence rate and percentage of persistent absentees) and the percentage of students with special educational needs (overall, with and without Statements or EHC plans) (see Table 12.4c). These factors are presumed to have played influential roles in determining schools' ratings.

The remaining sub-groups, which report upon the prevalence of female students, disadvantaged students, non-mobile students and the percentage of students that spoke English as an additional language, displayed lower levels of association that were less consistent across the analyses. Despite these variables accounting for high proportions of variation during some of the analyses (see for example the percentage of variance explained by gender in the 2017-18 change analyses, or the percentage of the 2018 Progress 8 scores explained by the percentage of disadvantaged students at each school) it is therefore concluded that within the sample these variables had lesser but not inconsequential impact upon schools' performance ratings. It would appear therefore that within the sample of 9 northern schools, Progress 8's method of controlling for the differences in students' prior attainment removed much but not all of the bias that these factors introduced.

The majority of associations were consistent with our pre-established expectations (66.3%); 3/10 relationships in the 2017 models, 7/10 from the 2018 models and 9/10 in the change analysis. The ability to predict the direction of variables effects therefore provides further support for concluding that schools' effectiveness ratings do reflect genuine differences within schools and whilst the rate of unexpected associations was noticeably higher than amongst the examination entry variables, all of these have credible explanations.

The most common anomaly was for the differences in schools' Progress 8 results to have a negative association with the percentage of students that spoke English as an additional language (5/6 observations unpredicted). The expectation was that these variables would have a positive relationship with school performance as the language skills of these students are expected to improve markedly during their secondary education. Within the sample, however, the percentage of EAL students per school was substantially lower than within the overall population (3.1% as opposed to 16.3%). This may conceivably have impacted upon the results or indicate that when the percentage of EAL students that typically attend a school is smaller, schools are less adept at meeting their unique learning needs. It may likewise account for these variables having a low level of influence within the analysis (see discussion above). An alternative explanation is that a shortfall in the operationalization of these variables might have been responsible. That is to say, that since the proficiency of students' language skills was not directly assessed, it is impossible to say for certain that these students started Key Stage 2 with below average literacy skills. Without this presumption, there would be less reason to suspect that these students would progress more than other students. What is more, the lower the number of EAL students per school and the fewer schools within the sample, the more likely it is that the average English speaking proficiency of EAL students would not have been significantly different from non-EAL students. There may equally have been an interaction between ethnicity, EAL status and performance that was unacknowledged by the research methodology. On several fronts, therefore, a positive relationship between these variables and the sample from detailed regression could be legitimised

Another set of unexpected associations related the assessment of the SEN students. Whilst the majority of these variables had a negative association with school performance that theoretically reflects the additional challenges that these pupils have to overcome, there were two instances of these indicators having positive associations with schools' ratings. More specifically, within the sample, schools' with more favourable ratings had higher proportions of students with special educational needs (with or without a Statement/EHC plan) and higher proportions of students with special educational needs but no Statement or EHC plan. If one is prepared to assume that students who had a Statement or EHC plan represent the most disadvantaged SEN group, then these results becomes more explicable. In fact, it may not only indicate that when adequate support is available SEN students are capable of success, but that the successful remediation of their initial disadvantage may cause these individual reach higher levels of Key Stage 4 attainment than students with the same Key Stage 2 prior-attainment scores. The mechanisms underlying the relationship between SEN status and student progression may not therefore be entirely dissimilar from those that underpin the association between English as an additional language status and performance. The persistent negative association between the percentage of statemented SEN pupils and schools' ratings would then reflect the severity of the barriers that the most disadvantaged pupils have to overcome and the difficulty of addressing them. It is important to emphasise, however, that whilst this explanation is in-line with Carrol's (1963) assertion that almost all students are capable of achieving academic success, and ties together the inconsistences within the results, this interpretation is highly speculative and requires validation from further research.

The only remaining anomalies were then; the positive associations between schools' 2017 Progress 8 scores and the percentage of disadvantaged students (variance explained = 0.0% 1dp), the relationship between school's 2017 Progress 8 scores and the percentage of the percent of female students (variance explained = 1.5%), and the negative correlations the percentage of non-mobile students and schools' Progress 8 scores from 2017 and 2018 (variance explained in 2017 = 0.0% 1dp, variance explained in 2018 = 0.0% 1dp). Given the limited effect size of these variables, the sample size and the lack of statistical controls, it seems reasonable to ascribe these differences to chance associations that may or may not reflect the causal-relationship between the variables and schools' results.

All of this suggests that the differences in schools' intake have a concerning level of influence upon the value-added ratings that schools' receive. Once again, however, it is important to consider several methodological decisions that will have impacted upon the results.

The first was the decision to treat school absence rates as an extraneous factor that was outside of schools' control. This stance was adopted because the choice as to whether to attend compulsory education is ultimately mediated by the actions of students and/or their parents. Whilst it is recognised that teachers' actions and/or school policies may impact upon the these variables, it was presumed that these factors would account for a lesser portion of the variation, and that encouragement from school practitioners may not in some instances be sufficient to overcome the initial differences. In acknowledgement of the fact that teachers and schools may exert some degree of influence over these matters, however, the average percentage of variance that was accounted for by intake variables, when these two absence variables are excluded, is also reported in Table 12.4a. The true influence of the intake bias is likely to have fallen within these two extremes. Within this analysis however the differences between the two accounts makes little difference to the substantive conclusions.

The second factor was that the two measures of student absence were operationalised as school-level variables. These indicators were therefore unique in that they reported upon the behaviour of all students within the school, rather than referring specifically to the schools' Year 11 cohort. This imprecision may have lowered the percentage of variance that absence rates could account for, especially within the change analysis where the magnitude of any inaccuracies would be larger in relation to the measurement scale.

Instructional Practices:

Differences in schools' instructional practices had moderate associations with school performance, though they still ranked as the third most influential category of variable during the annual analyses and the least influence group within the change analysis.

Table 12.4d: The average variance in Progress 8 scores explained by instructional practices during the 2017, 2018 and 2017-18 Detailed Regression Analyses

Sub-group of variables	2017 Analysis	Rank	2018 Analysis	Rank	2017-18 Average	Rank	2017-18 Change	Rank
Orientation	14.3%	3	3.3%	7	8.8%	6	7.0%	7
Structuring	12.6%	5	10.1%	5	11.4%	4	20.0%	3
Questioning	18.1%	2	17.2%	3	17.7%	3	13.5%	6
Teacher-modelling	14.0%	4	21.6%	1	17.8%	2	23.1%	2
Application	10.2%	6	8.3%	6	9.3%	5	16.4%	4
Classroom learning environment	32.1%	1	17.5%	2	24.8%	1	15.6%	5
Classroom assessment	5.1%	7	11.3%	4	8.2%	7	29.9%	1

* Average scores refer to the mean two annual scores reported in table.

Across the two annual analyses, classroom learning environment variables accounted for the highest percentages of the variance in schools' Progress 8 results[38], followed by the utilisation of teacher-modelling activities and questioning. Whilst the use of structuring tasks, application, orientation, classroom assessments accounted smaller portions of the scores. Teachers' ability to manage lesson time was only evaluated in a single set of analyses, but the available evidence suggests its impact was modest (average variance explained = 11.7%) (see Table 12.4d). The factors that were most useful in predicting within-year differences in schools' Progress 8 scores were not, however, the factors that were best at explaining the variation in schools' ratings over time.

It is notable, though, that the relationship between each instructional behaviour and Progress 8 ratings deviated considerably across the analyses. This is unsurprising given the sample size and the fact that all instructional behaviours were evaluated without the benefit of statistical controls. Nevertheless, it suggests that the reported effect sizes have been influenced by extraneous variables. One should therefore be cautious in interpreting the individual parameter estimates.

That being said, the majority of the variables correlated with schools' performance ratings in the expected manner (69.6% of the variables with assessable directional effects). Specifically, 22/31 of the relationships observed in the 2017 models, 22/35 of the relationships observed in the 2018 models and 50/69 assessable directional effects from the change analysis[39]. This helps to legitimise the figures.

[38] With the classroom disruptions accounting for more variance on average than the nature of student-student interactions, and the characteristics of students-student interactions explaining more variation than the differences in teacher-student interactions. This was evident from the 2017-2018 average effect sizes, which were 18.9%, 13.0% and 12.1% respectively.

[39] Seven of the instructional variables from the change analyses exhibited no variation across our sample. These associations are therefore excluded from the stated statistic as their directional effect could not be evaluated.

The average effect sizes of each factor are also consistent with past research. There is, for example, a r = 0.643 correlation between the effect sizes of the 2017 instructional factors and the average effect sizes reported in Creemers and Kyriakides (2008) meta-analysis of classroom-level effectiveness factors, and an r = 0.534 correlation between the effect sizes of the 2018 instructional factors and the same figures. One can therefore be assured that the reported effect sizes reflect the underlying relationships between the factors.

The results therefore imply that teaching behaviours had a systematic impact upon schools' performance ratings. The proportion of variation that these variables explained, however, suggests that examination entry differences and school intakes had a greater impact upon the scores than differences in teachers' practices. Taken at face value, this outcome is highly concerning and suggests that Progress 8 ratings provide a vastly unfair and biased appraisal of school effectiveness.

The dimensional aspect of the assessment, however, permits a more optimistic appraisal. Specifically it showed that when the different dimensions of instructional practices were dissected from one another there was evidence to suggest that the quality of instructional practices had a greater impact than the regularity of teaching behaviours, their focus or timing (see Table 12.4e). In fact, the quality of instruction variables had a comparable average effect size to the examination entry variables and explained more variation than the intake variables.

Table 12.4e: The average variance in Progress 8 scores explained by each dimension of the instructional variables during the 2017, 2018 and 2017-18 Detailed Regression Analyses

Dimension	Average variance explained		
	2017	2018	Change analysis
Frequency	33.2%	23.1%	20.0%
Focus	18.8%	8.2%	9.2%
Stage	9.3%	7.3%	17.0%
Quality	Not assessed	35.1%	23.7%
Differentiation	Not assessed	Not assessed	11.9%

This observation is consistent with Creemers and Kyriakides' (2008) empirical appraisal of their measurement framework, which concluded that all dimensions of effectiveness factors helped to increase the percentage of students' performance that could be accounted for. The rank order of the dimension's influence was also similar in both studies (see Section 7.4.3, part 2 for further details).

Had the sample been large enough to have modelled the collective effect of variables we may therefore have observed that the combined influence of several quality variables was similar to or greater than the bias introduced by intake or examination entry differences. So whilst the analysis suggests that Progress 8 ratings are biased by unchecked differences in school intake, it is possible that the relative influence of school-related and non-school factors may not be as damming as it first appeared. This interpretation however is speculative and requires verification from future research.

Details of unexpected associations

Across the three analyses 30.4% of the relationships between instructional variables and performance were inconsistent with the pre-specified expectations. Given the sample size used in this analysis, the lack of statistical controls and the fact that the unexpected associations accounted for lower percentages of variance than the expected relationships, it is logical to assume that in some instances the causal effect of the variables were overwhelmed by extraneous influences. There were however three consistencies within the data to which the reader's attention will now be draw. Collectively these explanations could account for a substantial portion of the unanticipated results.

Firstly, over 1/3 of the unexpected associations involved variables which evaluated the parts of the lesson or year in which activities took place. It is probably not a coincidence therefore that the timing of activities was one of the more difficult concepts to capture using a questionnaire that was filled out at a specific moment in time. If these questions were to have provided an ineffective or overly generalised account of these differences, this would account for the higher degree of unexpected associations in the area. The potential shortfall may also have artificially lowered the percentage of variation that these variables explained.

Another significant deviation from the pre-specified expectations was the negative association that was observed between *the frequency with which teachers responded to classroom disruptions* and schools' Progress 8 ratings from 2017 and 2018. As these variables were intended to report upon the proportion of classroom disruptions that teachers responded to, it was assumed that more regular intervention would maximise active learning time and therefore schools' performance. There are, however, at least two explanations that could account for the negative associations. The first theory, which is emphasised within teaching strategies such as active teaching (Brophy and Good, 1986), is that attempts to address off-task behaviour can elicit negative effects if the interventions themselves disrupt the flow of the lesson. Such approaches therefore encourage teachers to use their body language, non-verbal gestures and their positioning within the room to discourage inappropriate behaviour without delaying instruction. Unfortunately, the questionnaire used in this study did not distinguish between different types of intervention so it is impossible to discern whether this was the case. A second explanation is that the wording of the question (see variable name cited above) may have been too ambiguous and therefore have led respondents to comment upon the frequency of the disruptions themselves. Whilst the intended meaning was stated more explicitly within the item's sub-heading, it is possible that some respondents skim read the questionnaire and missed the clarification. This latter justification does not however account for the variable having a positive association with performance within the change analysis.

One outcome that defied explanation, however, is that the 4 variables which evaluated the consistency of teachers' instructional practices exhibited strong negative associations with school performance (see results of 2018 and change analysis for further details). This is confusing given that 3 of the items were operationalised in a way that implied that greater uniformity would always be advantageous. Whilst variety in 'the teaching style(s) used by teachers' may conceivably have allowed teachers the freedom to be more creative in addressing students' needs, there is no obvious explanation for large deviations in instructional quality, teachers' coverage of the curriculum and utilisation of lesson time appearing to have been advantageous. The best defence for the irregularity is therefore to point out that these variables were not specified within the Dynamic Model of Educational Effectiveness. These variables were added to the analysis to help compensate for classroom-level data being unavailable. Strictly speaking the observed relationships did not therefore conflict with past research, merely this researcher's assumptions.

School Policies:

Overall school policy variables were the least information category. These factors were, however, more useful in predicting the changes in schools' performance over time than the differences between schools' performance.

Table 12.4f: The average variance in Progress 8 scores explained by school policy factors in 2017, 2018 and 2017-2018 Detailed Regression Analyses

Sub-group of variables	2017 Analysis	Rank	2018 Analysis	Rank	2017-18 Average	Rank	2017-18 Change	Rank
School teaching policies	8.6%	2	10.1%	2	9.4%	2	22.5%	1
SLE policies	6.8%	3	3.1%	4	5.0%	3	20.1%	2
Evaluation of the school teaching policies	10.8%	1	22.4%	1	16.6%	1	13.6%	3
Evaluation of the school learning environment	1.8%	4	4.9%	3	3.4%	4	5.9%	4

* Average scores refer to the mean two annual scores reported in table.

Whilst the rank-order of factors' influences deviated across the analyses, there were clear patterns in the results (see Table 12.4f). Within the annual analyses, for example, the evaluation of school teaching polices always accounted for the largest percentages the variation in schools' results, followed by the schools' teaching policies. The policies for establishing an effective learning environment and the mechanisms used to evaluate the school learning environment explained comparable but lesser portions of the scores. Similarly, in the change analysis the policies and evaluation procedures that regulated classroom instruction accounted for more variation, on average, than the procedures that govern stakeholders' behaviour outside of lessons. Though in this instance it was the school policies that were the more effective predictors. These observations are logical as classrooms are the locus of the educational experience (Creemers, 1994; Scheerens, 1992). It could also be argued that it makes sense that changes to the schools' evaluations had a lower impact upon the next year's performance ratings as any adaptations will take time to elicit there full effect (Creemers and Kyriakides, 2008), and that this is especially true for the evaluations procedures which can only impact upon teachers' behaviours indirectly by influencing other practices.

The analysis also revealed that the quality, focus and level of differentiation within school policies had a far greater association with school performance than the number of policies that schools produced (see Table 12.4g).

Table 12.4g: The average variance in Progress 8 scores explained by each dimension of the instructional variables during the 2017, 2018 and 2017-2018 Detailed Regression Analyses

Dimension	Average variance explained		
	2017	2018	Change analysis
Frequency	3.2%	3.4%	4.4%
Focus	9.2%	11.8%	18.7%
Stage	5.5%	8.2%	14.8%
Quality	9.5%	14.9%	18.7%
Differentiation	Not assessed	Not assessed	22.1%

This is not only rational in the sense that the presence of a policy document does not in itself guarantee changes in school practice but is also something that was foreseen by the creators of the Dynamic Model.

The results therefore imply that differences in school policy do impact upon schools' performance ratings. In the majority of analyses, however, the effects attributed to school-level procedures was surpassed by the average effect size of non-school factors and it is therefore plausible for schools' ratings to be overwhelmed by extraneous influences. Furthermore, whilst the average effect attributed to the frequency and quality dimension of schools' instructional practices often exceeded the mean impact of intake and/or examination differences, implying that some aspects of teachers' behaviours had the potential to overcome the influence of extraneous sources of bias, this was not the case here. In fact, the range of predictive capacities expressed in Table 12.4g serves mainly to emphasise that school policies could have an even smaller effect if their content is inadequate.

In the interest of producing a comprehensive report, it is noted that whilst the majority of the associations from this section were logical and in line with the theoretical arguments outlined within the Dynamic Model, the rank order of factors' mean effects was inconsistent with past research. There was therefore a r = -0.628 correlation between average effect sizes reported in the 2017 assessment and the average effect sizes reported in Creemers and Kyriakides (2008) meta-analysis of school-level effectiveness factors, and a r = -0.796 association between the 2018 average effect sizes and the same figures. This may be attributable to the sample size, the small and comparable effect of school policy variables or the lack of statistical controls within the current analysis. The change analysis results were more in-line with the meta-analysis and had an r = 0.418 correlation with effect sizes reported for schools' policies.

A connected issue is that this subsection of analyses had the lowest rate of anticipated interactions (66.0%). Specifically, 23/40 of the relationships from the 2017 models, 17/35 of the assessable directional effects from the 2018 models and 58/75 of the assessable directional effects from the change analysis were consistent with the effect that the variables have upon students' raw performance. This suggests that construct irrelevant variance might have impacted upon the measures. The ratio of expected to unexpected associations was still favourable, though, which helps to legitimise the study's conclusion.

Some of these anomalies could also be explained by shortfalls in the operationalisation of the independent variables. Nearly 1/3 of the unexpected directional associations related to variables that evaluated either the number of years that school policies had been implemented, or the average number of years between modifications of the policies. The foremost of these items was intended to reflect Creemers' and Kyriakides (2008) supposition that for policies to have the largest effect they need to be implemented consistently throughout students educational career, in this instance, from the end of Key Stage 2 when the prior-attainment measures were taken to the end of Key Stage 4 when students' final attainment level was evaluated. The latter reflects the need for flexibility in the schools' approach (see, for example, Cousins and Leithwood, 1986; Thomas, 2001). Whilst the operationalised measures were intended to differentiate between the establishment of policies and their modification, it is conceivable that in practice this distinction is not clear cut and that the reported associations may have been influenced by the resulting ambiguity. This possibility enhances the creditability of the results by suggesting that there may have been rational reasons for many of the unexpected directional associations.

12.5 Conclusion

This section was intended to establish whether variation in school-level Progress 8 ratings is indicative of genuine differences in school effectiveness.

Whilst all of the regression models agree that correlates from educational effectiveness research were able to account for meaningful proportions of the variation in schools' value-added scores, the evidence collected in this section suggests that external factors such as the differences in school intakes and examination entry practices have a significant influence upon schools' scores. In fact, if the regression outputs are accepted at face value, then the average effect of deviations in school intake and examination entry practices exceeds the mean effect of instructional and policy variables. This does not, of course, prove that the combined influence of school-related variables is surpassed by the effects of non-school factors but is nevertheless indicative of problems within the calculation.

That being said, there are several reasons for suspecting that the situation may not be quite as bad as the results imply. These include the caveats discussed in the Shallow Regression Analyses. Namely that the percentage of variance explained by the examination entry variables would reduce if differences in schools' intake, instructional practices and policies were accounted for. It is also possible that the use of cohort-level data, school-leader questionnaires and the decision to classify absence as a extraneous variable may have reduced the effect attributed to schools, whilst any reciprocal relationship that exists between examination entry variables and performance would have exaggerated the variance that non-school factors could explain. Furthermore, a portion of the effect that was attributed to intake bias may in fact stem from the compositional effects, which would count toward the schools' Type A effect (the overall effect of attending one school over another) but not their Type B effects (the differences which are attributable to the quality of the schools' provisions). All of which would increase the percentage of variance that was attributed to non-school factors, whilst reducing the variance ascribed to schools'.

A unique contribution of this section, however, was to evaluate the effect of alternative dimensions of effectiveness. That is to say, whether acknowledging the focus, timing, quality and differentiation of school policies and practices would noticeably increase the percentage of variance that school-related factors can explain. The analysis suggests this is the case. In particular, the results suggest that the quality of teachers' actions have a greater impact upon performance than the regularity of particular behaviours. Whilst the characteristics of school polices (i.e. their specificity, purpose, implementation, quality and the level of differentiation that is encouraged) have a greater impact than the number of policies produced or the number of areas that they cover. Were it possible for the effect of multiple factors to have been modelled simultaneously, the analysis may therefore have found that school-related variables account for a more reassuring portion of the variance in schools' results. This would not negate the bias introduced by intake and examination entry variables, but justifies the attention of future research.

Addendum: A note on the plausibility of non-linear relationships

Creemers and Kyriakides (2008) assert that all of the factors within their Dynamic Model have a positive association with student attainment[40]. The frequency and focus variables however are unique in that they may have an optimum point beyond which further increases would become counter-productive. These non-linear effects could potentially explain some of the unanticipated associations observed within this section and/or increase the percentage of variance that was attributed to school-related factors.

The decision to evaluate the linear effect of variables is defended on three grounds. First, because it made the assessment more objective. Whilst conducting this analysis we attempted to evaluate whether the results for frequency and focus variables were more consistent with the pre-specified expectations if the relationships were modelled using a quadratic function. The methodology was ultimately rejected, however, because almost any relationship could be interpreted as being consistent with the hypothesised inverted-U association if the researcher was willing to assume that they were looking at a segment of a larger distribution. That is to say, the start, middle or end of an inverted-U association. Moreover, if a stringent criterion was used, then there was little change in the percentage of compliant functions. Similarly, within the change analysis we tested a procedure for interpreting whether increases and decreases in the potency of a variable had the expected effect, based on the distributions observed in the annual analyses. In other words, whether an improvement would have moved the school towards or away from the theorised optimum point. Similar problems were encountered including how to make the distinction if the distributions observed within the 2017 and 2018 models conflicted. It was therefore decided that it was better to have a clear and objective criterion for specifying whether a relationship was in-line with our expectations. Second, with only nine observations a quadratic function inevitably increased that percentage of variation that variables could account for, whether the function was in-line with our expectations or not. Utilising a quadratic model would therefore have risked artificial inflating the percentage of variance that factors could explain and distorting the results of the analysis. Finally, non-linear effects occur infrequently within non-experimental studies that are conducted within a single country and/or context (see section 7.4). The decision to focus upon the linear effect of variables was therefore unlikely to have a substantive influence upon the analysis' results.

[40] Note that this is not the case within our empirical analyses because some of variables were operationalised as the inverse of the factor they describe. The percentage of disadvantaged students, for example, is the inverse of socio-economic status.

13. Conclusions

13.1 Chapter Introduction

This chapter concludes the thesis by collating the evidence from the four empirical sections. A final judgement is then made as to the validity of Progress 8 assessments and the implications for research, policy and practice. In particular, these discussions address the question of whether Progress 8 provides a valid and reliable indicator of schools' contribution (Type B school effect) that assists parents in selecting an effective school for their child (Type A school effect)

13.2 Summary of research findings

Prediction Analysis:

The first empirical section assessed the validity of Progress 8 by drawing upon the knowledge of school leaders. As experts of their institutions it is reasonable to expect that these individuals would have an intermit knowledge of the factors that school effectiveness is normally attributed to, and would therefore be able to anticipate the rating that their school would receive in advance. Despite a base level of agreement between leaders' estimates and their schools' official ratings, the analysis concluded that school leaders' insight was minimal. There are two explanations for this. Either schools' progression ratings were valid, but school effectiveness was so volatile that the year-to-year change in other schools' scores prevented meaningful predictions from being made. Or there was sufficient construct irrelevant variance within the scores to render them an ineffective measure of schools' performance.

Both scenarios are problematic. If school effectiveness is presumed to change drastically each year then it is questionable whether any performance measure could provide parents with reliable information about the education their child will receive. Moreover, if head-teachers, armed with the schools' most recent value-added rating, detailed information their schools' practices, data on students' current attainment levels and the previous years' attainment averages cannot make reliable predictions about their schools' performance immediately before students' sit their GCSE examinations, how can parents possibly be expected to make the same determination six years in advance? The first and most definitive conclusion of this thesis is therefore that, even if Progress 8 ratings are assumed to be valid, they do not provide meaningful insight into schools' long-term performance. It is also problematic that leaders cannot anticipate cohorts' scores as this implies that it is difficult for them to make proactive school improvement decisions, rather than simply responding to the ratings of students that have left the school.

It is very likely, however, that at least a portion of the instability was artificial. That is to say that school leaders' inability to foresee their value-added results may hint at their being problems with the underlying calculation. More specifically, with the way extraneous influences upon students' performance are controlled. Such a conclusion would suggest that Progress 8 provides not only an unreliable measure of schools' long-term performance, but also an inaccurate measure of their current effectiveness. The phrase 'very likely' is used here because even the academics who developed the first DfE value-added measures argued that statistical controls could never account for all of the differences between school-cohorts, and that this type of assessment would therefore provide at best an estimate of school performance (Fitz-Gibbon, 1997). The remaining section of the thesis explored the severity of this problem.

Thought Experiment:

During the previous analysis it was assumed that all school leaders would have access to students' Key Stage 2 prior-attainment fine levels. This would not, therefore, have interfered with their ability to predict Progress 8 ratings in advance. In real-world situations, however, this information would not be 100% accurate. The reasons for this were discussed in Chapter 4. Assuming though, that what the specification really requires is a perfect assessment of students' aptitude (i.e. a statement of their initial knowledge and ability), then it follows that even the best evaluations would contain a proportion of measurement error because many students' would perform better or worse than normal on test day.

This analysis therefore considered the implications of there being measurement error within students' KS2 and KS4 data. The results suggest that the effects are not comparable. In fact, the impact of KS2 error can up to 2.5 times greater. This is because the scores of students with comparable prior-attainment fan out over time, most likely because of the differential effect that schools have upon students of different ability levels. If any of the KS2 measurement errors were to be non-random, then the capacity for them to impact upon schools' Progress 8 ratings is substantial. To our knowledge, this is a unique observation that has not been addressed in past research. Additional studies are needed to explore the implications.

Shallow Regression Analysis:

The Shallow Regression Analysis was the heart of this thesis. In this assessment the Progress 8 scores of a nationally representative sample of 125 schools were regressed upon effectiveness correlates from school effectiveness research. The results suggest that non-school factors had an unacceptable level of influence upon schools' value-added scores.

To be more specific, across the three datasets (2017 data, 2018 data, and 2017-2018 changes) and the three types of regression model (simple linear, forward and hierarchical), two categories of extraneous variable had a close and persistent associations with schools' performance ratings; differences in school intake and differences in schools' examination entry practices.

When the underlying structure of the data was acknowledged, the regression outputs suggest that differences in school intake could account for more than 40% of the variation in schools' annual performance ratings (see results of hierarchical models). If this evidence is accepted then it follows that Progress 8 ratings provide an extremely biased appraisal of school performance that advantages schools with particular intakes. These variables also help to explain the changes in schools' performance over time, though the effect recorded within this study was more modest (5.4% of variance was explained by changes in the three most influential intake characteristics).

Examination entry variables also correlated with performance. As one would expect, the more Attainment 8 slots that students filled the higher schools performance ratings was likely to be. The effect was sizeable, however, even after differences in schools' intake, instructional practices and policies had been taken into account (10.4%-12.5% of the variance in schools' annual performance ratings was explained by these factors, and 20.2% of the variance in schools' ratings over time)(Again see results of hierarchical analyses). It is argued here therefore than these variables may also play too great a role in the determination of schools' Progress 8 scores, especially when one considers the percentage of variance accounted for by other non-school factors. This conclusion can be disputed, however, as it was the DfE's intention to use the school-level value-added scores as a means of encouraging schools' to provide students with an education that covers particular areas. The degree of impact that these variables should have is therefore a qualitative decision.

Within the analysis these two groups of variables accounted for more variance, on average and collectively, than the operationalised aspects of schools' provisions. This includes factors associated with schools' teaching policies, policies on the school learning environment, the policies for evaluating schools' performance, as well as 8 aspects of teachers' instructional behaviour. Though a substantial portion of the variation in schools' ratings remained unexplained, and could therefore be attributed to school-related or extraneous variables, this is a concerning finding which suggests that the bias within Progress 8 measures has the capacity to overwhelm the genuine differences in schools' effectiveness.

Detailed Regression Analysis:

The final empirical section took a closer look at the performance of 9 schools. This time schools' ratings were regressed upon a far wider range of school-related variables, including measures that assessed the frequency of schools' practices and policies, their focus (specificity and purpose), their timing, quality and the level of differentiation that took place. The analysis thereby expanded upon the previous assessment by considering dimensions of effectiveness that were previously neglected.

The results showed that school effectiveness factors could account for a higher percentage of variance when these perspectives are considered. The quality of teachers' instructional behaviour, for example, was of particular importance as it explained more of the variation in schools' performance than the frequency of specific actions and/or activities. Likewise all aspects of school policy were shown to be more influential that the number of documents that a school had in place.

The substantive findings of the analysis were, however, no different. Intake and examination variables still accounted for a higher percentage of variance, on average, than the effect of school-related variables.

It should be noted, though, that whilst the rank-order of examination entry variables importance was fairly regimented across the analyses and absence was consistently the most predictive intake factor, the association between the other intake variables deviated between the two sets of analyses (shallow and detailed). For example, the percentage of disadvantaged students per cohort had a more consistent association with schools' performance ratings in the shallow analyses, whereas the variables associated with special education needs and English as an additional language status explained more variation in the detailed analysis. As the mean percentage of students per cohort with these characteristics deviated substantially across the analyses[41], the intuitive explanation is that the differences reflect the characteristics of the respective samples. The associations within the larger nationally representative sample are therefore assumed to be more generalizable.

[41] In 2018, the percentage of disadvantaged, unstatemented SEN and EAL students was 20.4%, 9.9% and 10.7% respectively within the nationally representative shallow regression analysis sample, and 28.2%, 14.9% and 2.6% within the case study of 9 northern schools. The percentage of disadvantage and unstatemented SEN student was therefore substantially lower in the larger nationally representative sample, whilst the percentage of EAL pupils was much higher.

13.3. Overall Conclusion

This thesis therefore concludes that Progress 8 does not provide a valid and reliable measure of school performance. The scores are more volatile than some researchers expected and whilst this could theoretically be explained by genuine changes in school performance, the evidence suggests that this is not the case. Of greatest concern is the impact that school intake appears to have upon schools' ratings. These factors, most noticeably differences in students' attendance and socio-economic status, have close associations with school performance that punishes schools with educationally disadvantaged cohorts. There is also evidence suggest that errors in students' prior-attainment have the potential to impact upon schools' ratings and that schools' examination entry practices have too great a sway over the results. What is more, school-related factors accounted for a surprisingly low percentage of the variation in the scores, and whilst sceptical readers may attribute part of this to the authors' methodological decisions (see discussion of limitations below), it is notable that school leaders were no more successful in explaining the results.

From a technical perspective it should be reiterated that the methodology of the two regression analyses was more adept at assessing the Type B effect of schools (the quality of institutions' contributions), than Type A effects (the overall effect attending one school over another), as the analyses would have interpreted any compositional effects as error. The legitimacy of compositional effects however is debated. Moreover, from a policy perspective this is somewhat of a null point as the volatility of Progress 8 ratings alone was sufficient to invalidate the notion that the ratings can provide parents with reliable insight into the effect that a school would have upon their child's development. Assuming, that is, that that they transition between schools at the traditional points.

13.4 Limitations

When designing the empirical investigations of this thesis, a conscious effort was made to ensure that the validity of Progress 8 was evaluated in an objective manner using robust research designs. The process of measuring school effectiveness cannot, however, be reduced to an entirely technical matter. At several points methodological decisions were made that may have impacted upon specific results or the meaning that was derived from them.

The most divisive choices were identified within the discussion sections of the respective analyses but include the supposition that school leaders should be able judge whether the standard of the their school's provisions has improved or declined, the decision to classify student absence as a non-school factor that was predominantly outside of the influences schools, the decision to measure differences in teaching practice indirectly using a questionnaire and the belief that curricular decisions should not have an overriding influence upon the performance rating that a school receives. Each of these stances informed the interpretation of research evidence.

Though each of the aforementioned discussions outlined why the adopted positions are defensible, it is recognised that alternative perspective also have merit. The sections therefore go on to discuss the implications of these assumptions being rejected (see individual sections for more specific information).

Whilst the refutation of one or more of these assumptions would lessen the claims of invalidity, it should be noted that there are limits to the effect. Just as the instability of Progress 8 rating is sufficient to undermine some applications of Progress 8 irrespective where the variation originates, neither the fact that schools' exert some influence over attendance or the likelihood of the operationalisation of school-related variables being imperfect detracts from the relationship that the remaining intake characteristics had with school performance. So whilst a sceptical interpretation of this research evidence might conclude that the model is more accurate and fairer than implied, and that there is scope for school-related variables to explain more variation than the observed sources of bias, it would be difficult to contest the assertions that Progress 8 residuals are too volatile to make reliable long-term predictions about the performance of individual schools or that the calculation is vulnerable to forms of intake bias that are likely to advantage particular types of school.

Another confounding variable is the fact that all analyses were forced to rely upon school-level data (see Section 8.3). In theory the failure to acknowledge the clustering of pupils within classrooms and schools may have led to biased regression co-efficients. It is argued, however, that this is unlikely to have been the case as the substantive conclusions of this study are roughly in-line with other critiques of the DfE value-added models. Leckie and Goldstein (2019), for example, concluded that Progress 8 results are unfairly biased by differences in school intake. Whilst Perry (2016a; 2019) drew similar conclusion about the comparable Best 8 model.

13.5. Implications for Policy, Research and Practice

Policy implications:

The evidence collated within the thesis suggests that the DfE's use of Progress 8 should be reconsidered. To put it bluntly, the measure is not valid or reliable enough to be used to make high-stakes decisions. The ratings are too biased to provide a fair measure of schools' contributions and too unstable to provide parents with dependable information about the effect of attending one school over another.

Particular objection is taken to the decision to ignore differences in pupils' demographic and socioeconomic characteristics. This is profoundly unfair. It has long been established that different sub-groups of student have different levels of mean achievement, and that the reasons for this inequality extend far beyond schools (Coleman *et al.*, 1966; Jencks *et al.*, 1972; EPI, 2017). Moreover, it is widely acknowledged that schools' have a limited capacity to address this disparity or the wider inequalities within society (Teddlie and Reynolds, 2000). In fact, in most cases the pre-existing gaps tend to widen rather than diminish during students' education because the underlying reasons for the disparity persist (Thomas *et al.*, 1997a). Whilst most people would agree that schools bear some of the responsibility for addressing this inequity, failing to acknowledge broader societal influences and their effect upon students' achievement essentially credits or blames schools for the educational affluence of the populations that they serve and overlooks the broader societal influences. Many researchers have therefore asserted that uncontextualised value-added models, such as Progress 8, are likely to reward and punish the wrong schools (Leckie and Goldstein, 2019) and the results of this thesis support this supposition.

The inaccurate classification of schools, however, is not only unjust. It has the capacity to undermine the national accountability system and any effects that it has upon students' learning. Furthermore, when used in high-stakes situations uncontextualised value-added models may discourage schools from admitting particular types of pupil, or encourage them to find ways of excluding them from examinations and therefore the value-added figures. Indeed, since the introduction of Progress 8 there has been a notable rise in pupil exclusions (DfE 2018), which has been partially attributed to schools attempting to game the accountability system (Leckie and Goldstein, 2019). It is likewise important to remember that students' progression scores also reflect upon teachers. So, if disadvantaged schools are more likely to receive negative ratings, there is therefore incentive for effective teachers to relocate to advantaged schools where the efforts and skillsets are more likely to be recognised. Both side effects would exacerbate existing social inequalities.

In terms of accountability, the only defensible application of the measure would be for it to be used as a screening device to identify schools that justify more intensive scrutiny. The intake of the institution could then be considered, along with other aspects of the schools' performance.

Implications for research:

Despite the problems with the measure, there was some evidence that schools' value-added residuals still reflected the impact of schools' practices. In less consequential contexts such as research, the use of Progress 8 and/or comparable models of effectiveness is therefore more defensible, though one suspects that most academics will favour alternative specifications. Whilst a full review of alternative indicators was beyond the scope of this thesis, growth models and regression discontinuity designs have the potential to negate many of the flaws associated with the DfE measure. Their approach is conceptually superior as it removes intake bias by design, rather than relying upon flawed statistical controls.

Further studies are also needed to assess the impact that examination-entry variables have upon composite value-added scores. The empirical investigations of this thesis suggest that curricular decisions had a substantial impact upon schools' Progress 8 ratings. The magnitude of this influence was insufficient to invalidate the measure on its own, but was nevertheless sizeable and added an additional source of construct irrelevant variance to a measure that already rests upon a dubious assumption (that any variance between students' attainment that cannot be explained by differences in their Key Stage 2 fine-levels is attributable to schools). The scope of these analyses, however, was limited. Alternative specifications of model were not considered and it is therefore impossible to say whether the stringent inclusion criteria of Attainment 8 increased or decreased the percentage of variance that these decisions can explain. Whilst it is recognised that reduced level of flexibility was successful in limiting the opportunity for schools to exploit discrepancies in the workload associated with different types of qualification, it is important to understand the cost of these safeguards.

Implications for practitioners and parents:

Schools may wish to use the measure, or an alternative specification of value-added model, as a tool for self-evaluation. Though it should be stressed that the reliability of the data is likely to deteriorate the more the figures are broken down. Departmental ratings will therefore be less dependable than the school-level ratings, and teacher-level ratings less dependable than departmental ratings.

In terms of parents and the matter of school choice, it remains defensible for parents to consider schools' Progress 8 scores when making educational decisions. It should be acknowledged, however, that the figures are only estimates of school performance and that the confidence intervals that are currently attached to the measure provide an ineffective summary of the potential for error. The ratings are also time-specific and do not necessarily represent schools' future performance. The author's advice would therefore be to use the information that is available but to interpret it alongside more hands-on information. Visits to schools, observations of lessons, comparisons of school curricular and discussions with teachers, for example, can provide insight into schools' ethos and learning environment that may not be reflected in attainment or progression figures.

Appendix A: Short Questionnaire

A study of national school improvement approaches

A Durham University study

Introduction

Thank you for taking part in our study on national school improvement approaches. This questionnaire consists of a single question and will therefore take seconds to complete. Your responses will be used to investigate how much of the year-to-year variation in schools educational provisions is picked up by value added assessments. We would like to emphasize that our research is focused on the government's policy for assessing secondary schools and not the performance of individual schools. Any data disclosed will be kept confidential and stored securely until it is destroyed. Schools and individuals will not be identifiable in the write up.

Section A: Self-evaluation of your school's educational provisions

Please estimate the value-added score that your school will receive in 2018*

In the space below please predict what your school's value added score will be in 2018. This is often referred to as the Key Stage 2 to Key Stage 4 Progress 8 score. You should provide an exact number, not a range of values. When devising this estimate it may help to consider the score that your school received last academic year and any changes that have occurred within your school since that time. Possible changes include variations in the quantity of instruction students received, the school curriculum, the quality of instruction and the school learning environment. Factors that influence these variables such as the school's funding, differences between the year 11 cohorts, changes in the teaching staff or school leadership and exam entry procedures may also be relevant. You may consider factors we have not mentioned.

....

Thank you for providing this information. If you wish you may submit your response now. If you are willing, however, there are a couple of extra questions we'd like you to complete. These should take approximately 3 minutes.

- ○ Continue
- ○ Submit questionnaire

Section B: Additional questions
Part 1: School policies

Have any of the following changed since last academic year?

Select N/A if your school does not have a policy for a particular area.

	N/A	Large decrease	Small decrease	No change	Small increase	Large increase
The amount of instruction time that the school policies provided students with (consider timetables, homework, cancelled lessons and your policy regarding absenteeism)	○	○	○	○	○	○
The alignment between the school curriculum and the content assessed in students' Key Stage 4 courses.	○	○	○	○	○	○
The quality of the polices on teachers' instructional behaviour (inc. the support and/or training available)	○	○	○	○	○	○
The quality of the policies that regulate the school learning environment (students' behaviour outside the classroom, teacher collaboration, school partnerships, the provision of learning resources and teachers'/students' attitudes towards learning).	○	○	○	○	○	○

Since last academic year, has the school made any changes to its exam entry policies?

Examples include variations in the number of students entered for the EBacc mathematics qualification, the EBacc English language and English literature qualifications, other EBacc subjects, non-EBacc GCSE's and non-GCSE qualifications, and the number of early entry AS level qualifications. Differences in the total number of examinations students enter may also be relevant. If any such changes have occurred please identify these below and predict whether each will have a) a large negative influence , b) a small negative influence, c) no influence, d) a small positive influence or e) a large positive influence, on your school's value added score.

....

Part 2: The evaluation of school policies

The term 'evaluation of school policies' refers to the school's own procedures for monitoring and assessing their educational provisions. These evaluations will usually consider pupil attainment levels as well as data collected about specific aspects of policy, such as attendance levels or teacher evaluations. For the purpose of this questionnaire, a distinction has been drawn between the evaluation of teaching (which considers the quantity of instruction, the appropriateness of the school curriculum and teachers' instructional behaviour) and the evaluation of the school learning environment (which includes students' behaviour outside the classroom, teacher collaboration, school partnerships, the provision of learning resources and student/teachers values about learning).

Since last academic year, has there been a change in the quality of the school's internal evaluation procedures?

High quality evaluation mechanisms are be reliable and appropriate for their proposed use. They assess all aspects of teaching or school learning environment, draw upon data from several sources and provide information that is useful for making managerial decisions. It is also expected that these mechanisms operate continuously throughout the year, not merely at the end of certain periods. Select N/A if your school does not conduct these evaluations.

	N/A	Large decrease	Small decrease	No change	Small increase	Large increase
The quality of the school's procedures for evaluating the school teaching policies (inc. policies on the quantity of instruction, teachers' instructional behaviour and the school curriculum)	○	○	○	○	○	○
The quality of the school's procedures for evaluating the school learning environment (inc. student behaviour, teacher collaboration, school partnerships, resource allocation and teacher/student values about learning	○	○	○	○	○	○

Were changes to following policies based on these evaluations?

If these polices have not been changed since last academic year, please consider whether the decision to retain the existing policy was based on the school's evaluation data.

	No	Yes
The changes to the school teaching policies (inc. policies relating to quantity of instruction, teachers' instructional behaviour and the school curriculum)	○	○
The changes to the policies on the school learning environment (inc. policies on students' behaviour outside of the classroom, teacher collaboration, school partnerships, the provision of learning resources and teachers'/students' values about learning)	○	○

Part 3: Classroom instructional behaviours

How frequently did the following behaviours occur LAST academic year?

	1. Almost Never	2. Infrequently	3. Routinely	4. Frequently	5. Very Frequently
Orientation tasks	○	○	○	○	○
Structuring tasks	○	○	○	○	○
Questioning	○	○	○	○	○
Teacher-modelling (teaching problem solving skills)	○	○	○	○	○
Application tasks (seat-work or small group tasks)	○	○	○	○	○
Task-related Teacher-student interactions	○	○	○	○	○
Task-related student-student interactions	○	○	○	○	○
Classroom assessments	○	○	○	○	○
Classroom disruptions caused by poor student behaviour (please note that this is the only negative behaviour in this list, a score of 5 is therefore poor)	○	○	○	○	○

How frequently did the following behaviours occur THIS academic year?

	1. Almost Never	2. Infrequently	3. Routinely	4. Frequently	5. Very Frequently
Orientation tasks	○	○	○	○	○
Structuring tasks	○	○	○	○	○
Questioning	○	○	○	○	○
Teacher-modelling (teaching problem solving skills)	○	○	○	○	○
Application tasks (seat-work or small group tasks)	○	○	○	○	○
Task-related Teacher-student interactions	○	○	○	○	○
Task-related student-student interactions	○	○	○	○	○
Classroom assessments	○	○	○	○	○
Classroom disruptions caused by poor student behaviour (please note that this is the only negative behaviour in this list, a score of 5 is therefore poor)	○	○	○	○	○

Has the quality of teachers' instructional behaviour changed since last academic year?

Good quality instruction is clear and influences students learning.

	N/A	Large decrease	Small decrease	No change	Small increase	Large increase
Quality of instructional behaviour	○	○	○	○	○	○

Since last academic year, has there been a change in the proportion of lesson time that was used for teaching?

Please note that this refers to the amount of time that the class were engaged in learning activities. This includes time dedicated to orientation and structuring tasks, time spent in classroom discussions and/or listening to the teacher lecture about a topic. It does not include time spent on classroom management (e.g. organizing the group or dealing with disruptive behaviour) or social activities.

	N/A	Large decrease	Small decrease	No change	Small increase	Large increase
The proportion of lesson time used for teaching	○	○	○	○	○	○

Since last academic year, have teachers covered the school curriculum to greater or lesser extent?

	N/A	Large decrease	Small decrease	No change	Small increase	Large increase
Teachers' coverage of the school curriculum	○	○	○	○	○	○

Part 4: Student-level changes

Were there any differences between the year 11 cohorts of students?

In the space below please identify any characteristics that were more or less common in the current cohort of year 11 students than they were in last year's 11 students. As a minimum please consider potential differences in the proportion of students who are a particular gender, socio-economic class, ethnicity, or have personalities and thinking styles that are suited to secondary education. Differences in students' intelligence, prior attainment levels, motivation and expectations may also be relevant. As is the number of students with special educational needs. For every difference you identify please specify whether the proportion of students with the characteristic has a) decreased dramatically, b) decreased slightly, c) increased slightly, d) increased dramatically.

Finally, since last academic year have there been any other changes which may influence your school's 2018 value added score?

If so please identify these below and specify whether you expect them to have a) a large negative influence, b) a small negative influence, c) no influence, d) a small positive influence or e) a large positive influence on your school's value added score.

Concluding message:

That concludes this questionnaire. Thank you for your participation

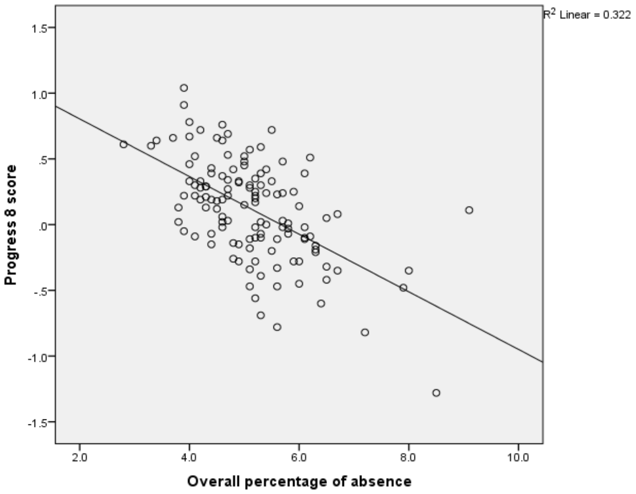
Appendix B: School Leaders' Predictions of their institution's 2018 Progress 8 Ratings

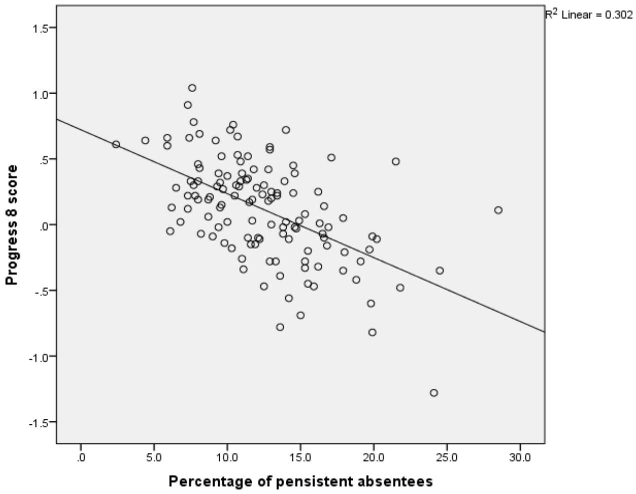
School number	Progress 8 score (2018)	Predicted Progress 8 score (2018)
1	-1.58	-1.2
2	-1.52	-0.9
3	-0.73	-0.3
4	-0.62	-0.4
5	-0.58	-0.25
6	-0.52	-0.1
7	-0.5	-0.24
8	-0.49	-0.1
9	-0.48	-0.45
10	-0.46	-0.4
11	-0.46	0.1
12	-0.42	0.06
13	-0.42	0.1
14	-0.41	-0.1
15	-0.41	-0.2
16	-0.4	-0.2
17	-0.4	0.2
18	-0.38	-0.15
19	-0.37	-0.45
20	-0.35	0
21	-0.32	0.09
22	-0.31	0.02
23	-0.3	-0.1
24	-0.28	0.1
25	-0.28	0.1
26	-0.25	-0.1
27	-0.24	-0.1
28	-0.24	-0.1
29	-0.23	0.28
30	-0.22	0.2
31	-0.21	0.3
32	-0.21	-0.04
33	-0.18	0
34	-0.17	-0.3
35	-0.17	0.2
36	-0.16	-0.25
37	-0.16	-0.19
38	-0.15	0.1
39	-0.15	0.01
40	-0.13	0.15
41	-0.11	0.3
42	-0.1	0.03
43	-0.1	0.12
44	-0.09	-0.1
45	-0.09	0.15
46	-0.09	0.01
47	-0.08	0.068
48	-0.08	0.01
49	-0.07	0.1
50	-0.07	0
51	-0.07	0.1
52	-0.06	0.03
53	-0.05	-0.35
54	-0.05	-0.1
55	-0.04	0.1
56	-0.04	-0.2
57	-0.04	-0.1
58	-0.03	0.1
59	-0.03	0.25
60	-0.03	-0.1
61	-0.01	-0.3
62	-0.01	0.42
63	-0.01	0.3
64	0	0.1
65	0	0.15
66	0	-0.1
67	0.01	-0.35
68	0.01	0
69	0.01	0.17
70	0.02	-0.2
71	0.02	0.15
72	0.03	0.2
73	0.03	0.1
74	0.04	0.16
75	0.04	0.07
76	0.05	0.38
77	0.05	0.1
78	0.06	0.2
79	0.06	0.39
80	0.06	-0.02
81	0.06	0.24
82	0.1	0.25
83	0.1	0.1
84	0.13	0.2
85	0.14	0.25
86	0.14	0
87	0.14	0.1
88	0.15	0.25
89	0.15	0.1
90	0.16	0.15
91	0.16	0.05
92	0.17	0.38
93	0.17	0.15
94	0.18	0.4
95	0.19	0.12
96	0.19	0.25
97	0.19	0.3
98	0.2	0.25
99	0.2	0.1
100	0.21	0.35
101	0.21	0.3
102	0.22	0.4
103	0.22	0.14
104	0.23	0
105	0.23	0.4
106	0.23	0.35
107	0.23	0.15
108	0.23	0.2
109	0.24	0.2
110	0.25	0.2
111	0.25	0.02
112	0.25	0.6
113	0.26	0.3
114	0.26	0.15
115	0.26	0.1
116	0.27	0.2
117	0.28	-0.2
118	0.28	0.4
119	0.28	0.35
120	0.29	0.35
121	0.3	0.28
122	0.31	0.2
123	0.31	0.25
124	0.32	0.02
125	0.33	0.55
126	0.33	0.3
127	0.35	0.2
128	0.35	0.35
129	0.36	0.15
130	0.37	0.4
131	0.38	0.11
132	0.39	0.5
133	0.4	0.4
134	0.4	0.3
135	0.41	0.2
136	0.41	0.35
137	0.42	0.34
138	0.42	0.9
139	0.43	0.4
140	0.44	0.4
141	0.45	0.45
142	0.45	0.48
143	0.46	0.41
144	0.46	-0.2
145	0.46	0.37
146	0.47	0.25
147	0.49	0.35
148	0.49	0.5
149	0.5	0.2
150	0.5	0.3
151	0.5	0.22
152	0.54	0.3
153	0.54	0.1
154	0.54	0.2
155	0.55	0.5
156	0.56	0.5
157	0.56	0.5
158	0.59	0.5
159	0.6	0.35
160	0.64	0.68
161	0.64	0.2
162	0.65	0.4
163	0.68	0.35
164	0.68	0.33
165	0.69	0.3
166	0.7	0.5
167	0.72	0.85
168	0.72	0.7
169	0.73	0.7
170	0.74	0.42
171	0.77	0.7
172	0.78	0.7
173	0.79	0.4
174	0.79	0.73
175	0.89	0.5
176	0.9	0.6
177	0.91	0.5
178	0.93	0.9
179	0.96	0.75
180	1.02	0.8
181	1.04	1.19
182	1.21	0.6

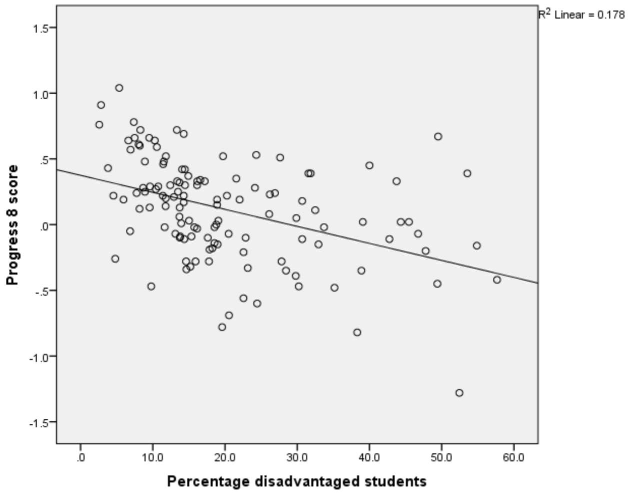
Appendix C: Scatter Graphs from the Shallow Regression Analysis

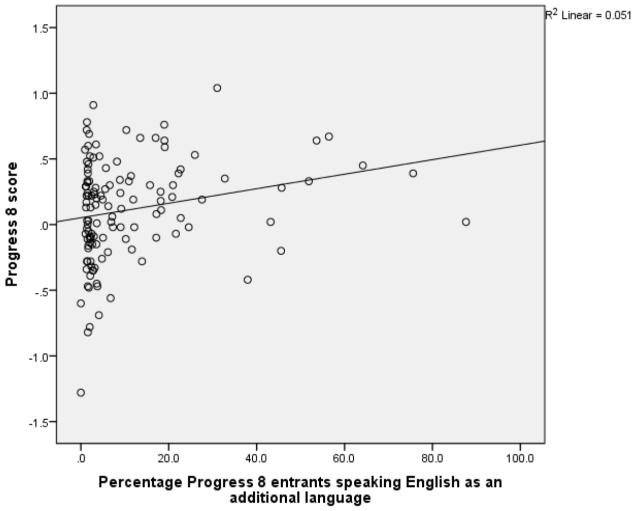
2017 Analysis:

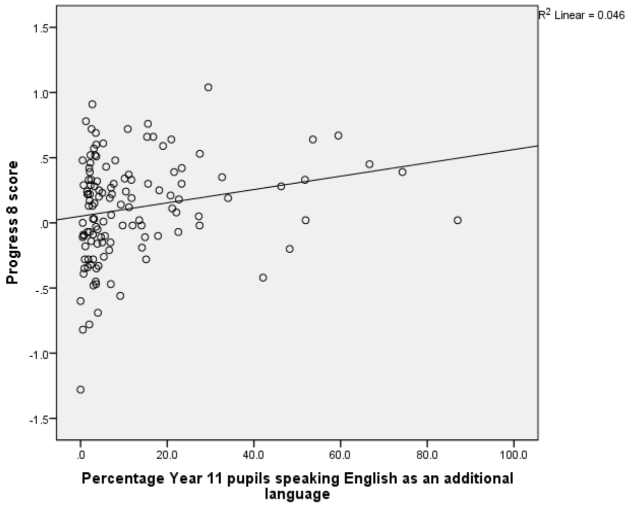
Intake variables:

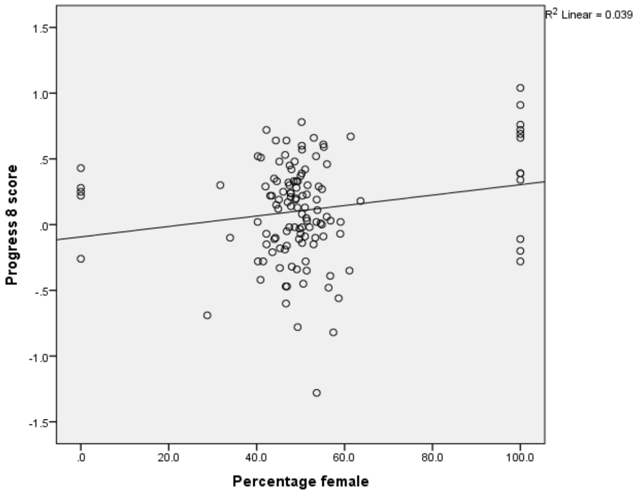


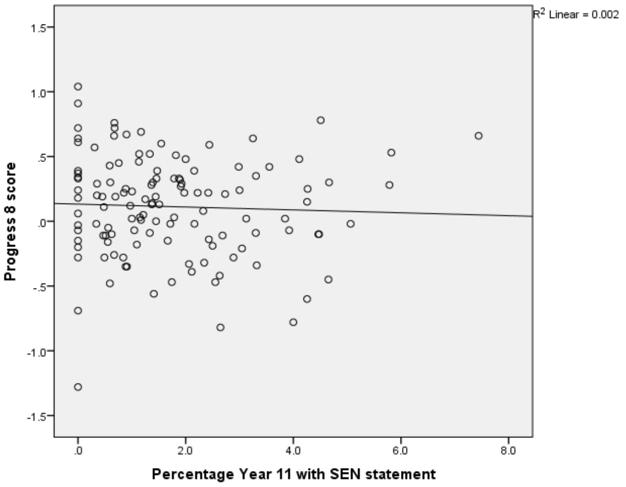


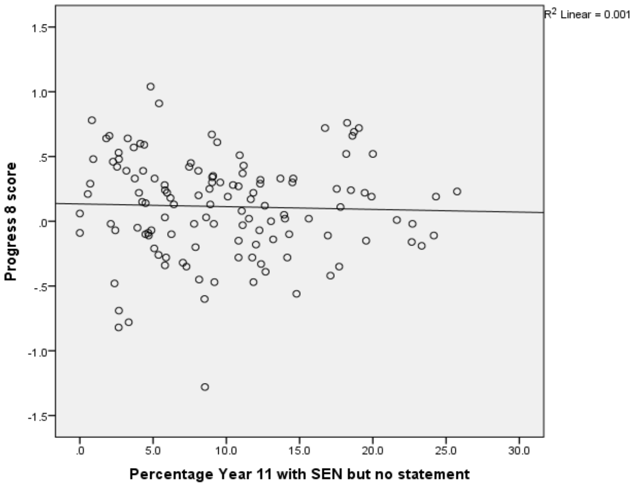


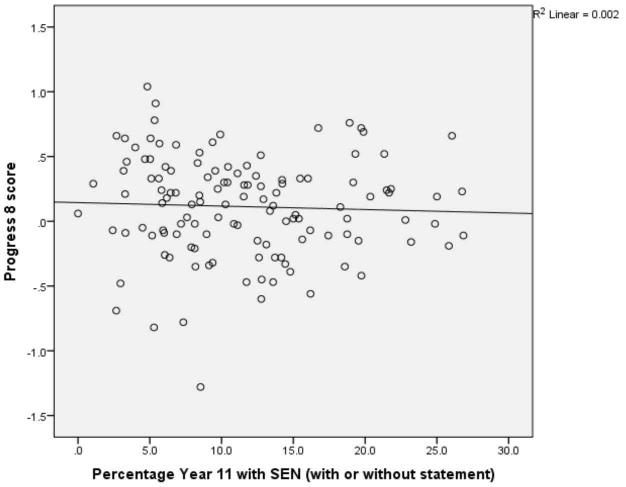


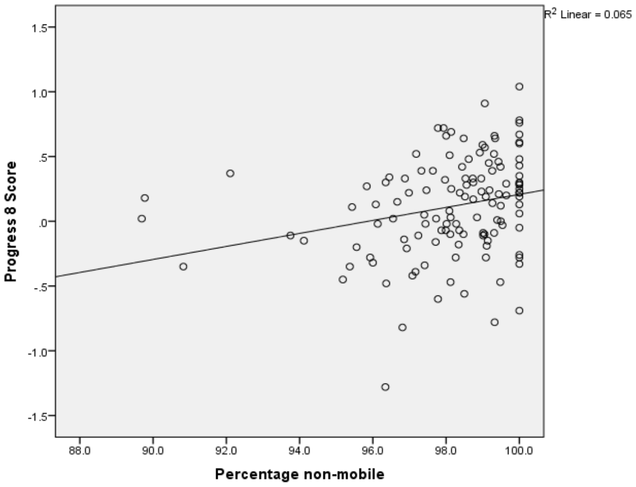




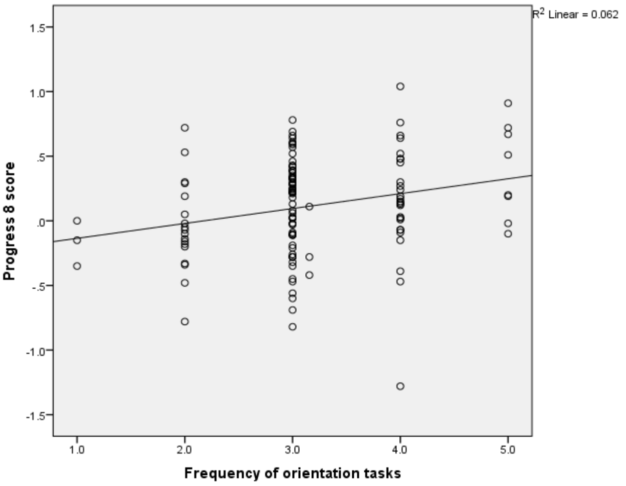


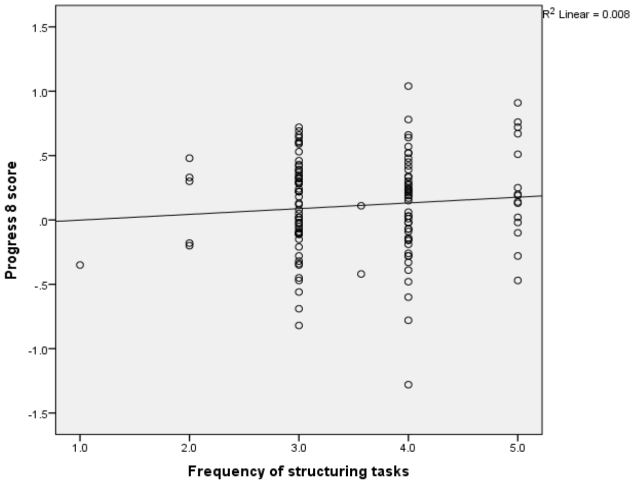


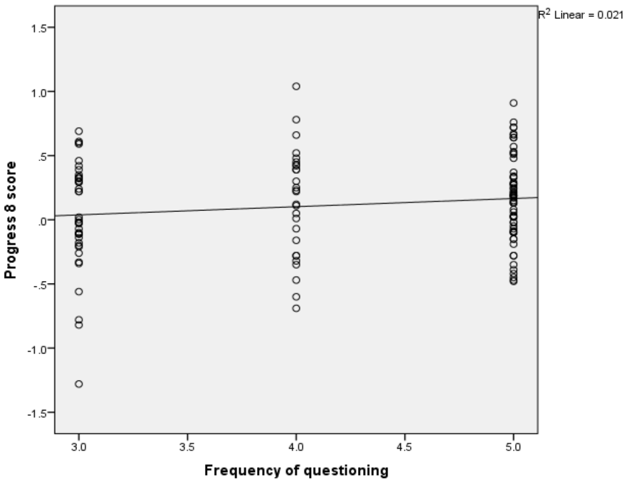


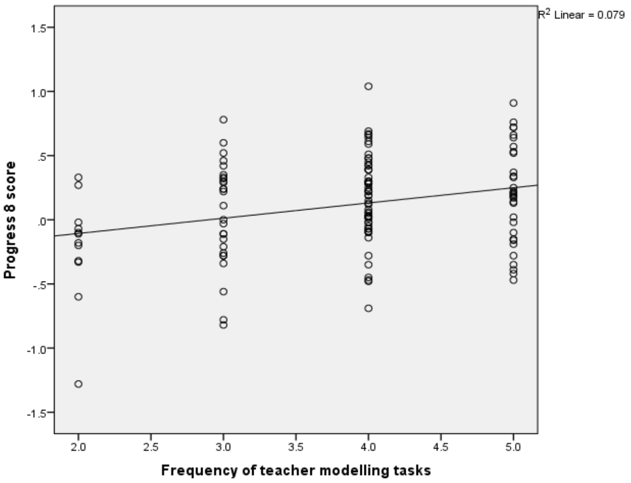


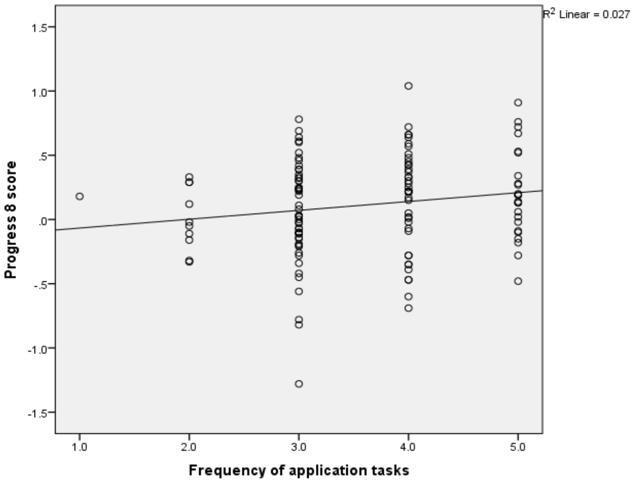
Instructional behaviours:

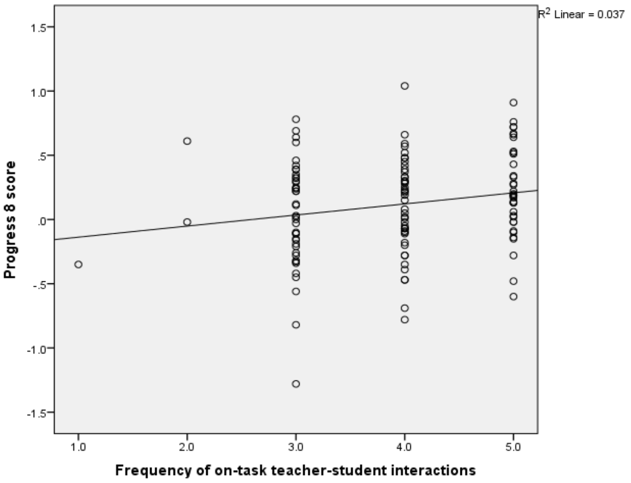


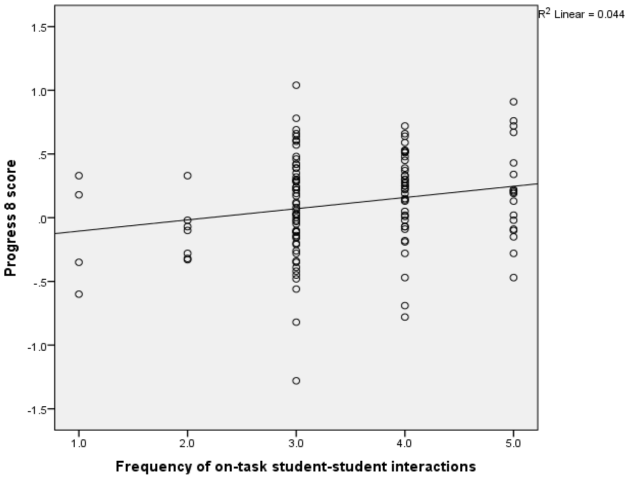


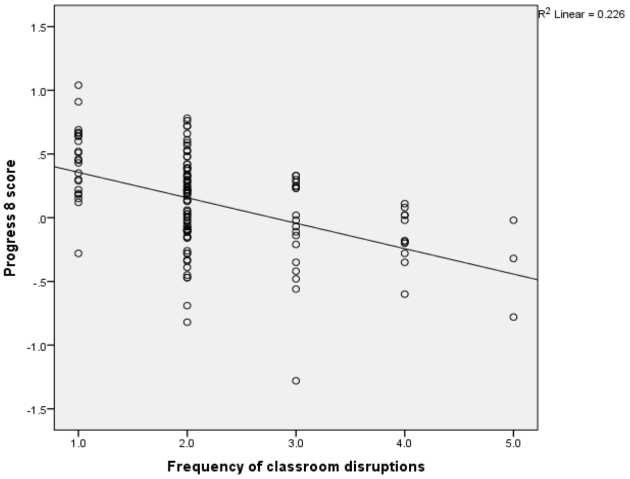

R^2 Linear = 0.021
Progress 8 score
1.5
1.0
.5
.0
-.5
-1.0
-1.5
3.0
3.5
4.0
4.5
5.0
Frequency of questioning

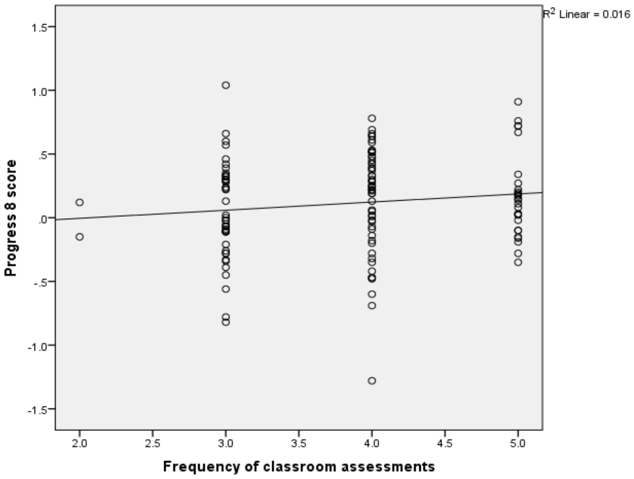

R^2 Linear = 0.079
Progress 8 score
1.5
1.0
.5
.0
-.5
-1.0
-1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Frequency of teacher modelling tasks


R^2 Linear = 0.027
Progress 8 score
1.5
1.0
.5
.0
-.5
-1.0
-1.5
1.0
2.0
3.0
4.0
5.0
Frequency of application tasks

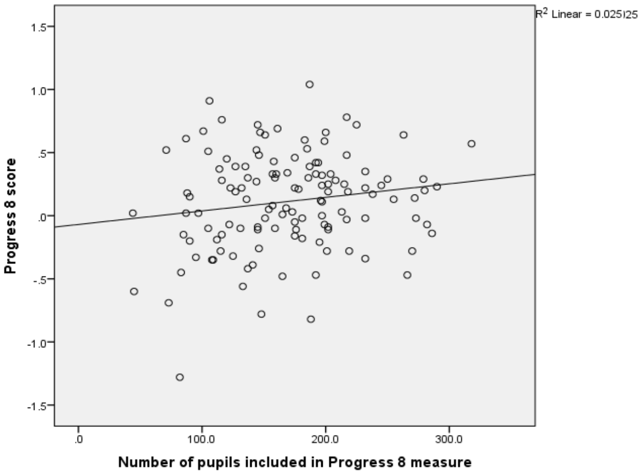

R^2 Linear = 0.037
Progress 8 score
1.5
1.0
.5
.0
-.5
-1.0
-1.5
1.0
2.0
3.0
4.0
5.0
Frequency of on-task teacher-student interactions

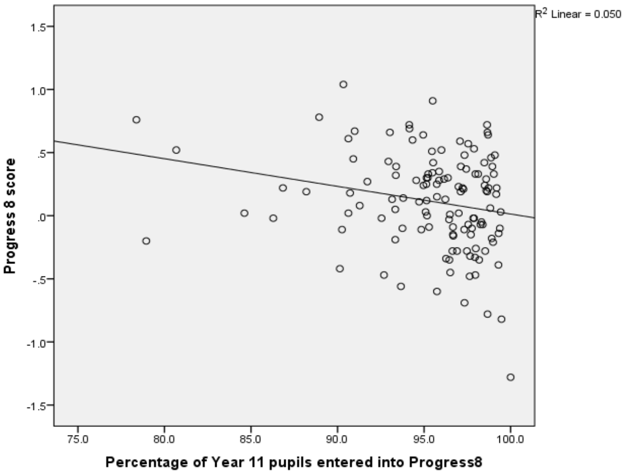

R^2 Linear = 0.044
Progress 8 score
1.5
1.0
.5
.0
-.5
-1.0
-1.5
1.0
2.0
3.0
4.0
5.0
Frequency of on-task student-student interactions

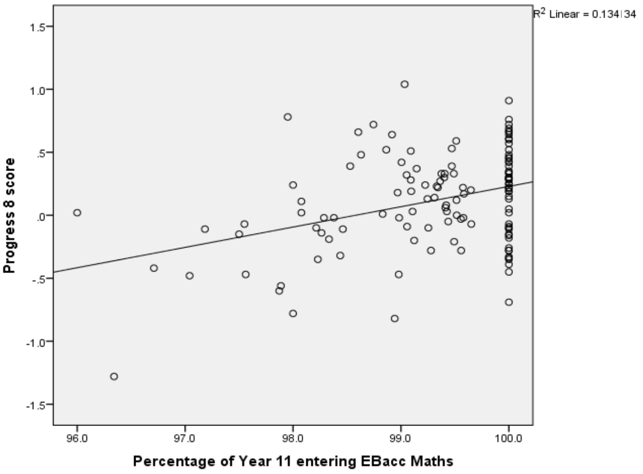

R^2 Linear = 0.226
Progress 8 score
1.5
1.0
.5
.0
-.5
-1.0
-1.5
1.0
2.0
3.0
4.0
5.0
Frequency of classroom disruptions

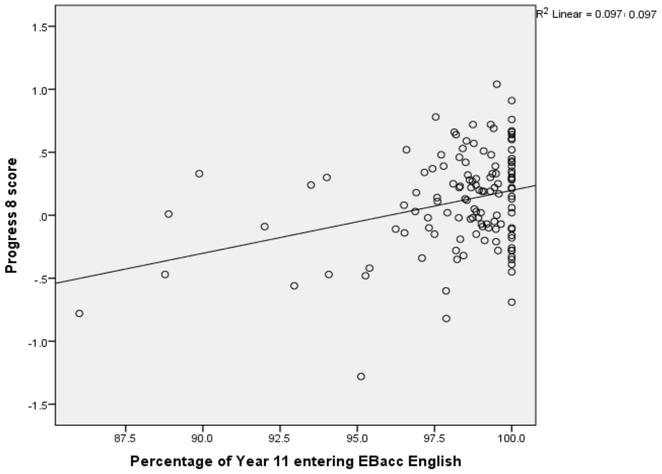

R^2 Linear = 0.016
Progress 8 score
1.5
1.0
.5
.0
-.5
-1.0
-1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Frequency of classroom assessments

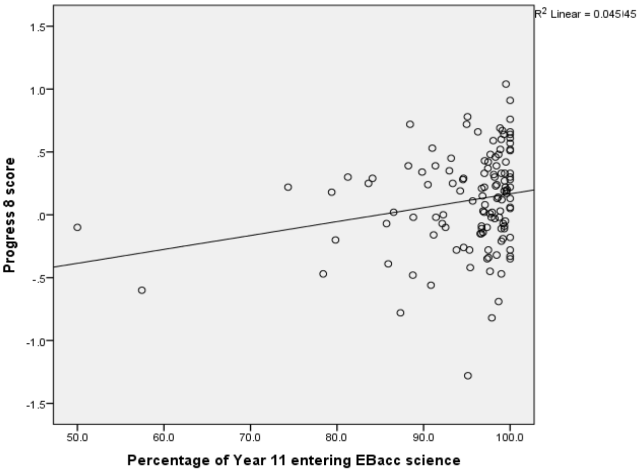
Examination entry variables:

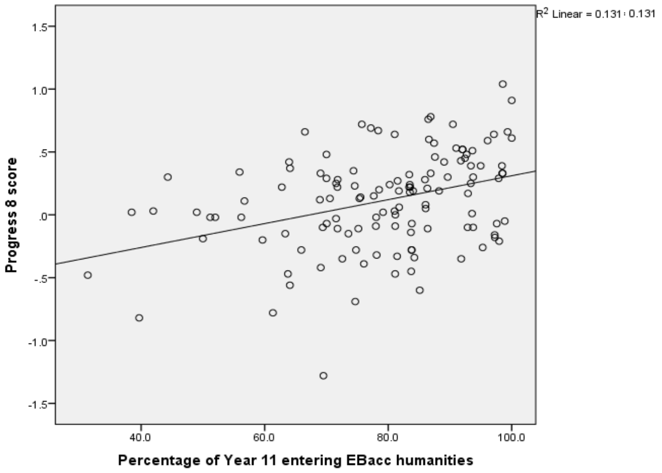


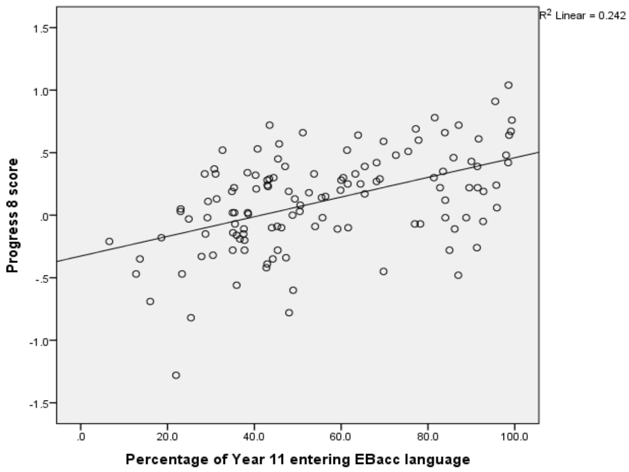


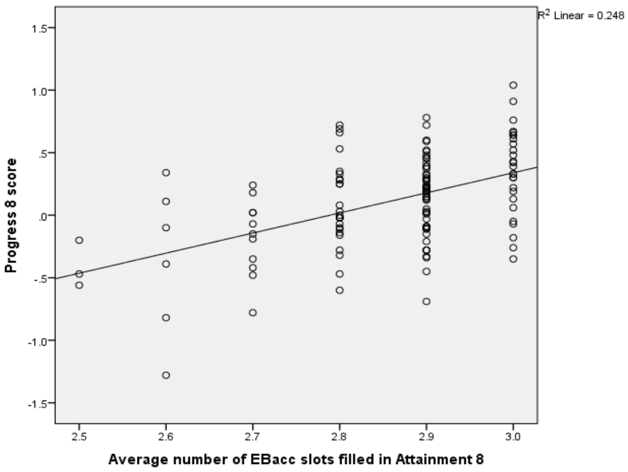


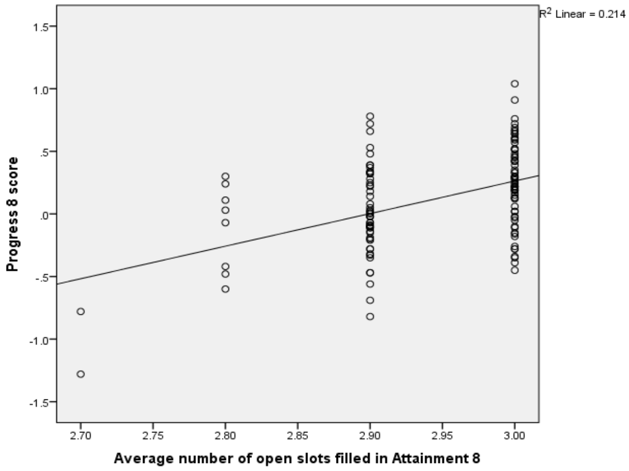


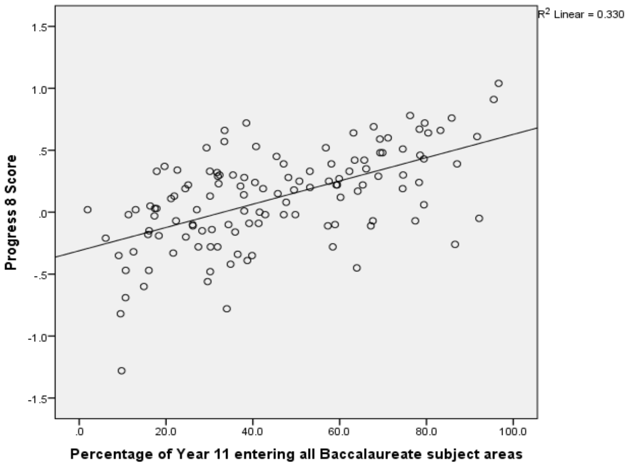


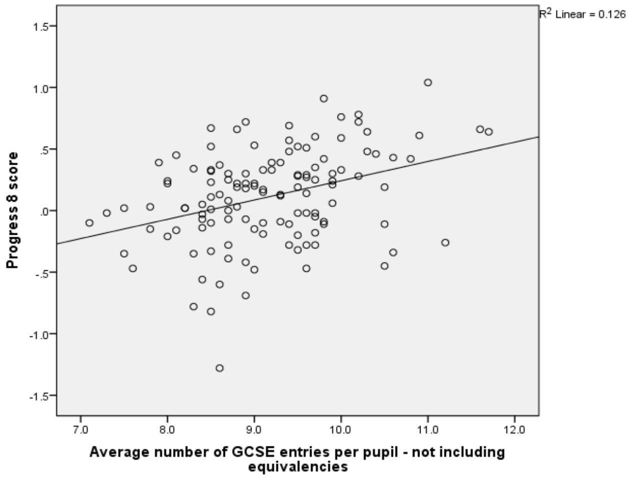


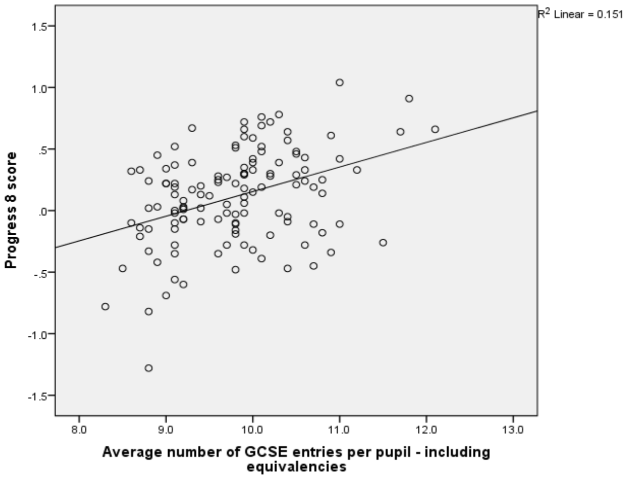

R^2 Linear = 0.242
Progress 8 score
1.5
1.0
.5
.0
-.5
-1.0
-1.5
.0
20.0
40.0
60.0
80.0
100.0
Percentage of Year 11 entering EBacc language


R^2 Linear = 0.248
Progress 8 score
1.5
1.0
.5
.0
-.5
-1.0
-1.5
2.5
2.6
2.7
2.8
2.9
3.0
Average number of EBacc slots filled in Attainment 8


R^2 Linear = 0.214
Progress 8 score
1.5
1.0
.5
.0
-.5
-1.0
-1.5
2.70
2.75
2.80
2.85
2.90
2.95
3.00
Average number of open slots filled in Attainment 8

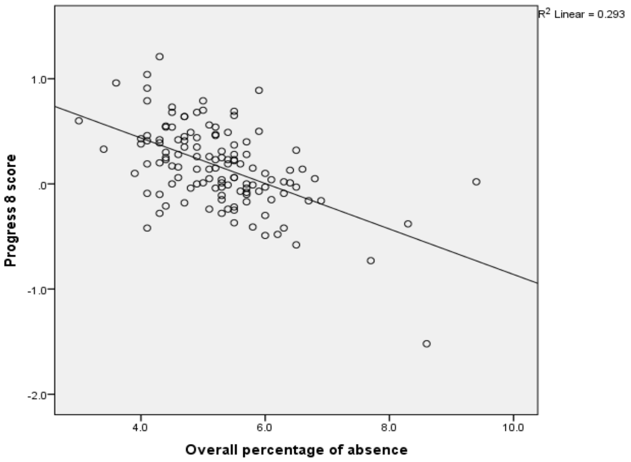

R^2 Linear = 0.330
Progress 8 Score
1.5
1.0
.5
.0
-.5
-1.0
-1.5
.0
20.0
40.0
60.0
80.0
100.0
Percentage of Year 11 entering all Baccalaureate subject areas

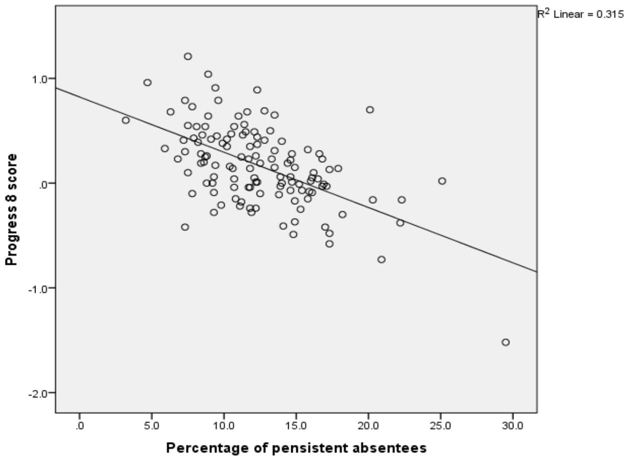

R^2 Linear = 0.126
Progress 8 score
1.5
1.0
.5
.0
-.5
-1.0
-1.5
7.0
8.0
9.0
10.0
11.0
12.0
Average number of GCSE entries per pupil - not including equivalencies

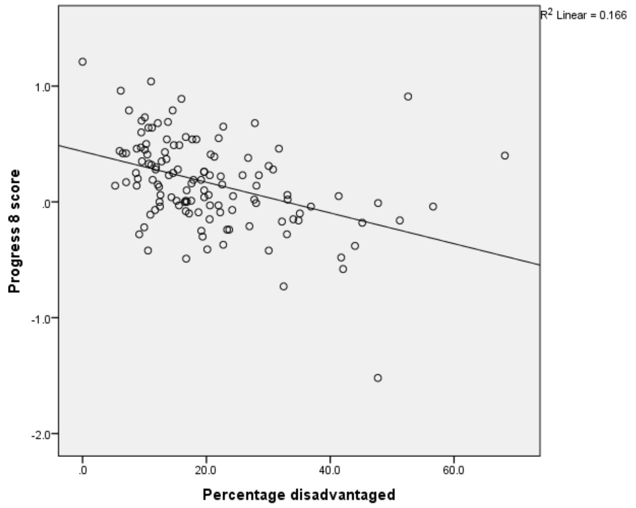

R^2 Linear = 0.151
Progress 8 score
1.5
1.0
.5
.0
-.5
-1.0
-1.5
8.0
9.0
10.0
11.0
12.0
13.0
Average number of GCSE entries per pupil - including equivalencies

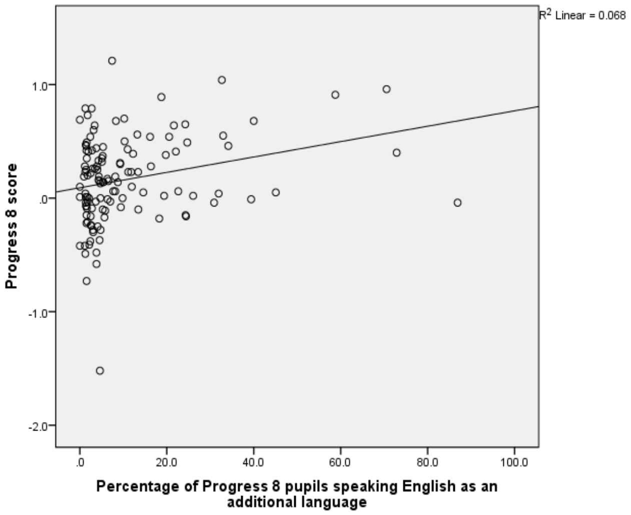
2018 Analysis:

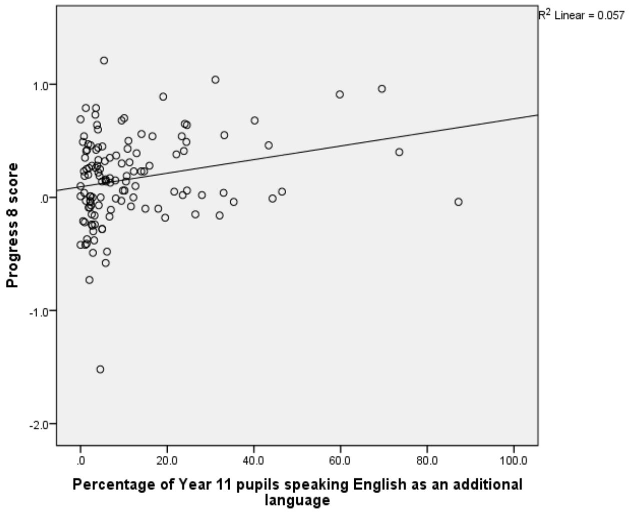
Intake variables:

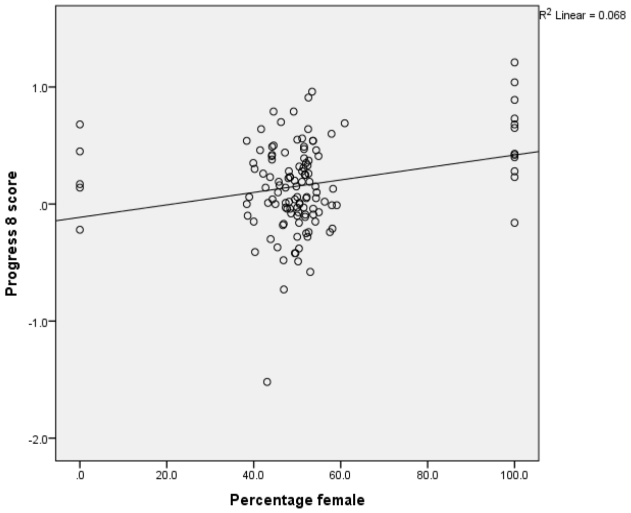


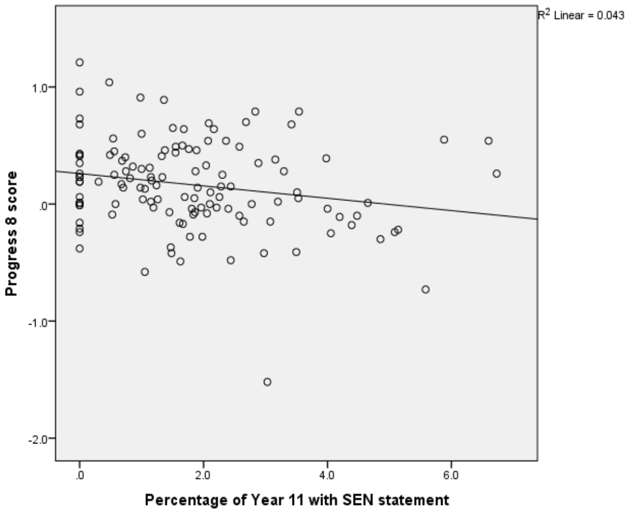


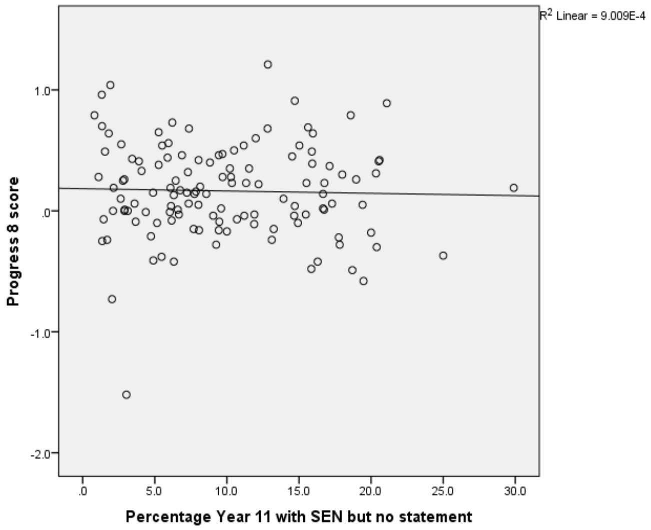


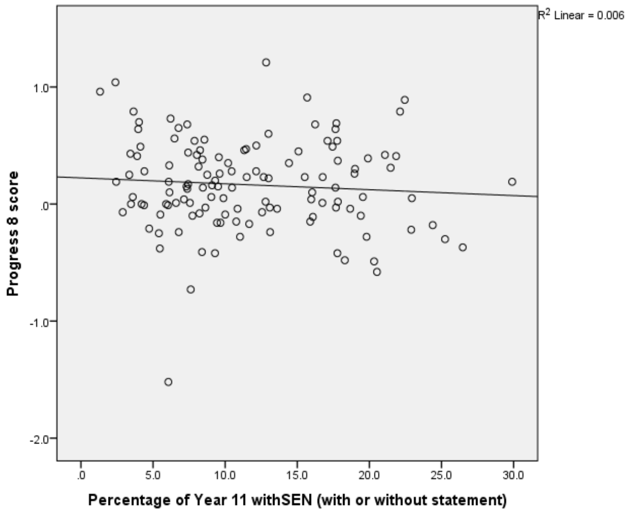


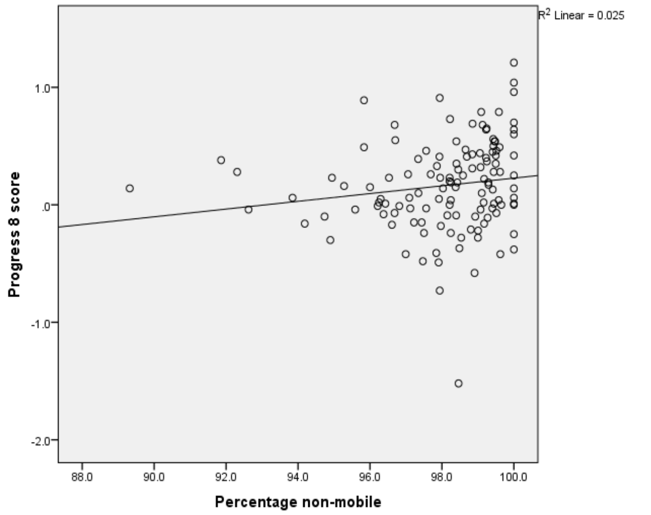




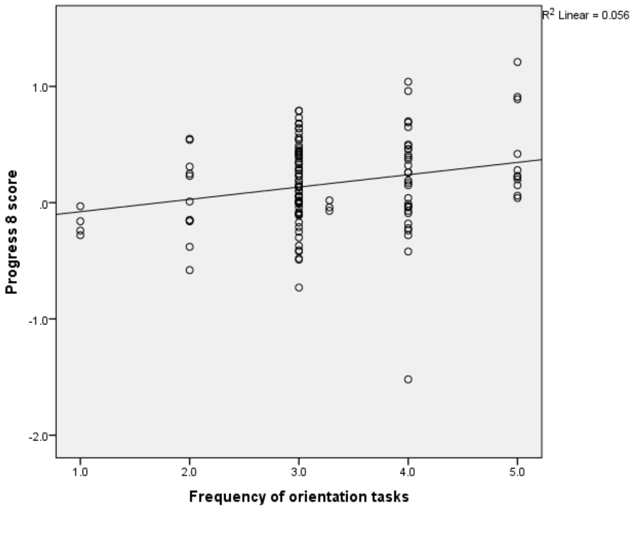


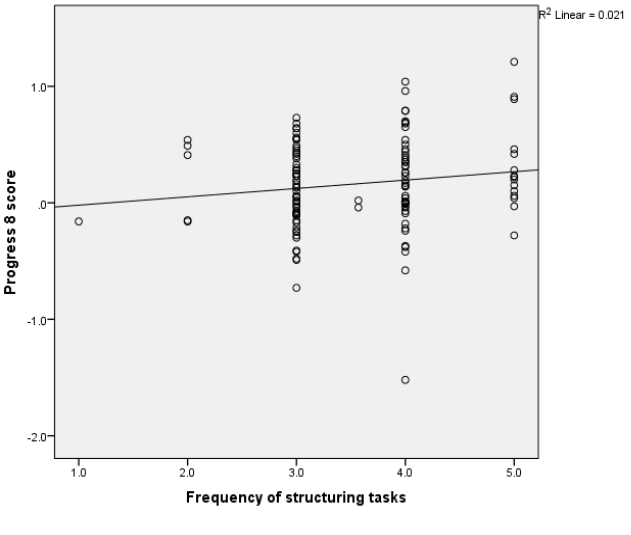


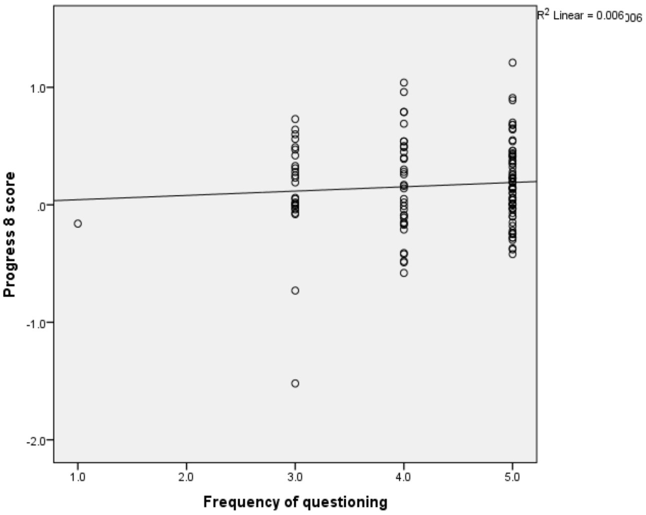


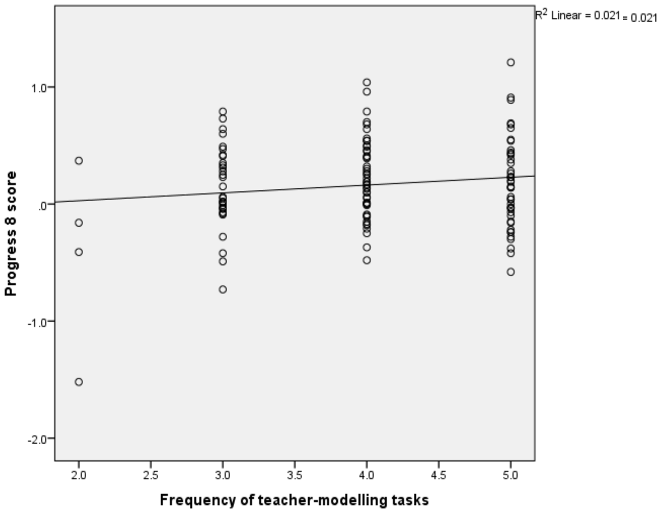


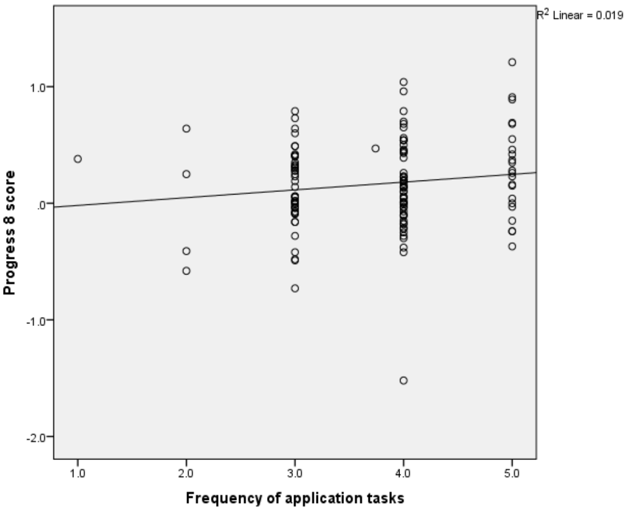
Instructional variables:

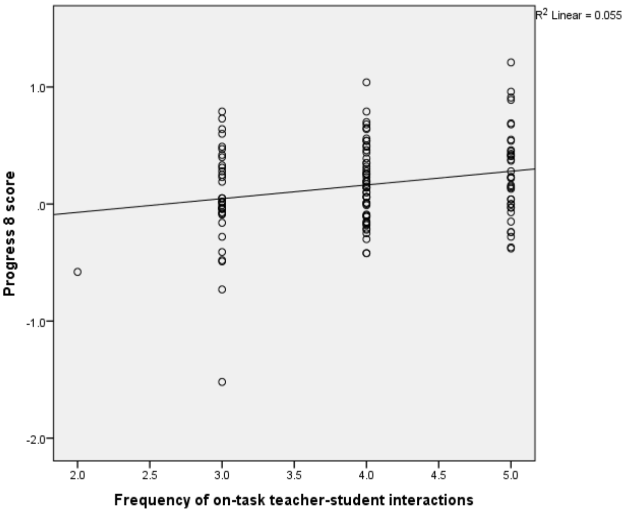


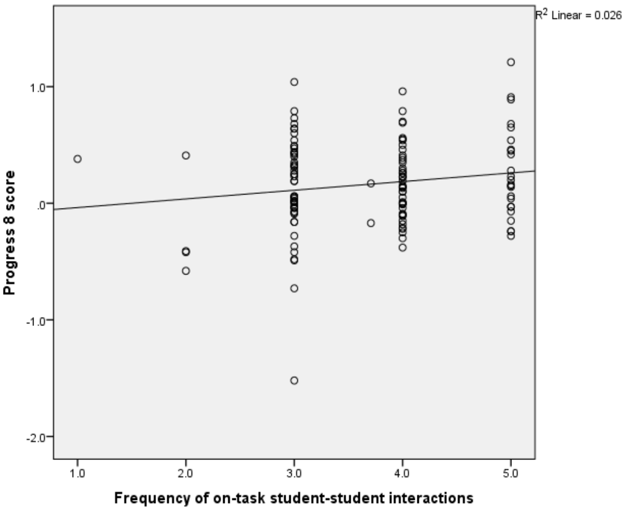


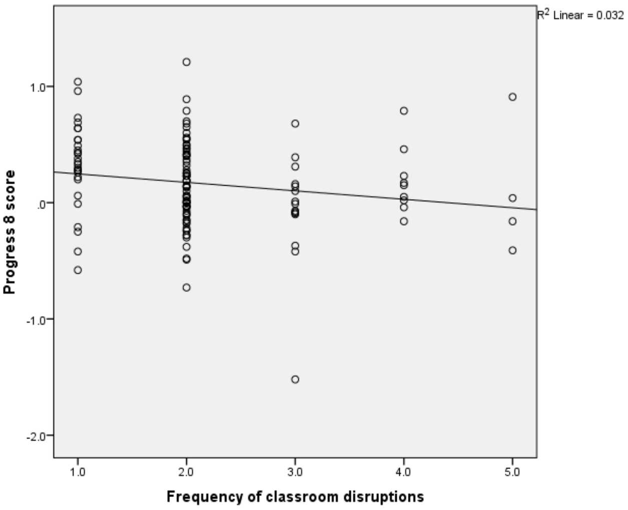

R^2 Linear = 0.006
Progress 8 score
1.0
.0
-1.0
-2.0
1.0
2.0
3.0
4.0
5.0
Frequency of questioning

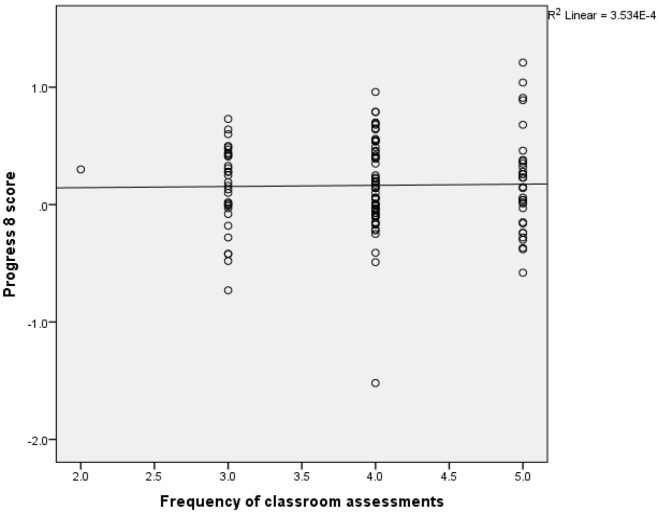

R^2 Linear = 0.021
Progress 8 score
1.0
.0
-1.0
-2.0
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Frequency of teacher-modelling tasks


R^2 Linear = 0.019
Progress 8 score
1.0
.0
-1.0
-2.0
1.0
2.0
3.0
4.0
5.0
Frequency of application tasks

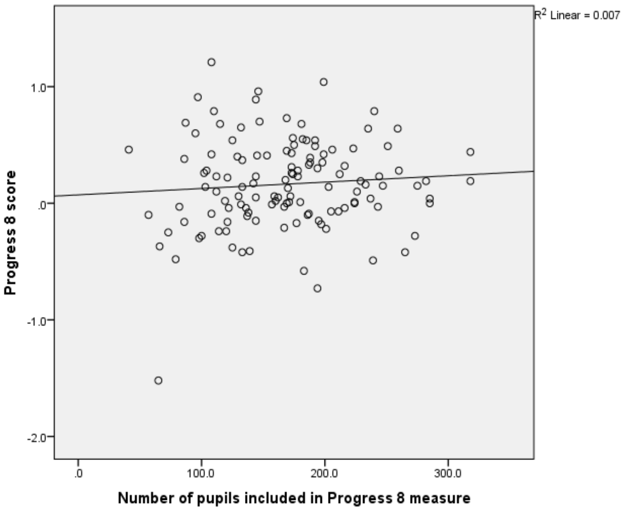

R^2 Linear = 0.055
Progress 8 score
1.0
.0
-1.0
-2.0
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Frequency of on-task teacher-student interactions

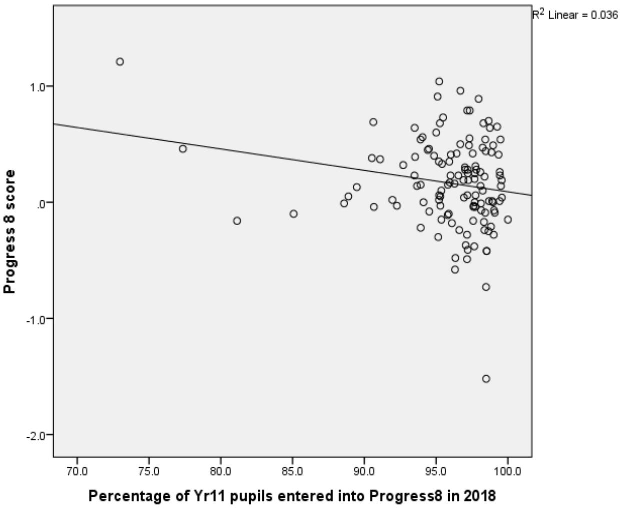

R^2 Linear = 0.026
Progress 8 score
1.0
.0
-1.0
-2.0
1.0
2.0
3.0
4.0
5.0
Frequency of on-task student-student interactions

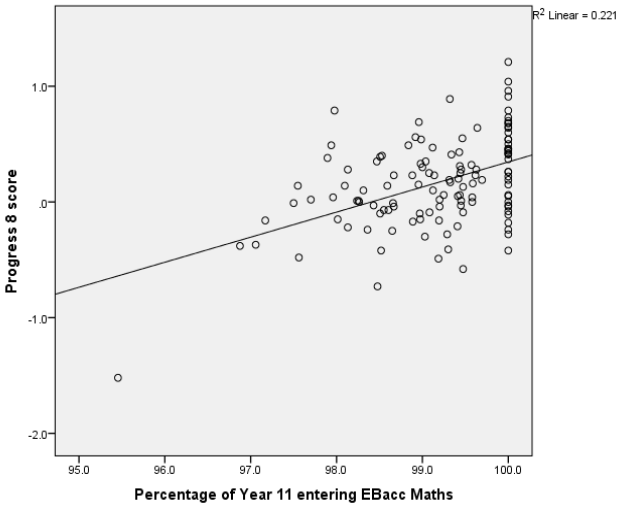

R^2 Linear = 0.032
Progress 8 score
1.0
.0
-1.0
-2.0
1.0
2.0
3.0
4.0
5.0
Frequency of classroom disruptions

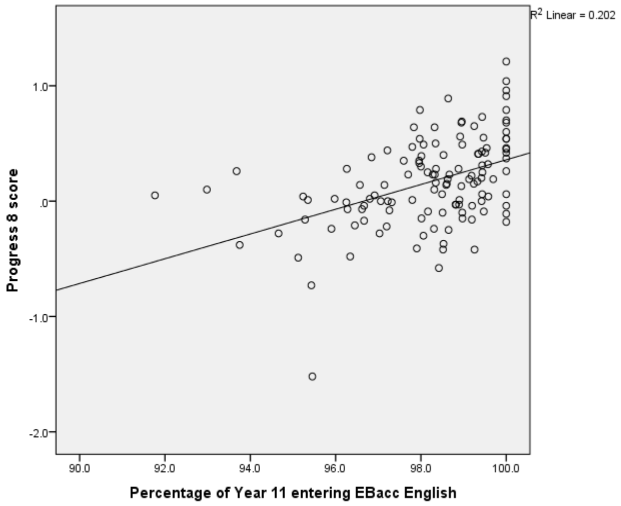

R^2 Linear = 3.534E-4
Progress 8 score
1.0
.0
-1.0
-2.0
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Frequency of classroom assessments

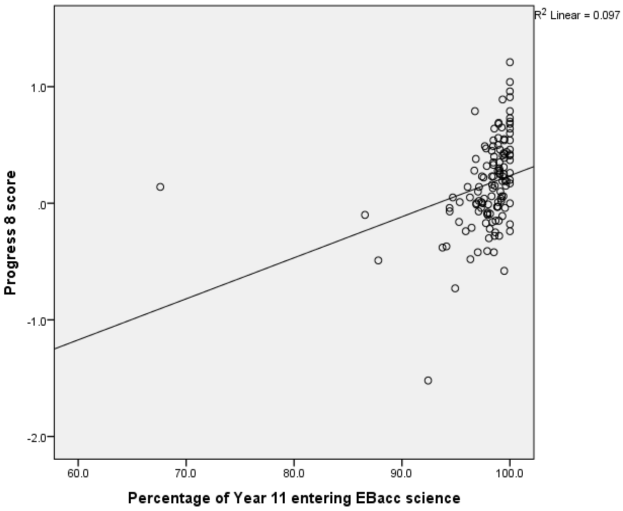
Examination entry variables:

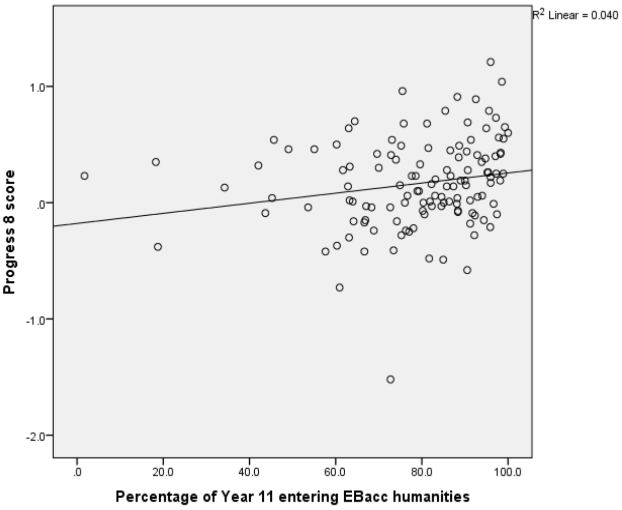


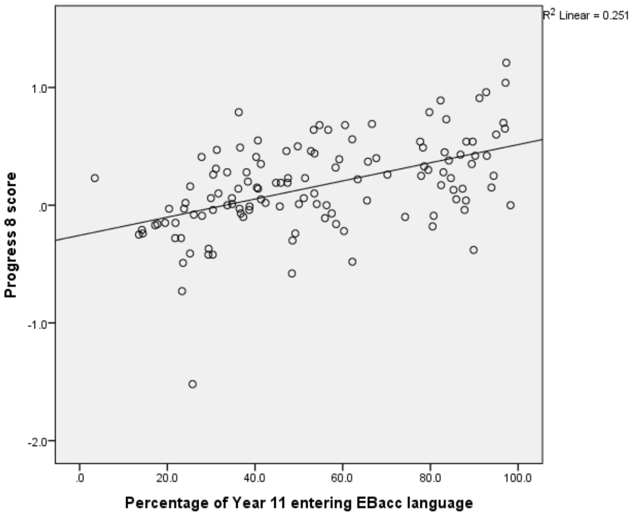


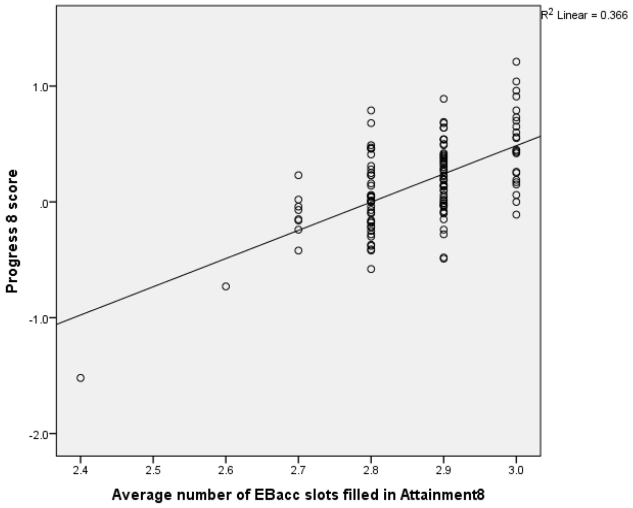


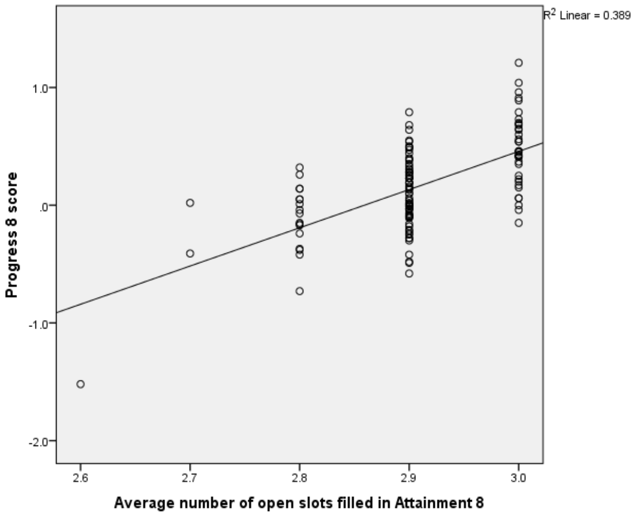


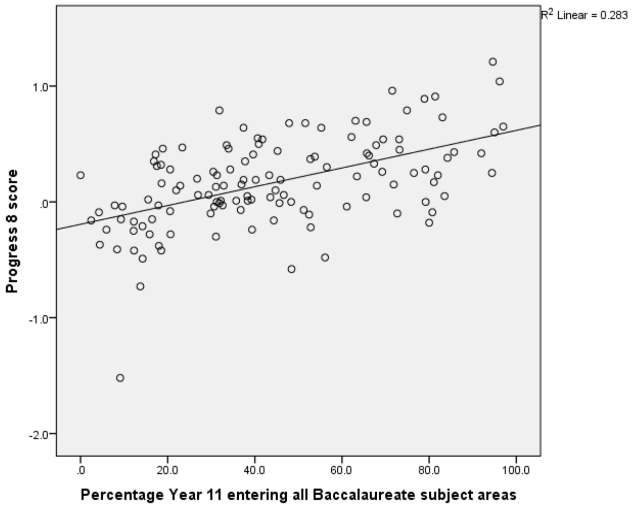


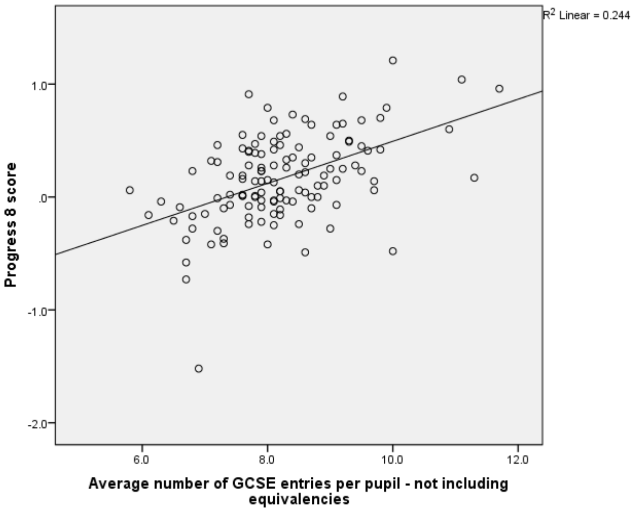


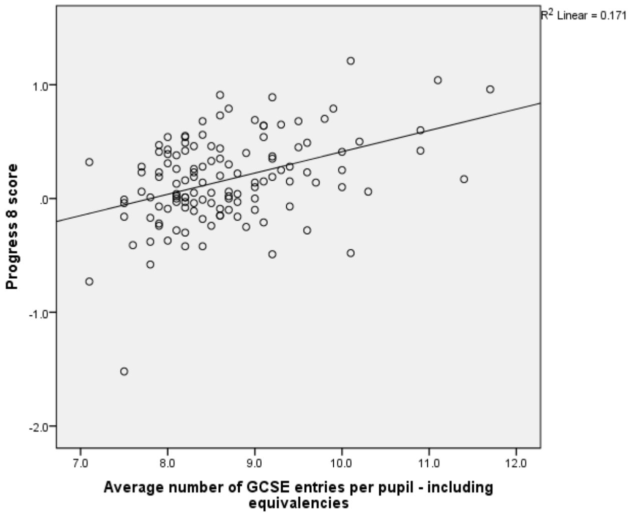

R2 Linear = 0.251
Progress 8 score
1.0
.0
-1.0
-2.0
.0
20.0
40.0
60.0
80.0
100.0
Percentage of Year 11 entering EBacc language


R2 Linear = 0.366
Progress 8 score
1.0
.0
-1.0
-2.0
2.4
2.5
2.6
2.7
2.8
2.9
3.0
Average number of EBacc slots filled in Attainment8


R2 Linear = 0.389
Progress 8 score
1.0
.0
-1.0
-2.0
2.6
2.7
2.8
2.9
3.0
Average number of open slots filled in Attainment 8

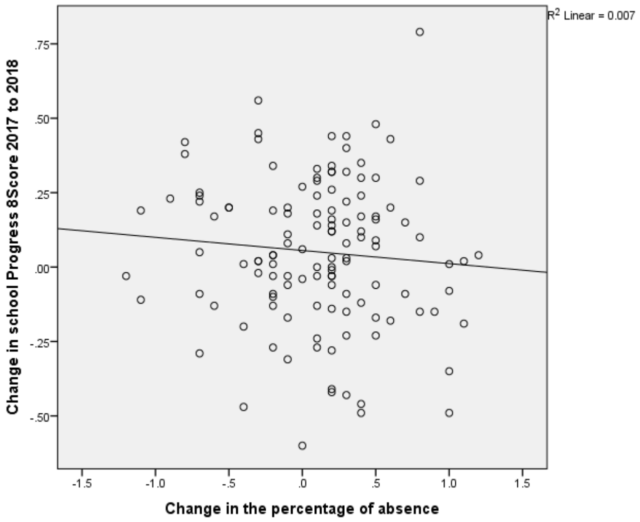

R2 Linear = 0.283
Progress 8 score
1.0
.0
-1.0
-2.0
.0
20.0
40.0
60.0
80.0
100.0
Percentage Year 11 entering all Baccalaureate subject areas

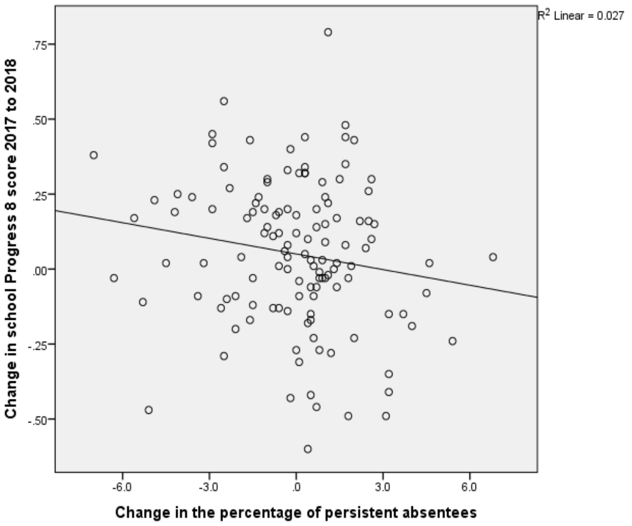

R2 Linear = 0.244
Progress 8 score
1.0
.0
-1.0
-2.0
6.0
8.0
10.0
12.0
Average number of GCSE entries per pupil - not including equivalencies

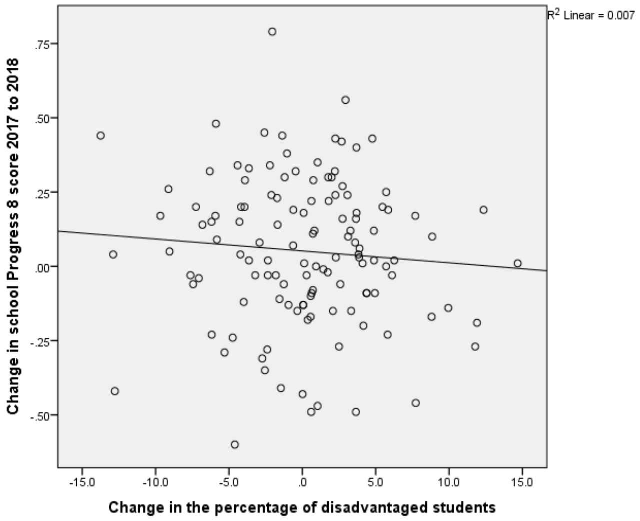

R2 Linear = 0.171
Progress 8 score
1.0
.0
-1.0
-2.0
7.0
8.0
9.0
10.0
11.0
12.0
Average number of GCSE entries per pupil - including equivalencies

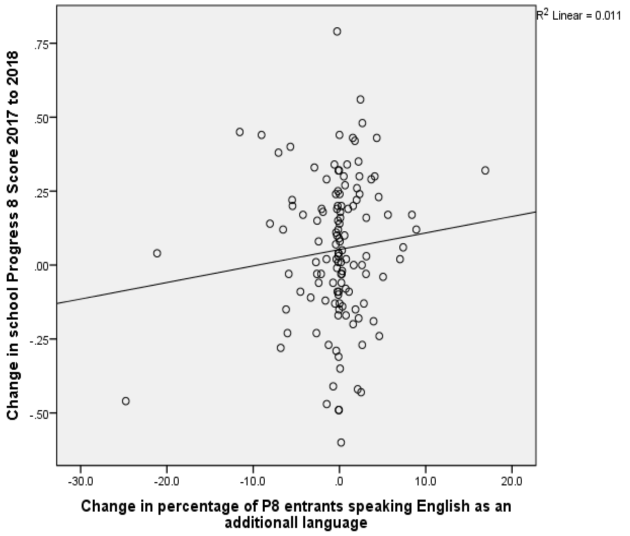
Change Analysis

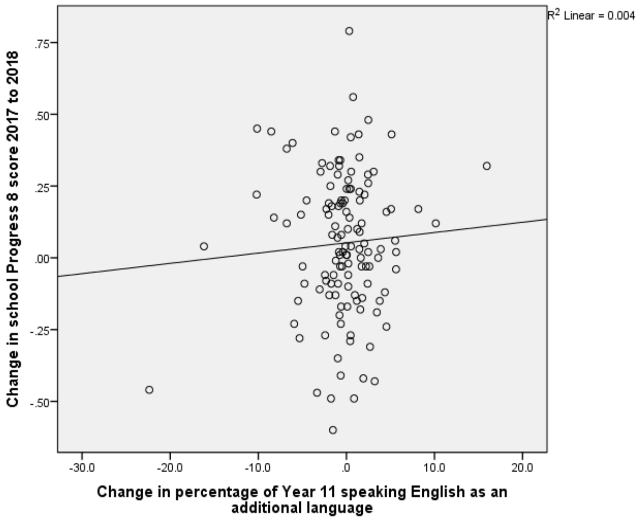
Intake variables:

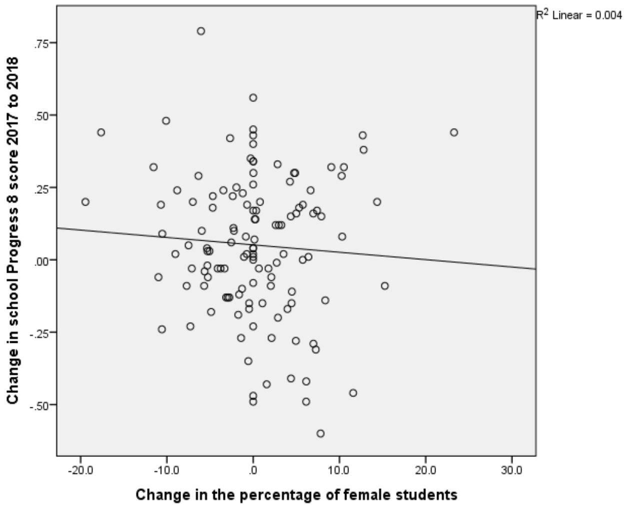


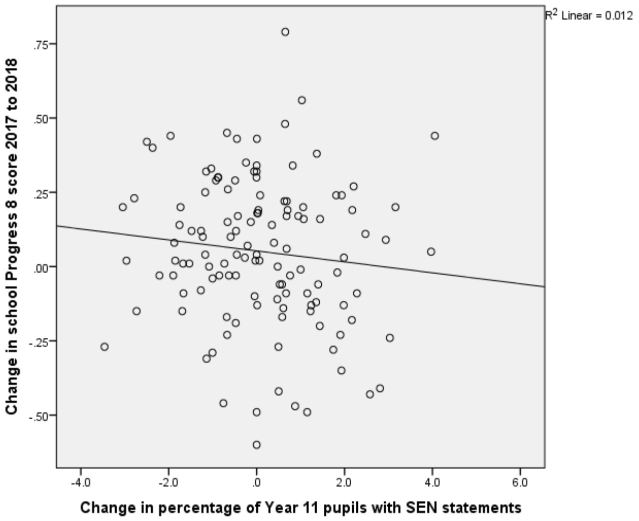


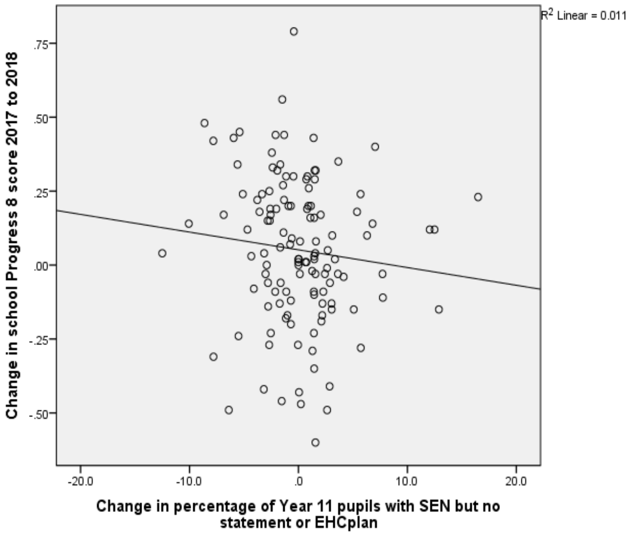


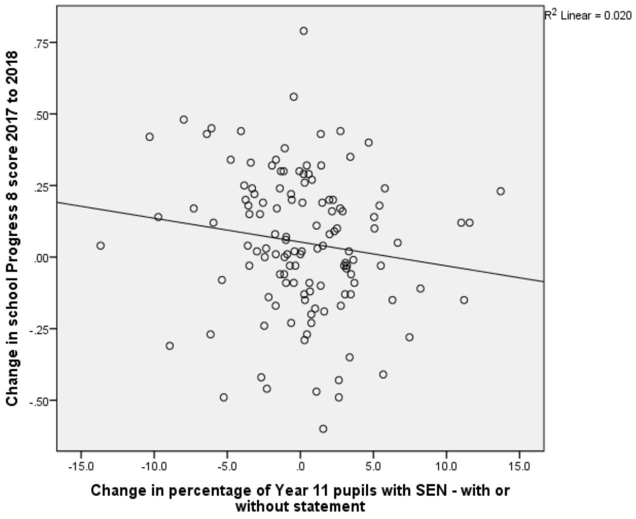


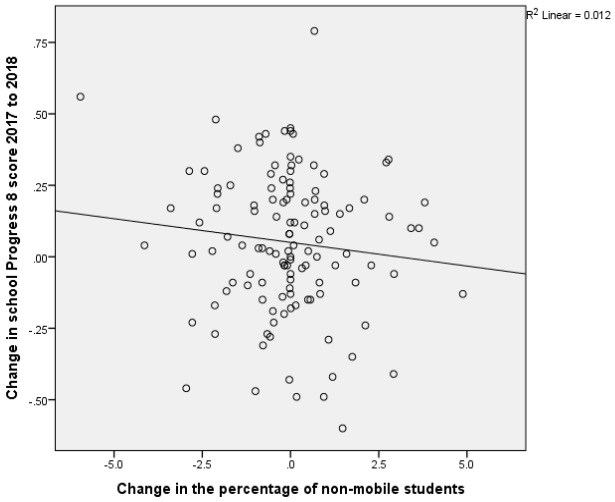




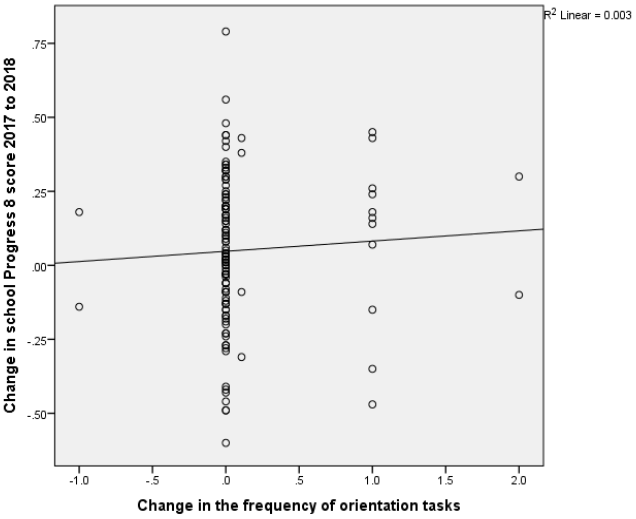

R^2 Linear = 0.012
Change in school Progress 8 score 2017 to 2018
.75
.50
.25
.00
-.25
-.50
-4.0
-2.0
.0
2.0
4.0
6.0
Change in percentage of Year 11 pupils with SEN statements

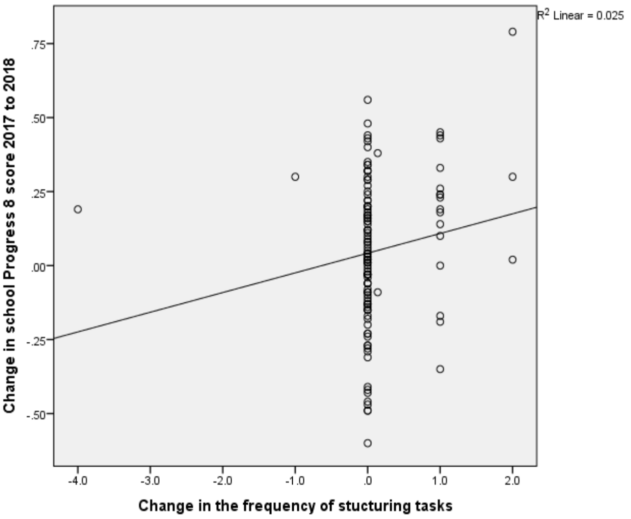

R^2 Linear = 0.011
Change in school Progress 8 score 2017 to 2018
.75
.50
.25
.00
-.25
-.50
-20.0
-10.0
.0
10.0
20.0
Change in percentage of Year 11 pupils with SEN but no statement or EHCplan

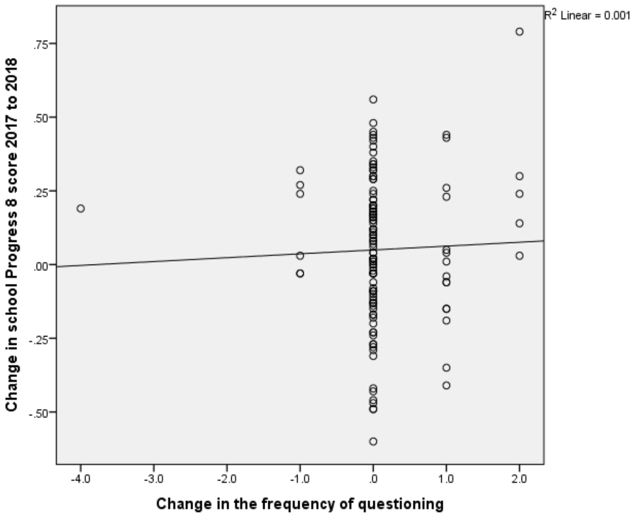

R^2 Linear = 0.020
Change in school Progress 8 score 2017 to 2018
.75
.50
.25
.00
-.25
-.50
-15.0
-10.0
-5.0
.0
5.0
10.0
15.0
Change in percentage of Year 11 pupils with SEN - with or without statement

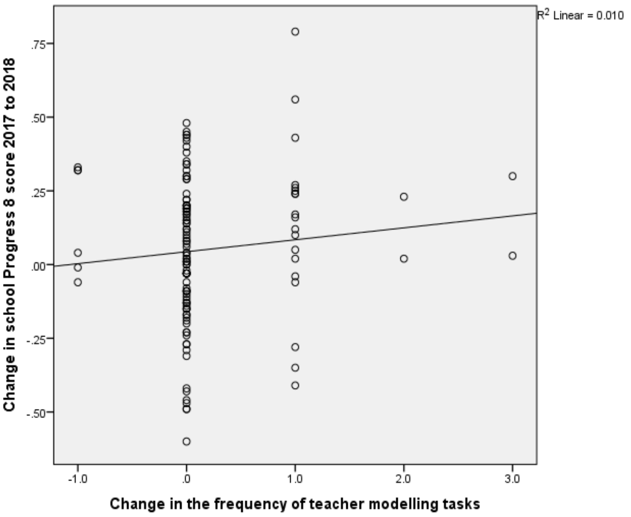

R^2 Linear = 0.012
Change in school Progress 8 score 2017 to 2018
.75
.50
.25
.00
-.25
-.50
-5.0
-2.5
.0
2.5
5.0
Change in the percentage of non-mobile students

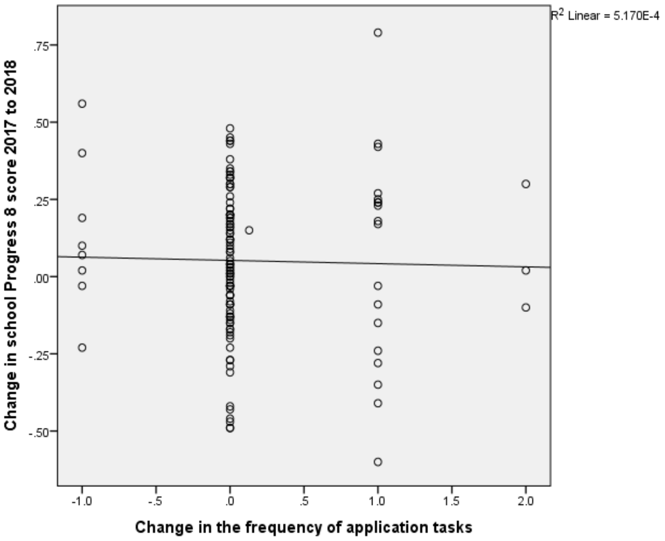
Instructional practices:

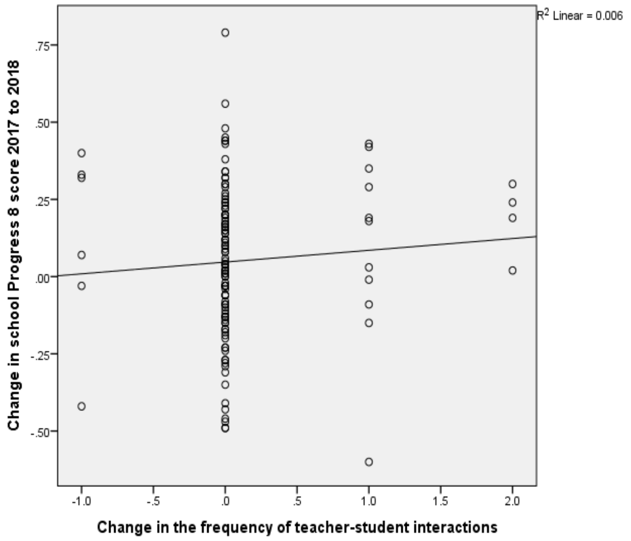


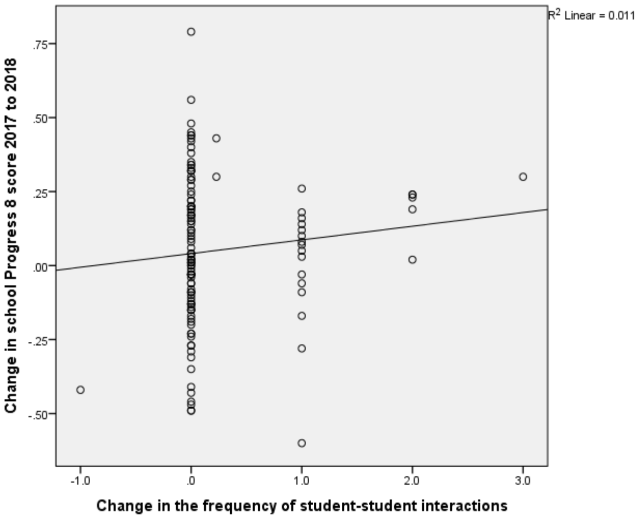


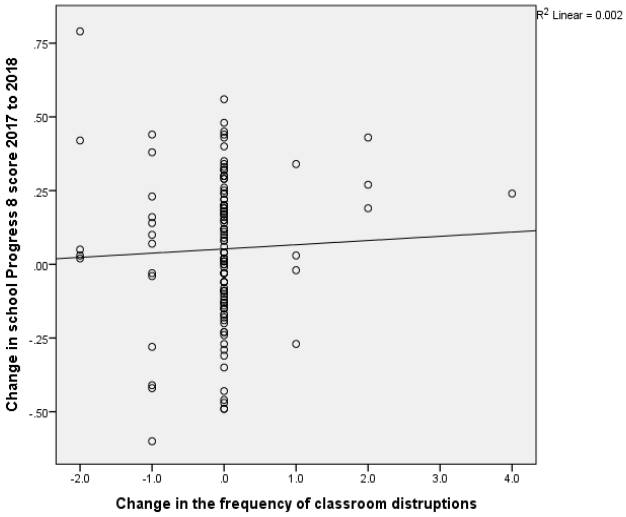


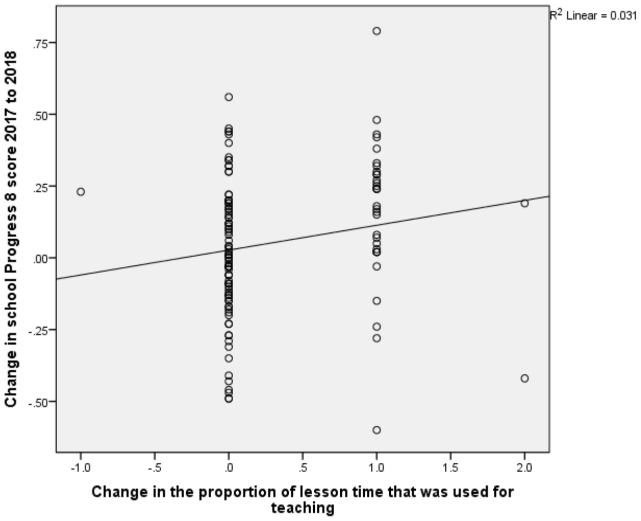


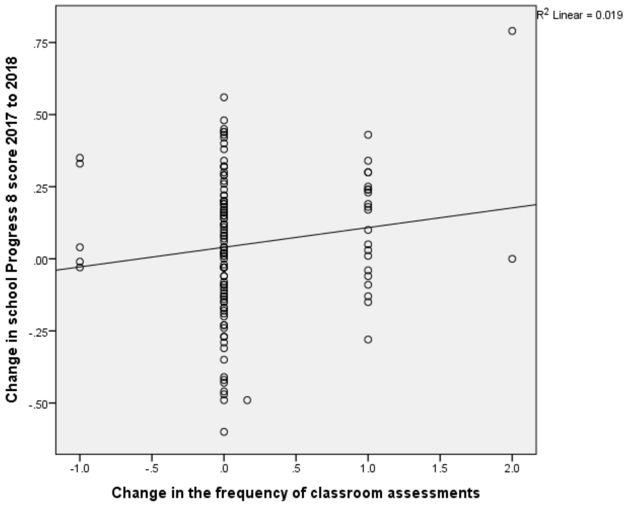


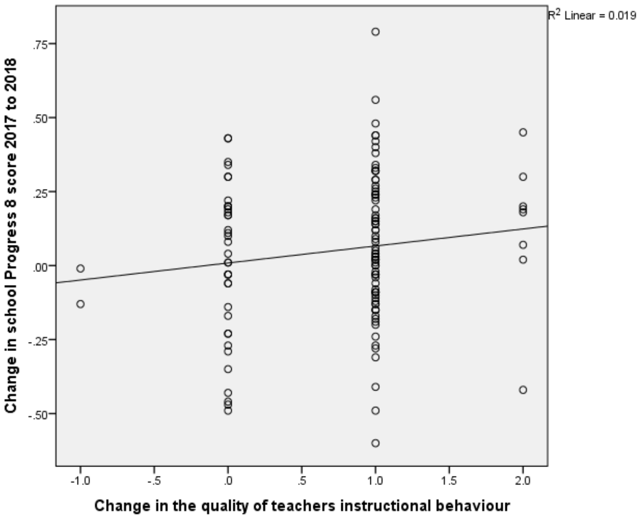


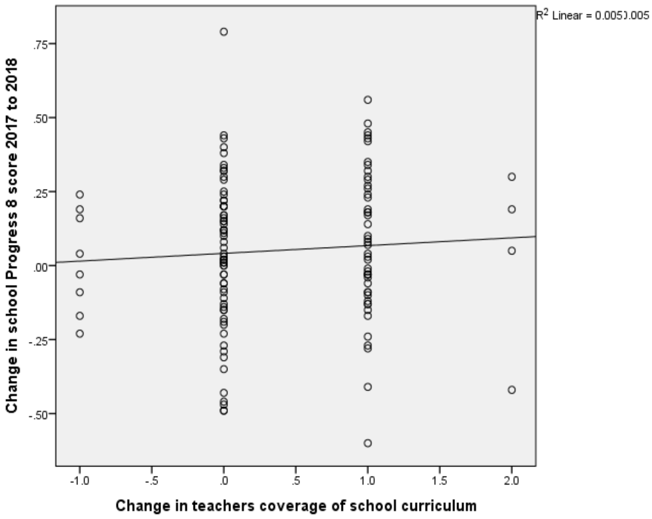

R2 Linear = 0.011
Change in school Progress 8 score 2017 to 2018
.75
.50
.25
.00
-.25
-.50
-1.0
.0
1.0
2.0
3.0
Change in the frequency of student-student interactions


R2 Linear = 0.002
Change in school Progress 8 score 2017 to 2018
.75
.50
.25
.00
-.25
-.50
-2.0
-1.0
.0
1.0
2.0
3.0
4.0
Change in the frequency of classroom distruptions

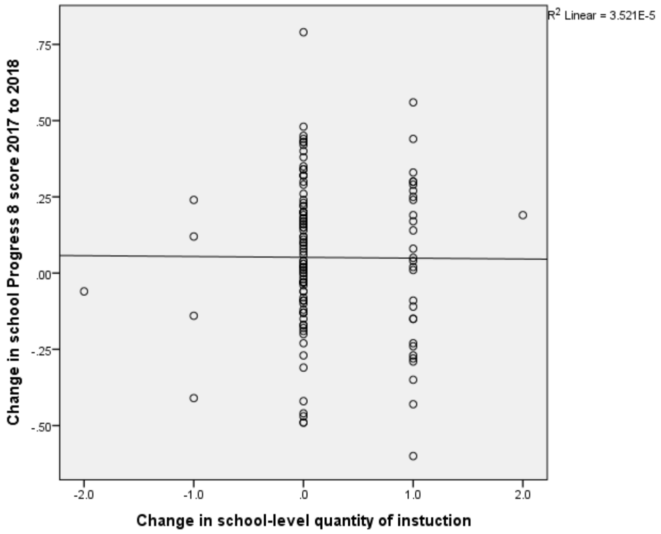

R2 Linear = 0.031
Change in school Progress 8 score 2017 to 2018
.75
.50
.25
.00
-.25
-.50
-1.0
-.5
.0
.5
1.0
1.5
2.0
Change in the proportion of lesson time that was used for teaching

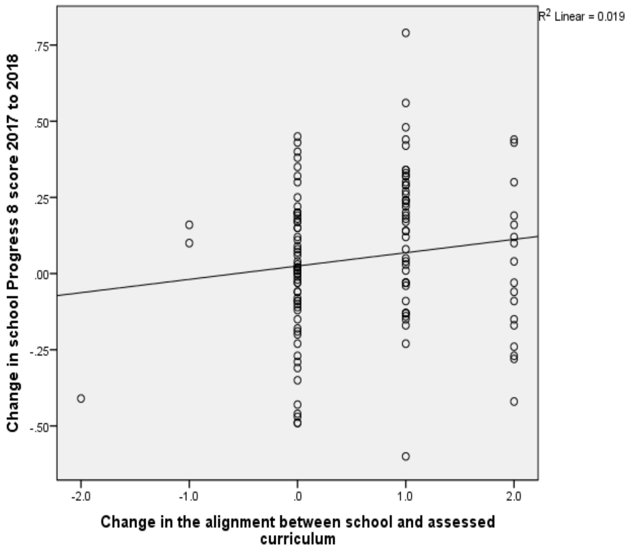

R2 Linear = 0.019
Change in school Progress 8 score 2017 to 2018
.75
.50
.25
.00
-.25
-.50
-1.0
-.5
.0
.5
1.0
1.5
2.0
Change in the frequency of classroom assessments

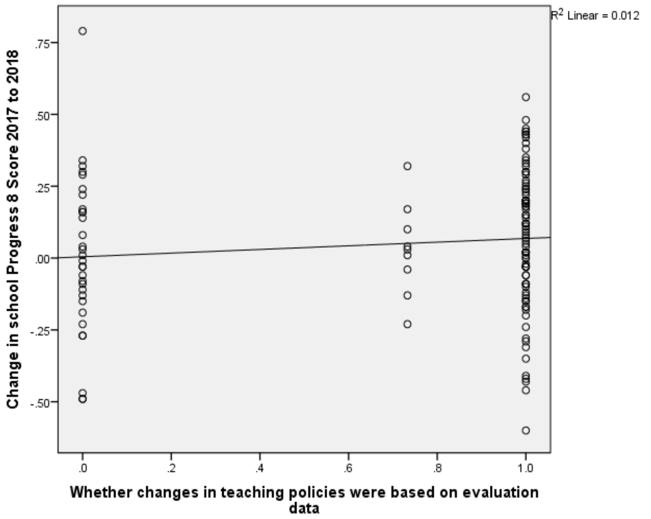

R2 Linear = 0.019
Change in school Progress 8 score 2017 to 2018
.75
.50
.25
.00
-.25
-.50
-1.0
-.5
.0
.5
1.0
1.5
2.0
Change in the quality of teachers instructional behaviour

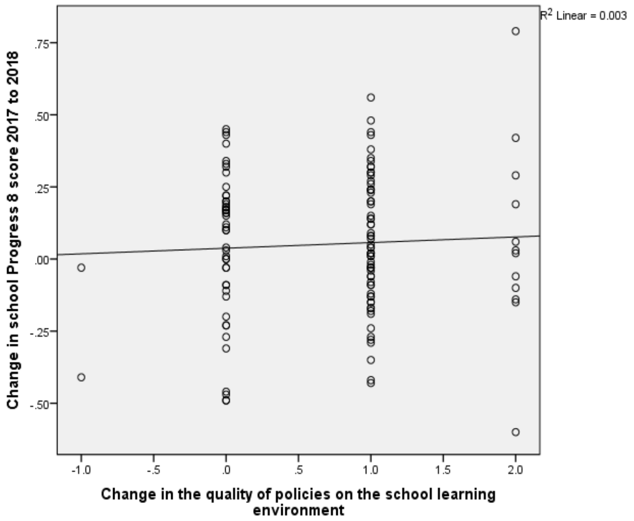

R2 Linear = 0.005
Change in school Progress 8 score 2017 to 2018
.75
.50
.25
.00
-.25
-.50
-1.0
-.5
.0
.5
1.0
1.5
2.0
Change in teachers coverage of school curriculum

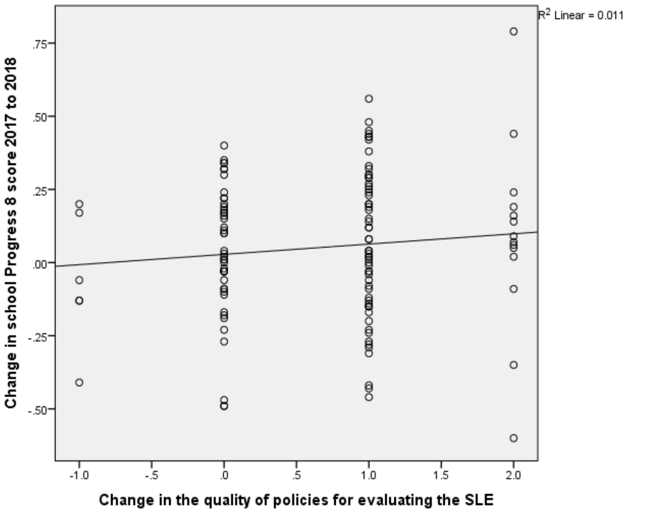
School policies:

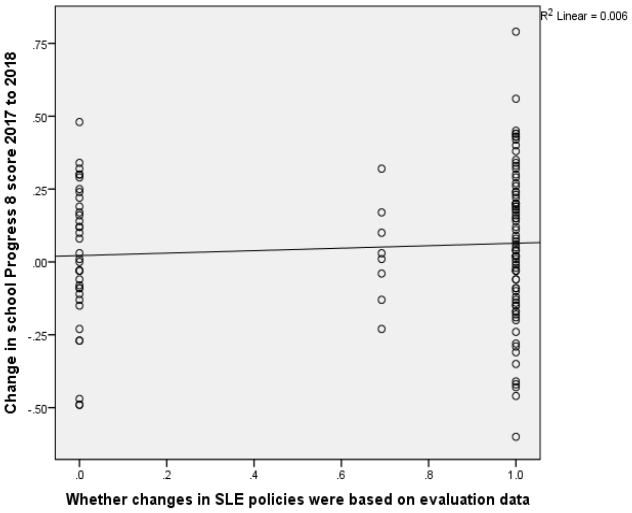




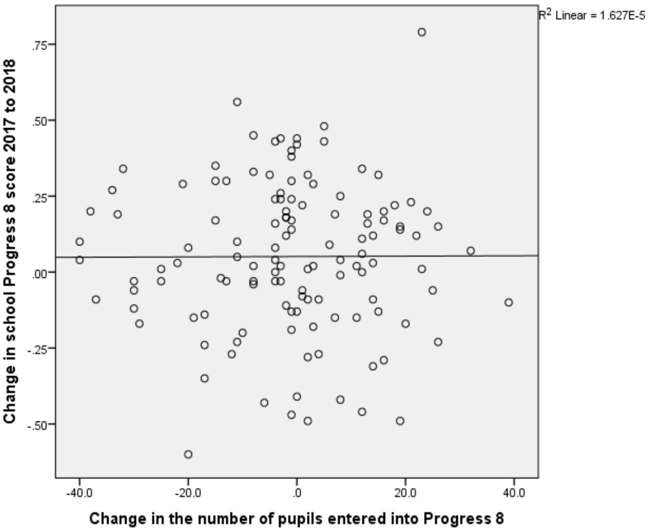


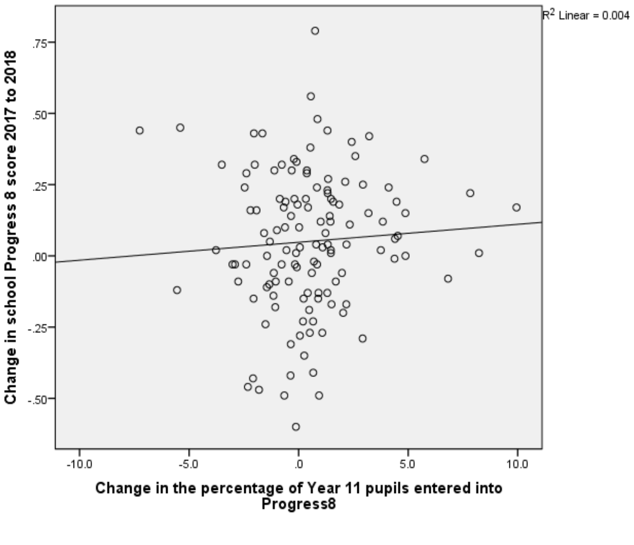


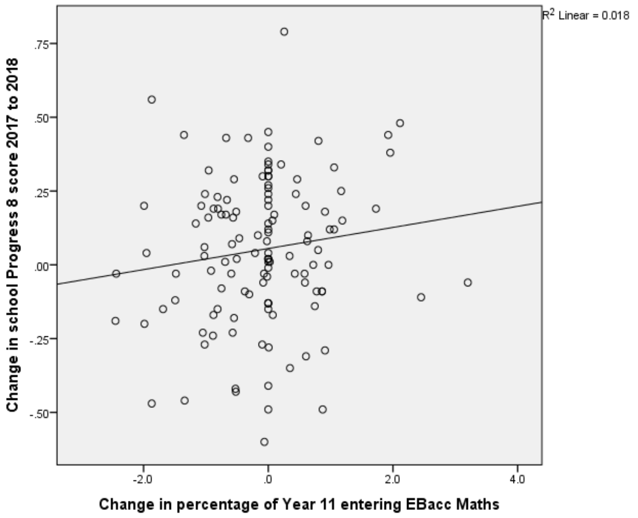


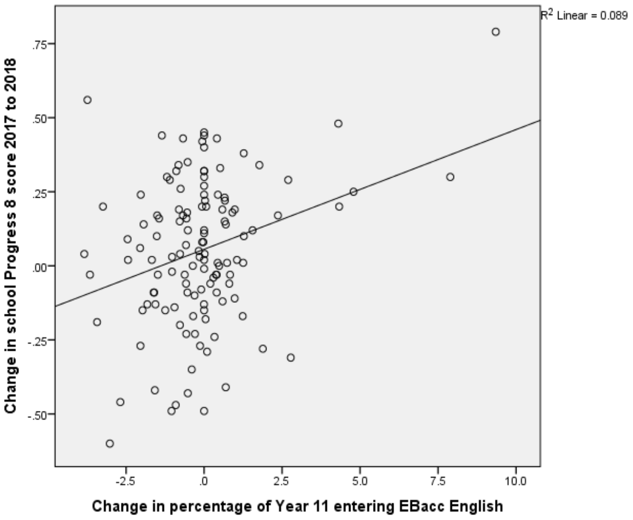


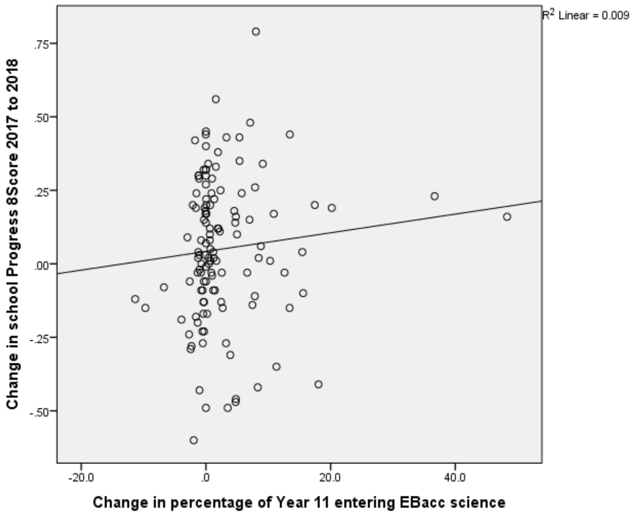
Examination entry variables:

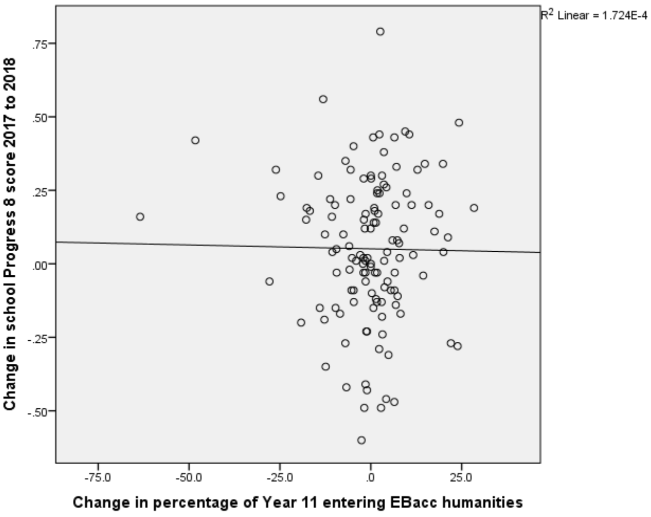


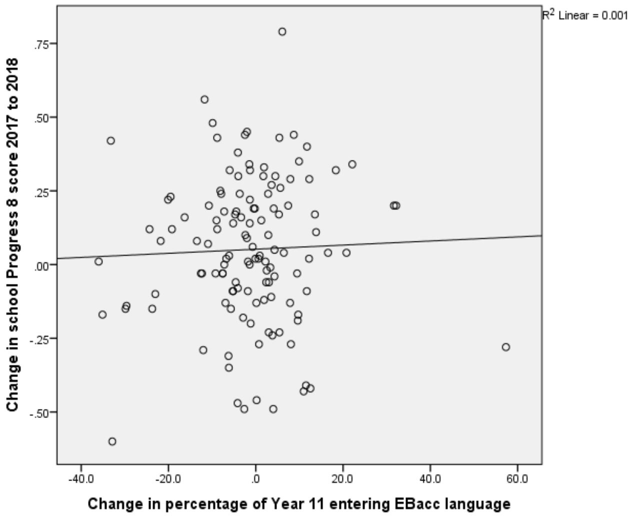


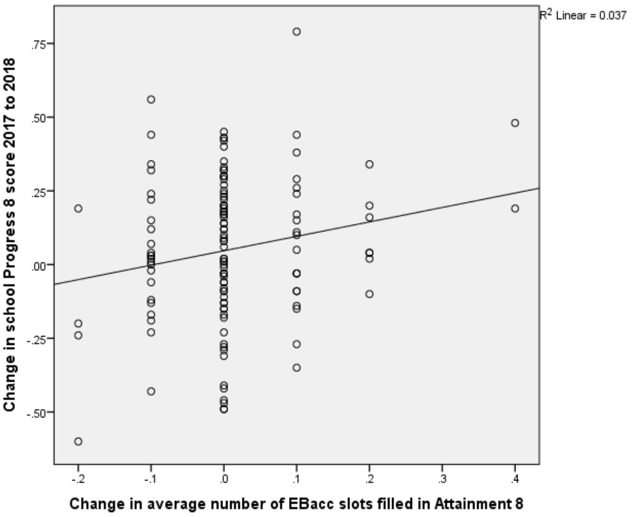

R2 Linear = 0.018
Change in school Progress 8 score 2017 to 2018
.75
.50
.25
.00
-.25
-.50
-2.0
.0
2.0
4.0
Change in percentage of Year 11 entering EBacc Maths


R2 Linear = 0.089
Change in school Progress 8 score 2017 to 2018
.75
.50
.25
.00
-.25
-.50
-2.5
.0
2.5
5.0
7.5
10.0
Change in percentage of Year 11 entering EBacc English


R2 Linear = 0.009
Change in school Progress 8Score 2017 to 2018
.75
.50
.25
.00
-.25
-.50
-20.0
.0
20.0
40.0
Change in percentage of Year 11 entering EBacc science


R2 Linear = 1.724E-4
Change in school Progress 8 score 2017 to 2018
.75
.50
.25
.00
-.25
-.50
-75.0
-50.0
-25.0
.0
25.0
Change in percentage of Year 11 entering EBacc humanities


R2 Linear = 0.001
Change in school Progress 8 score 2017 to 2018
.75
.50
.25
.00
-.25
-.50
-40.0
-20.0
.0
20.0
40.0
60.0
Change in percentage of Year 11 entering EBacc language


R2 Linear = 0.037
Change in school Progress 8 score 2017 to 2018
.75
.50
.25
.00
-.25
-.50
-.2
-.1
.0
.1
.2
.3
.4
Change in average number of EBacc slots filled in Attainment 8

Appendix D: Long Questionnaire

A study of national school improvement approaches

A Durham University study

Introduction

Thank you for taking part in our study of national school improvement approaches. This questionnaire will ask you to describe any changes that have occurred within your school since last academic year. We would like to emphasize that our research is focused on the government's policy for evaluating secondary school performance and not the performance of individual schools. Any data disclosed will be kept confidential and stored securely until it is destroyed. Schools and individuals will not be identifiable in the write up. The questionnaire will take approximately 15 minutes to complete and should be completed in one go (google forms does not allow you to save your responses).

Section A: Self-evaluation of your school's educational provisions

Please estimate the value-added score that your school will receive in 2018*

In the space below please predict what your school's value added score will be in 2018. This is often referred to as the Key Stage 2 to Key Stage 4 Progress 8 score. You should provide an exact number, not a range of values. When devising this estimate it may help to consider the score that your school received last academic year and any changes that have occurred within your school since that time. Possible changes include variations in the quantity of instruction students receive, the school curriculum, the quality of instruction, and the school learning environment. Factors that influence these variables such as the school's funding, changes to the student intake, teaching staff or leadership, the exam entry procedures and teaching to the test may also be relevant. You may take into account factors that we have not mentioned.

....

Section B: School Policies
Areas addressed by the school policies:

Below are 4 tables. Please indicate whether your school had a policy associated with the topics in column 1, in place, during the time periods identified in columns 2 (last year) and 3 (this year). Do this by ticking the appropriate boxes. Select any that apply. For the purpose of this questionnaire, a 'policy' includes any guidelines which help to make the school's approach more concrete to staff and students. A paragraph explaining the rules on attendance, within the school behaviour policy, would therefo re still count as a policy on attendance. School-organized actions/interventions that were intended to improve performance in the discussed areas should also be viewed as part of your school policies.

Policies that govern the quantity of instruction students receive

	Areas covered by last year's school policies	Areas covered by your current policies
Lesson schedules and school timetables	□	□
The protection of learning time (ensuring lessons start on time, and are not interrupted or cancelled due to school meetings/events).	□	□
Staff and pupil attendance	□	□
Homework	□	□

Policies on the provision of learning opportunities

Note that in this context the terms 'learning opportunities' and 'opportunity to learn' describe the extent to which your school provides students with access to content that is in line with the material assessed in students' Key Stage 4 courses.

	Areas covered by last year's school policies	Areas covered by your current policies
A school mission statement associated with the provision of learning opportunities	□	□
The content of the curriculum	□	□
Teaching aims associated with the provision of learning opportunities	□	□
The selection of appropriate textbooks	□	□
The use of additional learning resources	□	□
The school learning arrangements	□	□
Long-term planning of learning opportunities	□	□
Short-term planning of learning opportunities	□	□
Policies to provide additional support for students with additional learning needs	□	□

Policies on teachers' instructional behaviour

This study is interested in 8 aspects of teachers' instructional behaviour. These are listed below.

	Areas covered by last year's school policies	Areas covered by your current policies
Orientation tasks (identifying learning objectives)	□	□
Structuring tasks (outlining or reviewing content, calling attention to main ideas and/or signalling transitions)	□	□
Questioning	□	□
Teacher-modelling (teaching problem-solving skills)	□	□
Application tasks (seat-work/small group tasks)	□	□
The teacher's role in making the classroom a learning environment (keeping students on-task and minimizing disruptive behaviour in lessons)	□	□
Teachers' management of lesson time	□	□
Classroom assessments	□	□

Policies on the school learning environment:

Learning does not only occur in classrooms. This questionnaire considers 5 policies that influence learning outside of the classroom. These are listed in column one of the table below.

	Areas covered by last year's school policies	Areas covered by your current policies
Student behaviour outside of the classroom	□	□
Collaboration and interaction between teachers	□	□
Partnership policy (i.e. the relations of school with community, parents, and advisers)	□	□
Provision of sufficient learning resources to students and teachers	□	□
Values in favour of teacher and student learning	□	□

The focus of the school policies:

To what extent did the following policies dictate teachers' and students' actions LAST academic year?

I.e. How specific were the policies about what teachers and students must do. Please consider school-organized actions/interventions as well as the official policy documentation.

	N/A	1. Not at all	2. Slightly	3. Somewhat	4. Largely	5. Completely
Policies on quantity of instruction (i.e. timetables, attendance, the protection of lesson time, and homework policies)	○	○	○	○	○	○
Policies on the provision of learning opportunities (curriculum related policies from question 2 of this section)	○	○	○	○	○	○
Policies on teachers' instructional behaviour (see question 3 of this section)	○	○	○	○	○	○
Policies on the school learning environment (Teacher/student values and interactions outside of lessons - question 4)	○	○	○	○	○	○

To what extent did the following policies dictate teachers' and students' actions THIS academic year?

	N/A	1. Not at all	2. Slightly	3. Somewhat	4. Largely	5. Completely
Policies on quantity of instruction	○	○	○	○	○	○
Policies on the provision of learning opportunities	○	○	○	○	○	○
Policies on teachers' instructional behaviour	○	○	○	○	○	○
Policies on the school learning environment	○	○	○	○	○	○

On average how many objectives were pursued by the following policies LAST academic year?

IMPORTANT DEFINITION. An objective refers to an aspect of the school provisions that a policy is intended to improve. For example, a homework policy may aim increase instruction time and improve the school's relationship with parents. Pupil attainment is the overall aim of education and should not be seen as a separate objective. Once again, consider any actions/interventions that were associated with the discussed policies.

	N/A	1.	2.	3.	4.	5+
Each of the policies on quantity of instruction	○	○	○	○	○	○
Each of the policies on the provision of learning opportunities	○	○	○	○	○	○
Each if the policies on teachers' instructional behaviour	○	○	○	○	○	○
Each of the policies on the school learning environment	○	○	○	○	○	○

On average how many objectives were pursued by the following policies THIS academic year?

	N/A	1.	2.	3.	4.	5+
Each of the policies on quantity of instruction	○	○	○	○	○	○
Each of the policies on the provision of learning opportunities	○	○	○	○	○	○
Each if the policies on teachers' instructional behaviour	○	○	○	○	○	○
Each of the policies on the school learning environment	○	○	○	○	○	○

The introduction and modification of the school policies:

When were the following policies established?

Please do NOT consider revisions to policies. We are interested in how long your school has employed the same overall approach to each area. Minor modifications to the policies will be assessed shortly. In the absence of an official policy document a school-organized action/intervention may be considered its equivalent.

	N/A	This academic year	Last academic year	3 years ago	4 years ago	5 years ago	6+years ago
The school policy on quantity of instruction	○	○	○	○	○	○	○
The school policy on the provision of learning opportunities	○	○	○	○	○	○	○
The school policy on teachers' instructional behaviour	○	○	○	○	○	○	○
The school policy on the school learning environment	○	○	○	○	○	○	○

On average how frequently are modifications made to the following policies?

These changes may include minor modifications to documentation or the introduction of new actions/interventions that support or expand upon the official policy.

	N/A	This academic year	Last academic year	3 years ago	4 years ago	5 years ago	6+years ago
The school policy on quantity of instruction	○	○	○	○	○	○	○
The school policy on the provision of learning opportunities	○	○	○	○	○	○	○
The school policy on teachers' instructional behaviour	○	○	○	○	○	○	○
The school policy on the school learning environment	○	○	○	○	○	○	○

Please consider the changes that were made to the school policies LAST academic year. Were these changes based on the systematic evaluation of the school's existing policies?

A change in policy may involve either a change to the official documentation or an associated action/intervention. If the policy for a particular area WAS NOT CHANGED please consider whether the decision to retain the existing policy was based upon evaluation data.

	No	Yes
Changes made to the school policies on quantity of instruction	○	○
Changes made to the school policies on the provision of learning opportunities	○	○
Changes made to the school policies on teachers' instructional behaviour	○	○
Changes made to the school policies on the school learning environment	○	○

Please consider the changes that were made to the school policies THIS academic year. Were these changes based on the systematic evaluation of the school's existing policies?

A change in policy may involve either a change to the official documentation or an associated action/intervention. If the policy for a particular area WAS NOT CHANGED please consider whether the decision to retain the existing policy was based upon evaluation data.

	No	Yes
Changes made to the school policies on quantity of instruction	○	○
Changes made to the school policies on the provision of learning opportunities	○	○
Changes made to the school policies on teachers' instructional behaviour	○	○
Changes made to the school policies on the school learning environment	○	○

The quality of the school policies:

Since last academic year has there been a change in the clarity of the following policies?

Clear policies are unambiguous and outline steps to be taken if a problem is about to be created, e.g. if a teacher is sick. Again, consider interventions as well as official documentation.

	N/A	Large decrease	Small decrease	No change	Small increase	Large increase
Policies on quantity of instruction	○	○	○	○	○	○
Policies on the provision of learning opportunity	○	○	○	○	○	○
Policies of teachers' instructional behaviour	○	○	○	○	○	○
Policies on the school learning environment	○	○	○	○	○	○

Since last academic year has there been a change in the alignment of the following policies with the academic literature?

It is IMPORTANT TO NOTE that providing additional learning time or learning opportunities would count as increasing the alignment of the quantity of instruction and opportunity to learn policies with the literature, respectively. Once again please consider school actions/interventions as well as the official documentation.

	N/A	Large decrease	Small decrease	No change	Small increase	Large increase
Policies on quantity of instruction	○	○	○	○	○	○
Policies on the provision of learning opportunity	○	○	○	○	○	○
Policies of teachers' instructional behaviour	○	○	○	○	○	○
Policies on the school learning environment	○	○	○	○	○	○

Since last academic year has there been a change level of support provided to teachers and/or students to implement the following policies?

In addition to the official documentation, consider any actions/interventions associated with the policies.

	N/A	Large decrease	Small decrease	No change	Small increase	Large increase
Policies on quantity of instruction	○	○	○	○	○	○
Policies on the provision of learning opportunity	○	○	○	○	○	○
Policies of teachers' instructional behaviour	○	○	○	○	○	○
Policies on the school learning environment	○	○	○	○	○	○

Since last academic year has there been a change in the level of influence that the following policies have had on teacher and student behaviour?

In addition to the official documentation, consider any actions/interventions associated with the policies.

	N/A	Large decrease	Small decrease	No change	Small increase	Large increase
Policies on quantity of instruction	○	○	○	○	○	○
Policies on the provision of learning opportunity	○	○	○	○	○	○
Policies of teachers' instructional behaviour	○	○	○	○	○	○
Policies on the school learning environment	○	○	○	○	○	○

Have any of the following changed since last academic year?

	Large decrease	Small decrease	No change	Small increase	Large increase
The amount of instruction time that the school policies provided students with (consider timetables, homework, cancelled lessons and your policy regarding absenteeism)	○	○	○	○	○
The alignment between the school curriculum and the content assessed in students' Key Stage 4 courses.	○	○	○	○	○

The level of differentiation in the school policies:

Since last academic year, has there been a change in the level of differentiation present in the following policies?

This may include additional support or flexibility that takes into account the needs, personality and preferences of staff and/or pupils. In addition to the official documentation, consider changes in associated actions/interventions.

	N/A	Large decrease	Small decrease	No change	Small increase	Large increase
Policies on quantity of instruction	○	○	○	○	○	○
Policies on the provision of learning opportunity	○	○	○	○	○	○
Policies of teachers' instructional behaviour	○	○	○	○	○	○
Policies on the school learning environment	○	○	○	○	○	○

Since last academic year, has there been a change in the extent to which the teachers were encouraged to differentiate the following aspects of their behaviour?

This may include additional support or flexibility that takes into account the needs, interests, personality and preferences of pupils. In addition to the official documentation, consider changes in associated actions/interventions.

	N/A	Large decrease	Small decrease	No change	Small increase	Large increase
Policies on the provision of learning opportunity	○	○	○	○	○	○
Their use of the eight classroom behaviours	○	○	○	○	○	○

Key Stage 4 subject options and exam entry policy:

What was the average number of qualifications taken by year 11 students?

Please consider double-award qualifications as two qualifications.

	6 or less	7	8	9	10 or more
Last academic year	○	○	○	○	○
This academic year	○	○	○	○	○

Has the number of year 11 students entered for the following examinations changed since last academic year?

EBacc is the abbreviation of English Baccalaureate

	Large decrease	Small decrease	No change	Small increase	Large increase
The EBacc mathematics qualification	○	○	○	○	○
The EBacc English language and English literature qualifications	○	○	○	○	○
GCSEs in other EBacc subjects	○	○	○	○	○
Non-EBacc GCSE subjects and non-GCSE qualifications	○	○	○	○	○
Level 3 qualifications (AS levels)	○	○	○	○	○

Has the amount of instruction time dedicated to the following subjects changed since last academic year?

Note that changes in instruction time may include time spent on homework.

	Large decrease	Small decrease	No change	Small increase	Large increase
Mathematics	○	○	○	○	○
English language and English literature	○	○	○	○	○
The other EBacc subjects	○	○	○	○	○
Non-EBacc GCSE subjects and non-GCSE courses	○	○	○	○	○
Level 3 qualifications (AS levels)	○	○	○	○	○

Have there been any other changes to the school's exam entry procedures?

For example, the school may have chosen to emphasise a particular subject because, in the past, students taking the qualification have made especially high levels of progress. If any such changes have occurred please identify these below and predict whether they will have a) a large negative influence, b) a small negative influence, c) no influence, d) a small positive influence or e) a large positive influence, on your school's value added score.

....

Section C: The evaluation of school policies and actions to improve teaching

The term 'school evaluation policies' refers to the school's own procedures for monitoring and assessing their educational provisions. These evaluations will usually consider pupil attainment levels as well as data collected about specific aspects of policy, such as attendance levels or teacher evaluations. For the purpose of this questionnaire, a distinction has been drawn between the evaluation of teaching (which considers quantity of instruction, the appropriateness of the school curriculum and the quality of teachers instructional behaviours) and the evaluation of the school learning environment (which considers teacher and student interactions, their values about learning
and the provision of resources).

The collection of evaluation data:

Has the frequency with which you collect evaluation data changed since last academic year?

Note that the frequency of data COLLECTION and EVALUATION may not be the same.

	N/A	Large decrease	Small decrease	No change	Small increase	Large increase
Data on teaching policies and/or actions taken to improve teaching	○	○	○	○	○	○
Data on school learning environment and/or actions taken to the school learning environment	○	○	○	○	○	○

Do the following evaluations draw upon less, the same or a more sources of information than they did last academic year?

	N/A	Large decrease	Small decrease	No change	Small increase	Large increase
The evaluation of the teaching policies and/or related actions	○	○	○	○	○	○
The evaluation of the school learning environment policies and/or related actions	○	○	○	○	○	○

Which aspects of your school TEACHING policies were evaluated during the specified time periods?

	Last academic year	Current academic year
The clarity of the policies/actions	□	□
The alignment of the policy/actions with the literature	□	□
The relevance of the policy/actions to the problems encountered by teachers and students	□	□
The impact of the policy/action on school practice	□	□
The effect of policy/action on student outcomes	□	□
The ability of staff and students to implement the policy and/or actions	□	□

Which aspects of the SCHOOL LEARNING ENVIRONMENT were evaluated during the specified time periods?

	Last academic year	Current academic year
Student behaviour outside of the classroom	□	□
Teacher collaboration	□	□
The school's relationships with community, parents and advisors	□	□
The provision of learning resources	□	□
Teachers' and students' values about learning	□	□
Teachers' and students' ability to implement the policy	□	□

How detailed was the feedback from the school evaluations LAST academic year?

An example of very general feedback about the school's teaching policies may be that the provisions were satisfactory. An example of extremely specific feedback would be information on the strengths and weaknesses of each individual teacher.

	N/A	1. Very general	2. Fairly general	3. Moderately detailed	4. Very detailed	5. Extremely detailed
The evaluations of the school teaching policies and/or actions to improve teaching	○	○	○	○	○	○
The evaluations of the school learning environment and/or actions to improve it	○	○	○	○	○	○

How detailed was the feedback from the school evaluations THIS academic year?

	N/A	1. Very general	2. Fairly general	3. Moderately detailed	4. Very detailed	5. Extremely detailed
The evaluations of the school teaching policies and/or actions to improve teaching	○	○	○	○	○	○
The evaluations of the school learning environment and/or actions to improve it	○	○	○	○	○	○

The timing of the school evaluations:

Has the frequency with which the school evaluates the following data changed since last academic year?

Note that the frequency with which the school COLLECTS and EVALUATES data may be different.

	N/A	1. Very general	2. Fairly general	3. Moderately detailed	4. Very detailed	5. Extremely detailed
Data on the school teaching policies and/or related actions	○	○	○	○	○	○
Data concerning the school learning environment	○	○	○	○	○	○

The quality of school evaluations:

LAST year, did your school have a formalized process for reviewing the EVALUATION PROCEDURES listed below?

	No	Yes
The procedures for evaluating the school teaching policies	□	□
The procedures for evaluating the school learning environment	□	□

THIS year, did your school have a formalized process for reviewing the EVALUATION PROCEDURES listed below?

	No	Yes
The procedures for evaluating the school teaching policies	□	□
The procedures for evaluating the school learning environment	□	□

To what extent do you agree with the following statements about LAST YEAR'S EVALUATIONS of the school TEACHING policies?

N/A indicates that your school did not assess this aspect of the school evaluations.

	N/A	Strongly disagree	Disagree	50/50	Agree	Strongly agree
The reliability of school assessments was high	○	○	○	○	○	○
Every monitoring system that your school implemented was used to inform decisions about school practice (i.e. the data is used formatively)	○	○	○	○	○	○
There is strong evidence that the school evaluations could accurately assess the factors that they claimed to assess	○	○	○	○	○	○
There is strong evidence of a relationship between these factors and academic attainment	○	○	○	○	○	○
The benefits of monitoring greatly outweighed the drawbacks	○	○	○	○	○	○

To what extent do you agree with the following statements about THIS YEAR'S EVALUATIONS of the school TEACHING policies?

N/A indicates that your school did not assess this aspect of the school evaluations.

	N/A	Strongly disagree	Disagree	50/50	Agree	Strongly agree
The reliability of school assessments was high	○	○	○	○	○	○
Every monitoring system that your school implemented was used to inform decisions about school practice (i.e. the data is used formatively)	○	○	○	○	○	○
There is strong evidence that the school evaluations could accurately assess the factors that they claimed to assess	○	○	○	○	○	○
There is strong evidence of a relationship between these factors and academic attainment	○	○	○	○	○	○
The benefits of monitoring greatly outweighed the drawbacks	○	○	○	○	○	○

To what extent do you agree with the following statements about LAST YEAR'S EVALUATIONS of the SCHOOL LEARNING ENVIRONMENT policies?

N/A indicates that your school did not assess this aspect of the school evaluations.

	N/A	Strongly disagree	Disagree	50/50	Agree	Strongly agree
The reliability of school assessments is high	○	○	○	○	○	○
Every monitoring system that your school implemented was used to inform decisions about school practice (i.e. the data is used formatively)	○	○	○	○	○	○
There is strong evidence that the school evaluations could accurately assess the factors that they claimed to assess	○	○	○	○	○	○
There is strong evidence of a relationship between these factors and academic attainment	○	○	○	○	○	○
The benefits of monitoring greatly outweighed the drawbacks	○	○	○	○	○	○

To what extent do you agree with the following statements about THIS YEAR'S EVALUATIONS of the SCHOOL LEARNING ENVIRONMENT policies?

N/A indicates that your school did not assess this aspect of the school evaluations.

	N/A	Strongly disagree	Disagree	50/50	Agree	Strongly agree
The reliability of school assessments is high	○	○	○	○	○	○
Every monitoring system that your school implemented was used to inform decisions about school practice (i.e. the data is used formatively)	○	○	○	○	○	○
There is strong evidence that the school evaluations could accurately assess the factors that they claimed to assess	○	○	○	○	○	○
There is strong evidence of a relationship between these factors and academic attainment	○	○	○	○	○	○
The benefits of monitoring greatly outweighed the drawbacks	○	○	○	○	○	○

The emphasis placed on evaluating specific aspects of the school policies:

Has the emphasis that your school places on evaluating the under-preforming aspects of its provisions changed since last academic year?

	N/A	Large decrease	Small decrease	No change	Small increase	Large increase
The emphasis placed on evaluating weaker aspects of the school's teaching (inc. Quantity of instruction, the quality of teachers' instructional behaviours and the appropriateness of the school's curriculum)	○	○	○	○	○	○
The emphasis placed on evaluating weaker aspects of the school learning environment	○	○	○	○	○	○

Section D: Classroom Instruction

In the final section of this questionnaire we will discuss your teachers' use of 8 instructional behaviours. These are defined below:

1. Orientation:
Orientation tasks help students to appreciate the reason for acquiring particular knowledge or skills.

2. Structuring:
Achievement is maximized when teachers begin by outlining the lesson objectives and the content to be covered, signal transitions between parts of the lesson, highlight important concepts and summarize learning at the end of the lesson. These are structuring task.

3. Questioning:
Verbal interactions designed to assess and/or develop students understanding.

4. Teacher-modelling:
Teacher-modelling tasks are used to teach students problem-solving techniques and higher-order thinking.

5. Application:
Application tasks provide an immediate opportunity for students to apply new knowledge. This work can be carried out individually (seat-work) or in small groups.

6. The teachers' role in creating a learning environment in their classrooms:
This refers to the teachers' ability to establish on-task interactions and minimize classroom disruptions.

7. Teachers' management of time:
The teachers' ability to use lesson time effectively.

8. Classroom assessments:
Evaluations of learning that take place in the classroom.

We recognize that teaching standards and styles will vary across the school, however, we would like you describe the typical or average case.

Teachers' use of the eight instructional behaviours:

How frequently did the following behaviours occur LAST academic year?

	1. Almost Never	2. Infrequently	3. Routinely	4. Frequently	5. Very Frequently
Orientation tasks	○	○	○	○	○
Structuring tasks	○	○	○	○	○
Questioning	○	○	○	○	○
Questioning that required an extended point	○	○	○	○	○
Teacher-modelling	○	○	○	○	○
Application tasks	○	○	○	○	○
Teacher-student interactions	○	○	○	○	○
Student-student interactions	○	○	○	○	○
Assessments tasks	○	○	○	○	○
Classroom disruptions caused by poor student behaviour (please note that this is the only negative behaviour in this list, a score of 5 is therefore poor)	○	○	○	○	○

How frequently did the following instructional behaviours occur THIS academic year?

	1. Almost Never	2. Infrequently	3. Routinely	4. Frequently	5. Very Frequently
Orientation tasks	○	○	○	○	○
Structuring tasks	○	○	○	○	○
Questioning	○	○	○	○	○
Questioning that required an extended point	○	○	○	○	○
Teacher-modelling	○	○	○	○	○
Application tasks	○	○	○	○	○
Teacher-student interactions	○	○	○	○	○
Student-student interactions	○	○	○	○	○
Assessments tasks	○	○	○	○	○
Classroom disruptions caused by poor student behaviour (please note that this is the only negative behaviour in this list, a score of 5 is therefore poor)	○	○	○	○	○

How frequently did teachers' respond to behaviour that disrupted the lesson?

Please consider the proportion of problems that the teacher responded to - not the frequency with which problems occurred.

	1. Almost Never	2. Infrequently	3. Routinely	4. Frequently	5. Very Frequently
How frequently did teachers respond to disruptive behaviour LAST academic year?	○	○	○	○	○
How frequently did teachers respond to disruptive behaviour THIS academic year?	○	○	○	○	○

Since last academic year, has there been a change in the proportion of lesson time that was used for teaching?

Please note that this refers to the amount of time that the class were engaged in learning activities. This includes time dedicated to orientation and structuring tasks, time spent in classroom discussions and/or listening to the teacher lecture about a topic. It does not include time spent on classroom management (e.g. organizing the group or dealing with disruptive behaviour) or social activities.

	Large decrease	Small decrease	No change	Small increase	Large increase
The proportion of lesson time used for teaching	○	○	○	○	○

Since last academic year, have teachers covered the school curriculum to greater or lesser extent?

	Large decrease	Small decrease	No change	Small increase	Large increase
Teachers' coverage of the school curriculum	○	○	○	○	○

The focus of teachers' instructional behaviours:

LAST year, did the following teaching behaviours typically refer to part of the lesson, a whole lesson or a series of lessons?

	N/A	Part of lesson	A whole lesson	A series of lessons
Orientation tasks	○	○	○	○
Structuring task	○	○	○	○
Questioning	○	○	○	○
Application tasks	○	○	○	○

THIS year, did the following teaching behaviours typically refer to part of the lesson, a whole lesson or a series of lessons?

	N/A	Part of lesson	A whole lesson	A series of lessons
Orientation tasks	○	○	○	○
Structuring task	○	○	○	○
Questioning	○	○	○	○
Application tasks	○	○	○	○

Has the number of objectives associated with the following activities changed since last academic year?

	N/A	Large decrease	Small decrease	No change	Small increase	Large increase
Orientation tasks	○	○	○	○	○	○
Structuring tasks	○	○	○	○	○	○
Questioning	○	○	○	○	○	○
Application tasks	○	○	○	○	○	○
Classroom assessment tasks	○	○	○	○	○	○

To what extent have the following features changed since last academic year?

	N/A	Large decrease	Small decrease	No change	Small increase	Large increase
The number of problem-solving strategies that can be applied to a number of circumstances (e.g. different lessons) [refers to teacher-modelling tasks]	○	○	○	○	○	○
The number of times that teachers discuss more than one strategy for solving a single problem [refers to teacher modelling task]	○	○	○	○	○	○
The proportion of teacher-student interactions that were related to the learning activities (on-task)	○	○	○	○	○	○
The proportion of student-student interactions that were related to learning activities (on-task)	○	○	○	○	○	○
The proportion of in-class behaviour issues that were the result of previously unresolved issues	○	○	○	○	○	○
The extent to which teachers have attempted to address the issues behind disruptions	○	○	○	○	○	○
The range of assessment methods that were used to evaluate students learning	○	○	○	○	○	○

The timing of teachers' instructional behaviours:

Recall the PREVIOUS academic year. Select any time periods when the activities in column 1 occurred consistently.

Tick any that apply

	The beginning of the lesson	The core of the lesson	The end of the lesson	The beginning of the school year	The middle of the school year	The end of the school year
Orientation tasks	□	□	□	□	□	□
Structuring tasks	□	□	□	□	□	□
Questioning	□	□	□	□	□	□
Application	□	□	□	□	□	□
Task-related teacher-student interactions	□	□	□	□	□	□
Task-related student-student interactions	□	□	□	□	□	□
Classroom assessments	□	□	□	□	□	□
Classroom disruptions (negative factor)	□	□	□	□	□	□

Recall the CURRENT academic year. Select any time periods when the activities in column 1 occurred consistently.

Tick any that apply

	The beginning of the lesson	The core of the lesson	The end of the lesson	The beginning of the school year	The middle of the school year	The end of the school year
Orientation tasks	□	□	□	□	□	□
Structuring tasks	□	□	□	□	□	□
Questioning	□	□	□	□	□	□
Application	□	□	□	□	□	□
Task-related teacher-student interactions	□	□	□	□	□	□
Task-related student-student interactions	□	□	□	□	□	□
Classroom assessments	□	□	□	□	□	□
Classroom disruptions (negative factor)	□	□	□	□	□	□

Have any of the following features changed since last academic year?

	N/A	Large decrease	Small decrease	No change	Small increase	Large increase
The extent to which teachers' orientation tasks take on board students' perspectives	○	○	○	○	○	○
The proportion of teacher-modelling tasks which introduce the strategy after the problem is encountered	○	○	○	○	○	○
The speed with which classroom assessments are analysed, reported and acted upon	○	○	○	○	○	○

Have any of the following changed since last academic year?

	N/A	Large decrease	Small decrease	No change	Small increase	Large increase
The clarity of orientation tasks	○	○	○	○	○	○
The influence that orientation tasks had on students' learning	○	○	○	○	○	○
The clarity of structuring tasks	○	○	○	○	○	○
The influence that structuring tasks had of students' learning	○	○	○	○	○	○
The extent to which lessons and schemes of work were structured so that easier tasks preceded more difficult ones	○	○	○	○	○	○
The clarity of questioning	○	○	○	○	○	○
The appropriateness of question difficulty	○	○	○	○	○	○
The extent to which teachers sustained their interactions with the original respondent by rephrasing and giving clues (during questioning)	○	○	○	○	○	○
The clarity with which problem-solving strategies were introduced [refers to teacher-modelling]	○	○	○	○	○	○
The extent to which application tasks expanded on the material that was taught in the lesson	○	○	○	○	○	○
The extent to which teachers' interventions were able to established the desired form of interaction (on-task behaviour)	○	○	○	○	○	○
The extent to which teachers' interventions solved the underlying issues behind classroom disruptions	○	○	○	○	○	○
The extent that classroom assessments measured what they were intended to measure	○	○	○	○	○	○
The amount of constructive feedback that was given to students after classroom assessments	○	○	○	○	○	○
The influence of assessments on students' learning	○	○	○	○	○	○

How much did the following vary across the school?

	Large amount of variation	Small amount of variation	No variation
The proportion of lesson time used for teaching	□	□	□
Teachers' coverage of the school curriculum	□	□	□
The quality of teaching	□	□	□
The style(s) of teaching adopted by different teachers	□	□	□

The level of differentiation present in teachers' instructional behaviours:

Since last academic year, have teachers become more or less able to adapt the following activities to meet students' individual needs?

	N/A	Large decrease	Small decrease	No change	Small increase	Large increase
Orientation tasks	○	○	○	○	○	○
Structuring tasks	○	○	○	○	○	○
Questioning	○	○	○	○	○	○
Teacher-modelling (problem-solving tasks)	○	○	○	○	○	○
Application tasks	○	○	○	○	○	○
Strategies for keeping students engaged in learning	○	○	○	○	○	○
Strategies for dealing with classroom disruptions	○	○	○	○	○	○
The allocation of lesson time	○	○	○	○	○	○
Classroom assessments and feedback	○	○	○	○	○	○

Student-level factors

Were there any differences between the year 11 cohorts of students?

In the space below please identify any characteristics that were more or less common in the current cohort of year 11 students than they were in last year's 11 students. As a minimum please consider potential differences in the proportion of students who are a particular gender, socio-economic class, ethnicity, or have personalities or thinking styles that are suited to secondary education. Differences in students' intelligence, prior attainment levels, motivation and expectations may also be relevant. As is the number of students with special educational needs. For every difference you identify please specify whether the proportion of students with the characteristic has a) decreased dramatically, b) decreased slightly, c) increased slightly, d) increased dramatically.

....

Additional factors

Since last academic year, have there been any other changes which may influence your school's value added score?

If so please identify these below and specify whether you anticipate they will have a) a large negative influence, b) a small negative influence, c) no influence, d) a small positive influence or e) a large positive influence on your school's value added score.

Concluding message:

That concludes this questionnaire. Thank you for your participation

Appendix E: Results Tables from the Detailed Regression Analysis

2017 Analysis:

The relationship between independent variables and school performance in 2017 – Detailed Regression Analysis

Rank	Classification	Dimension	Variable Name	Linear association	Linear R-squared
1	Instructional	Frequency	How frequently teachers responded to classroom disruptions	Negative	0.764
2	Instructional	Frequency	Frequency of student-student interactions	Positive	0.709
3	Instructional	Frequency	Frequency of teacher-student interactions	Positive	0.694
4	Exam entry	n/a	Number of pupils included in Progress 8 measure	Negative	0.656
5	Exam entry	n/a	Percentage of Year 11 entering all English Baccalaureate subject areas	Positive	0.628
6	Exam entry	n/a	Percentage of Year 11 entering Baccalaureate Language subject area	Positive	0.568
7	Intake	n/a	Percentage of persistent absentees at the school (greater 10% absence)	Negative	0.542
8	Exam entry	n/a	Average number of open slots filled in Attainment 8	Positive	0.459
9	Exam entry	n/a	Average number of GCSE entries per pupil (including equivalencies)	Positive	0.398
10	Exam entry	n/a	Percentage of Year 11 entering Baccalaureate Humanities subject area	Positive	0.384
11	Instructional	Stage	Stages of academic year in which classroom disruptions consistently took place	Negative	0.384
12	Exam entry	n/a	Average number of GCSE entries per pupil (not including equivalencies)	Positive	0.384
13	Instructional	Frequency	Frequency of classroom disruptions	Negative	0.381
14	Exam entry	n/a	Percentage of Year 11 entering Baccalaureate English subject area	Positive	0.38
15	Policies	Quality	Reliability of mechanisms that were used to evaluate the school teaching policies	Positive	0.374
16	Instructional	Focus	Whether orientation tasks typically referred to a series, whole or part of the lesson	Positive	0.367
17	Instructional	Frequency	Frequency of open-ended questions	Positive	0.358
18	Intake	n/a	Percentage of Year 11 pupils that had SEN and a Statement or EHC plan	Negative	0.352
19	Policies	Stage	Number of years that the current policies on teachers instructional behaviours have been implemented	Positive	0.334
20	Instructional	Frequency	Frequency of questioning	Positive	0.321
21	Intake	n/a	Overall percentage of absence at the school	Negative	0.32
22	Exam entry	n/a	Average number of EBacc slots filled in Attainment 8	Positive	0.316
23	Intake	n/a	Percentage of Year 11 with SEN but no Statement or EHC plan	Positive	0.273
24	Instructional	Focus	Whether structuring tasks typically referred to a series, whole or part of the lesson	Positive	0.267
25	Intake	n/a	Percentage of Progress 8 entrants speaking English as an additional language (EAL)	Negative	0.266
26	Intake	n/a	Percentage of Year 11 pupils speaking English as an additional language (EAL)	Negative	0.216
27	Policies	Focus	Number of objectives pursued by SLE policies	Negative	0.202
28	Policies	Focus	Number of objectives pursued by quantity of instruction policies	Negative	0.195
28	Policies	Focus	Number of objectives pursued by the policies on the provision of learning opportunities	Negative	0.195
30	Intake	n/a	Percentage of Year 11 pupils with SEN (With or without Statement/EHC plan)	Positive	0.187
31	Policies	Quality	Proportion of evaluation data that was used formatively; teaching evaluations.	Positive	0.173
32	Instructional	Stage	Stages of lesson in which orientation tasks consistently took place	Negative	0.17
32	Instructional	Stage	Stages of lesson in which structuring tasks consistently took place	Negative	0.17
34	Policies	Focus	Extent that the quantity of instruction policies dictated teachers' and students' actions	Negative	0.169
35	Instructional	Frequency	Frequency of application tasks	Positive	0.166
36	Exam entry	n/a	Percentage of Year 11 entering Baccalaureate Maths subject area	Positive	0.155
37	Policies	Focus	Extent that the policies on the provision of learning opportunities dictated teachers' and students' actions	Negative	0.146
38	Instructional	Frequency	Frequency of teacher-modelling tasks	Positive	0.14
39	Policies	Quality	Extent to which the benefits of evaluating the school teaching policies outweighed the drawbacks	Positive	0.127

40	Policies	Frequency	Coverage of quantity of instruction policy areas (4 policy areas)	Positive	0.121
41	Instructional	Focus	Whether application tasks typically referred to a series, whole or part of the lessons	Positive	0.117
42	Instructional	Stage	Stages of academic year in which questioning tasks consistently took place	Negative	0.113
43	Instructional	Stage	Stages of lesson in which questioning tasks consistently took place	Negative	0.112
44	Policies	Quality	Face validity of the mechanisms that used to evaluate the school teaching policies	Positive	0.108
45	Policies	Stage	Number of years that the current learning opportunity policy has been implemented	Negative	0.098
46	Policies	Stage	Number of years that the current SLE policies been implemented	Positive	0.087
47	Instructional	Stage	Stages of lesson in which classroom disruptions consistently took place activity takes place	Negative	0.086
48	Instructional	Stage	Stages of academic year in which application tasks consistently took place	Positive	0.084
48	Instructional	Stage	Stages of academic year in which teacher-student interactions consistently took place	Positive	0.084
48	Instructional	Stage	Stages of academic year in which student-student interactions consistently took place	Positive	0.084
51	Exam entry	n/a	Percentage of Year 11 entering Baccalaureate Science subject area	Positive	0.079
52	Policies	Focus	Number of objectives pursued by the policies on teachers' instructional behaviours.	Negative	0.077
53	Instructional	Frequency	Frequency of classroom assessment tasks	Positive	0.076
54	Instructional	Stage	Stages of lesson in which classroom assessments consistently took place	Positive	0.062
55	Policies	Focus	Extent that the SLE policies dictated teachers' and students' actions	Positive	0.06
56	Policies	Quality	Influence that the evaluation of the SLE polices had upon students' learning	Negative	0.056
57	Policies	Quality	Reliability of mechanisms that were used to evaluate the SLE	Positive	0.053
58	Policies	Quality	Influence that the evaluations of school teaching policies had upon students' learning	Positive	0.051
59	Policies	Stage	Average number of years between modifications of the quantity of instruction policies	Negative	0.047
60	Policies	Focus	Extent that the policies on teachers' instructional behaviour dictated teachers' and students' actions	Positive	0.045
61	Instructional	Frequency	Frequency of structuring tasks	Positive	0.042
62	Instructional	Stage	Stages of lesson during which application tasks consistently took place	Negative	0.042
63	Policies	Stage	Whether modifications in the learning opportunity policies were based upon data from systematic evaluations	Positive	0.032
63	Policies	Stage	Whether modifications to the instructional behaviour policies were based upon data from systematic evaluations	Positive	0.032
63	Policies	Stage	Were modifications in the SLE policies were based upon data from systematic evaluations	Positive	0.032
66	Instructional	Stage	Stages of the academic year in which orientation tasks consistently took place	Positive	0.029
67	Instructional	Stage	Stages of academic year in which structuring tasks consistently took place	Positive	0.026
68	Policies	Stage	Average number of years between modifications of the policies on teachers' instructional behaviours	Positive	0.025
69	Policies	Stage	Average number of years between modifications of the SLE policies	Positive	0.024
70	Instructional	Stage	Stages of lesson the teacher-student interactions consistently took place	Negative	0.02
71	Policies	Stage	Average number of years between modifications of the school's learning opportunity policies.	Negative	0.02
72	Policies	Stage	Whether there was a formalised procedure for evaluating the mechanisms that were used to assess the school teaching policies	Positive	0.017
72	Policies	Stage	Whether there was a formalised procedure for evaluating the mechanisms that were used to assess the SLE policies	Positive	0.017
74	Instructional	Stage	Stages of academic year in which classroom assessments consistently took place	Positive	0.016
75	Intake	n/a	Percentage of Progress 8 entrants that were female	Negative	0.015
76	Policies	Quality	Extent to which the benefits of evaluating the SLE policies outweighed the drawbacks	Negative	0.011
77	Policies	Focus	Aspects of the school teaching policies that were evaluated (6 policy areas)	Positive	0.011
78	Instructional	Frequency	Frequency of orientation tasks	Negative	0.005
79	Policies	Frequency	Coverage of the SLE policies (5 policy areas)	Negative	0.005
80	Policies	Stage	Number of years the current quantity of instruction policies have been implemented	Negative	0.004
81	Policies	Focus	Level of feedback generated by the evaluations of the SLE	Positive	0.002
82	Exam entry	n/a	Percentage of Year 11 pupils entered into Progress 8	Positive	0.002
83	Policies		How many of the 8 effective teaching behaviours are covered by	Negative	0.001

		Frequency	the policies on teachers' instructional behaviours (8 policy areas)		
84	Policies	Frequency	Coverage of learning opportunity policies (9 policy areas)	Positive	0.001
85	Policies	Quality	Proportion of evaluation data that was used formatively; SLE evaluations.	Negative	0.001
85	Policies	Quality	Face validity of the mechanisms that used to evaluate the SLE policies	Negative	0.001
87	Policies	Focus	Level of feedback generated by the evaluation of school teaching policies	Positive	0.001
88	Instructional	Focus	Whether questioning typically refer to a series, whole or part of the lessons	Negative	0.001
89	Intake	n/a	Percentage of Progress 8 entrants that were disadvantaged	Positive	0.000
90	Instructional	Stage	Stages of lesson in which student-student interactions consistently took place	Positive	0.000
91	Policies	Focus	Aspects of the SLE policies that were evaluated (6 aspects total)	Positive	0.000
92	Intake	n/a	Percentage of Progress 8 entrants that were non-mobile	Negative	0.000
93	Policies	Stage	Whether modifications to in the quantity of instruction policies were based upon data from systematic evaluations	Negative	0.000

*Shading in Column 5 signifies the direction of an association was inconsistent with our expectations.

2018 Analysis:

The relationship between independent variables and school performance in 2018 – Detailed Regression Analysis

Rank	Classification	Dimension	Variable Name	Linear association	Linear R-squared
1	Exam entry	n/a	Average number of open slots filled in Attainment 8	Positive	0.761
2	Intake	n/a	Overall percentage of absence at the school	Negative	0.741
3	Policies	Quality	Proportion of evaluation data that was used formatively; teaching evaluations.	Positive	0.723
4	Exam entry	n/a	Percentage of Year 11 entering Baccalaureate Maths subject area	Positive	0.695
5	Exam entry	n/a	Percentage of Year 11 entering Baccalaureate English subject area	Positive	0.615
6	Intake	n/a	Percentage of persistent absentees at the school	Negative	0.607
7	Exam entry	n/a	Average number of GCSE entries per pupil (including equivalencies)	Positive	0.574
8	Instructional	Frequency	Frequency of student-student interactions	Positive	0.553
9	Policies	Quality	Extent to which the benefits of evaluating the school teaching policies outweighed the drawbacks	Positive	0.544
10	Intake	n/a	Percentage of Progress 8 entrants that were disadvantaged	Negative	0.49
11	Instructional	n/a	Consistency in teachers' coverage of the school curriculum	Negative	0.436
12	Instructional	n/a	Consistency in the proportion of lesson time that was used for teaching	Negative	0.408
12	Instructional	n/a	Consistency in the quality of teachers' instruction	Negative	0.408
14	Intake	n/a	Percentage of Year 11 pupils with SEN (With or without Statement/EHC plan)	Negative	0.391
15	Exam entry	n/a	Percentage of Year 11 entering Baccalaureate Language subject area	Positive	0.374
16	Intake	n/a	Percentage of Year 11 pupils that had SEN and a Statement or EHC plan	Negative	0.369
17	Policies	Stage	Number of years that the current learning opportunity policies had been implemented	Negative	0.357
18	Instructional	Frequency	Frequency of open-ended questions	Positive	0.348
19	Instructional	Frequency	Frequency of classroom disruptions	Negative	0.344
20	Exam entry	n/a	Percentage of Year 11 entering Baccalaureate Science subject area	Positive	0.333
21	Instructional	Stage	Stages of academic year in which classroom disruptions consistently took place	Negative	0.319
22	Policies	Focus	Extent that the quantity of instruction policies dictated teachers' and students' actions	Negative	0.318
23	Policies	Focus	Extent that the policies on the provision of learning opportunities dictated teachers’ and students’ actions	Negative	0.318
24	Instructional	Frequency	Frequency of questioning	Positive	0.297
25	Policies	Focus	Level of feedback generated by the evaluation of school teaching policies	Negative	0.295
26	Exam entry	n/a	Percentage of Year 11 entering all English Baccalaureate subject areas	Positive	0.275
27	Exam entry	n/a	Average number of EBacc slots filled in Attainment 8	Positive	0.27
28	Policies	Stage	Number of years that the current quantity of instruction policies had been implemented	Negative	0.269
29	Policies	Focus	Level of feedback generated by the evaluations of the SLE and/or actions taken to improve it	Negative	0.256
30	Intake	n/a	Percentage of Year 11 with SEN but no Statement or EHC plan	Negative	0.255
31	Instructional	Frequency	Frequency of application tasks	Positive	0.226
32	Instructional	Frequency	Frequency of teacher-student interactions	Positive	0.217
33	Instructional	Frequency	Frequency of teacher-modelling tasks	Positive	0.216
34	Exam entry	n/a	Average number of GCSE entries per pupil (not including equivalencies)	Positive	0.212
35	Policies	Quality	Reliability of mechanisms that were used to evaluate the school teaching policies	Positive	0.209
36	Instructional	Focus	Whether structuring tasks referred to a series, whole or part of lessons	Positive	0.208
37	Instructional	Frequency	Frequency of classroom assessment tasks	Positive	0.184
38	Policies	Stage	Average number of years between modifications of the instructional behaviours policies	Positive	0.162
39	Instructional	n/a	Consistency of teaching style(s) used by teachers	Negative	0.153
40	Instructional	Stage	Stages of academic year in which classroom assessments consistently took place	Negative	0.132
41	Policies	Focus	Aspects of the SLE policies that were evaluated (6 aspects total)	Positive	0.119
42	Instructional	Stage	Stages of lesson in which orientation tasks consistently took place	Negative	0.115
42	Instructional	Stage	Stages of lesson in which structuring tasks consistently took place	Negative	0.115
44	Exam entry	n/a	Percentage of Year 11 entering Baccalaureate Humanities subject	Positive	0.109

			area		
45	Instructional	Stage	Stages of lesson in which teacher-student interactions consistently took place	Positive	0.105
46	Instructional	Stage	Stages of academic year in which questioning tasks consistently took place	Negative	0.101
47	Policies	Stage	Average number of years between modifications of the quantity of instruction policies	Negative	0.095
48	Exam entry	n/a	Number of pupils included in Progress 8 measure	Negative	0.095
49	Policies	Stage	How long the current instructional behaviour policies had been implemented	Positive	0.09
50	Instructional	Frequency	Frequency of structuring tasks	Positive	0.079
51	Instructional	Focus	Whether questioning tasks referred to a series, whole or part of lessons	Negative	0.079
52	Intake	n/a	Percentage of Year 11 pupils that spoke English as an additional language (EAL)	Negative	0.075
53	Policies	Stage	How long the current SLE policies had been implemented	Positive	0.073
54	Instructional	Frequency	How frequently teachers responded to classroom disruptions	Negative	0.072
55	Policies	Frequency	Coverage of quantity of instruction policy areas (4 policies areas)	Positive	0.07
56	Intake	n/a	Percentage of girls in the Progress 8 measure	Positive	0.069
57	Instructional	Stage	Stages of academic year in which application tasks consistently took place	Positive	0.057
57	Instructional	Stage	Stages of academic year in which teacher-student interactions consistently took place	Positive	0.057
57	Instructional	Stage	Stages of academic year in which student-student interactions consistently took place	Positive	0.057
60	Policies	Stage	Average number of years between modifications of the SLE policies	Positive	0.054
61	Intake	n/a	Percentage of Progress 8 pupils that spoke English as an additional language	Negative	0.048
62	Policies	Frequency	Coverage of learning opportunity policies (9 policy areas)	Negative	0.039
63	Instructional	Stage	Stages of lesson in which questioning tasks consistently took place	Negative	0.035
64	Policies	Focus	Number of objectives pursued by the policies on teachers' instructional behaviours.	Negative	0.034
64	Policies	Focus	Number of objectives pursued by SLE policies	Negative	0.034
66	Policies	Stage	Whether modifications to the learning opportunity policies were based upon data from systematic evaluations	Positive	0.034
67	Exam entry	n/a	Percentage of Year 11 pupils entered into Progress 8	Negative	0.026
68	Policies	Frequency	Coverage of the SLE policies (5 policy areas)	Negative	0.025
69	Instructional	Focus	Whether application tasks referred to a series, whole or part of lessons	Negative	0.025
70	Instructional	Stage	Stages of lesson in which application tasks consistently took place	Positive	0.023
70	Instructional	Stage	Stages of lesson in which student-student interactions consistently took place	Positive	0.023
72	Instructional	Stage	Stages of lessons in which classroom assessments consistently took place	Positive	0.022
73	Policies	Focus	Aspects of the school teaching policies that were evaluated (6 aspects total)	Negative	0.021
74	Instructional	Focus	Whether orientation tasks referred to a series, whole or part of lessons	Positive	0.015
75	Policies	Focus	Extent that the policies on teachers' instructional behaviour dictated teachers' and students' actions	Negative	0.012
76	Policies	Quality	Extent to which the benefits of evaluating the SLE policies outweighed the drawbacks	Positive	0.01
77	Policies	Stage	Average number of years between modifications of the learning opportunity policies	Negative	0.006
78	Policies	Focus	Number of objectives pursued by quantity of instruction policies	Positive	0.005
78	Policies	Focus	Number of objectives pursued by the policies on the provision of learning opportunities	Positive	0.005
80	Instructional	Stage	Stages of lesson in which classroom disruptions consistently took place	Negative	0.003
81	Policies	Quality	Face-validity of the mechanisms that were used to evaluate the school teaching policies	Positive	0.003
82	Instructional	Stage	Stages of academic year in which structuring tasks consistently took place	Positive	0.002
83	Policies	Quality	Face-validity of the mechanisms used to evaluate the SLE policies	Negative	0.002
84	Policies	Quality	Influence that the evaluation of the SLE polices had upon students' learning	Negative	0.002
85	Instructional	Stage	Stages of academic year in which orientation tasks consistently took place	Negative	0.002
86	Policies	Frequency	How many of the 8 teaching behaviours are covered by the policies on teachers' instructional behaviours (8 policy areas)	Negative	0.002
87	Policies	Stage	Whether modifications to the quantity of instruction policies were based upon data from systematic evaluations	Positive	0.001
88	Policies	Focus	Extent that the SLE policies dictated teachers' and students'	Negative	0.001

			actions		
89	Policies	Quality	Reliability of mechanisms that were used to evaluate the SLE	Positive	0.001
90	Instructional	Frequency	Frequency of orientation tasks	Positive	0.000
91	Intake	n/a	Percentage of Progress 8 entrants that were non-mobile	Negative	0.000
92	Policies	Quality	Proportion of evaluation data that was used formatively; SLE evaluations.	Positive	0.000
n/a	Policies	Stage	Whether modifications to the SLE policies were based upon data from systematic evaluations	N/A**	0
n/a	Policies	Stage	Whether there was a formalised procedure for evaluating the mechanisms that were used to assess the school teaching policies	N/A**	0
n/a	Policies	Stage	Whether there was a formalised procedure for evaluating the mechanisms that were used to assess the SLE policies	N/A**	0
n/a	Policies	Stage	Whether modifications to the instructional behaviour policies were based upon data from systematic evaluations	N/A**	0
n/a	Policies	Quality	Influence that the evaluations of school teaching policies had upon students' learning	N/A*	0

*The direction of these relationships could not be assessed due to a lack of variation in the independent variable.

** Shading in Column 5 signifies that the direction of an association was not consistent with our expectations.

Change Analysis (2017-2018):

The relationship between independent variables and school performance in Detailed Regression Analysis of 2017-18 changes

Rank	Classification	Dimension	Variable Name	Linear association	Linear R-squared
1	Instructional	Frequency	Change in the frequently with which teachers responded to classroom disruptions	Positive	0.678
2	Policies	Quality	Change in the level of influence that the quantity of instruction policies had on teachers and students behaviour	Positive	0.644
3	Policies	Quality	Change in the alignment between the quantity of instruction policies and the academic literature	Positive	0.635
4	School intake	n/a	Change in the percentage of girls in the Progress 8 measure	Positive	0.616
5	Policies	Stage	Number of years that the quantity of instruction policy has been implemented**	Negative	0.606
6	Instructional	Stage	Change in the speed with which classroom assessments are analysed, reported and acted upon.	Positive	0.584
7	Policies	Stage	Number of years that the policies for providing learning opportunities have been implemented**	Negative	0.579
8	Instructional	Stage	Change in the stages of lessons in which questioning tasks took place	Negative	0.576
9	Policies	Focus	Change in the number of objectives pursued by the policies on teachers instructional behaviour	Positive	0.553
10	Instructional	Quality	Change in the influence that structuring tasks had on students' learning	Positive	0.548
10	Instructional	Quality	Change in the extent to which lessons and schemes of work were structured so that the easier tasks preceded the difficult ones	Positive	0.548
10	Instructional	Quality	Change in the clarity with which problem-solving strategies were introduced	Positive	0.548
10	Instructional	Quality	Change in the extent to which teachers' interventions were able to establish the desired form of interaction (on-task behaviour)	Positive	0.548
10	Instructional	Differentiation	Change in teachers' ability to adapt teacher-modelling tasks to meet students' individual needs	Positive	0.548
15	Instructional	Stage	Change in the stages of academic year that questioning tasks took place	Negative	0.531
16	Policies	Differentiation	Change in the level of differentiation in the policies governing teachers' instructional behaviours	Positive	0.503
16	Policies	Differentiation	Change in the level of differentiation in the SLE policies	Positive	0.503
18	Instructional	Stage	Change in the stages of lessons in which classroom assessments took place	Positive	0.472
19	Instructional	Frequency	Change in the frequency of application tasks	Positive	0.468
20	Instructional	n/a	Consistency of teaching style(s) used by teachers	Negative	0.466
21	Exam entry	n/a	Change in the average number of EBacc slots filled in Attainment 8	Positive	0.465
22	Policies	Focus	Change in the number of objectives pursued by the SLE policies.	Positive	0.465
23	Instructional	Focus	Change in the number of objectives behind classroom assessments task	Positive	0.444
24	Policies	Quality	Change in the level of support provided to teachers and/or students to implement the quantity of instruction policies	Positive	0.443
25	Exam entry	n/a	Change in the number of pupils included in Progress 8 measure	Positive	0.437
26	Instructional	Stage	Change in the stages of academic year that classroom assessments took place	Negative	0.421
27	Policies	Quality	Change in the alignment between the policies on the provision of learning opportunities and the academic literature	Positive	0.415
28	Exam entry	n/a	Change in the percentage of Year 11 entering Baccalaureate Language	Positive	0.4
29	Policies	Quality	Change in the alignment between the SLE policies and the academic literature	Positive	0.367
30	Policies	Quality	Change in the extent to which evaluations of the school teaching policies assessed the factors that they were intended to assess (face validity)	Positive	0.361
31	Exam entry	n/a	Change in the percentage of Year 11 entering all English Baccalaureate subject areas	Positive	0.36
32	Policies	Quality	Change in the clarity of quantity of instruction policies	Positive	0.337
33	Instructional	Frequency	Change in the frequency of teacher-student interactions	Positive	0.333
34	Policies	Focus	Change in the number of objectives that were pursued by the quantity of instruction policies.	Positive	0.332
35	School intake	n/a	Change in the percentage of Year 11 pupils with SEN (with or without Statement/EHC plan)	Negative	0.331
36	Instructional	Frequency	Change in the frequency of classroom assessment tasks	Positive	0.323

37	Policies	Focus	Change in the number of objectives pursued by the policies for the provision of learning opportunities	Positive	0.322
38	Policies	Quality	Change in the level of influence that the policies on the provision of learning opportunities had on teachers' and students' behaviour	Positive	0.317
39	Policies	Quality	Change in the level of support provided to teachers and/or students to implement the policies on the provision of learning opportunities	Positive	0.312
40	Policies	Focus	Change in the level of feedback generated by the evaluations of the school teaching policies	Positive	0.296
41	Policies	Quality	Change in the level of influence that the SLE policies had on teachers' and students' behaviour	Positive	0.291
42	Policies	Differentiation	Change in the extent to which teachers were encouraged to differentiate the learning opportunities that they offer to students	Positive	0.286
43	Policies	Quality	Change in the reliability of the mechanisms/processes that evaluate the school teaching policies	Positive	0.284
44	Instructional	Differentiation	Change in teachers' ability to adapt application tasks to meet students' individual needs	Positive	0.272
45	School intake	n/a	Change in the percentage of persistent absentees at the school	Negative	0.269
46	Instructional	Stage	Change in the stages of lessons in which orientation tasks took place	Negative	0.259
46	Instructional	Stage	Change in the stages of lessons in which structuring tasks took place	Negative	0.259
46	Exam entry	n/a	Change in the average number of open slots filled in Attainment 8	Positive	0.259
49	Policies	Stage	Change in whether the modifications to the quantity of instruction policies were based upon evaluation data (formative use of evaluation data)	Positive	0.254
49	Policies	Stage	Change in whether the modifications to the policies on teachers' instructional behaviours that were based upon evaluation data (formative use of evaluation data)	Positive	0.254
49	Policies	Stage	Change in whether the modifications to the SLE policies that were based upon evaluation data (formative use of evaluation data)	Positive	0.254
49	Policies	n/a	Change in the instruction time dedicated to Non-EBacc GCSEs and Non-GCSEs	Positive	0.254
53	Instructional	Quality	Change in the extent to which teachers' interventions solved the issues underlying classroom disruptions	Positive	0.253
54	Policies	Stage	Average number of years between modifications of the SLE policies**	Positive	0.247
55	Instructional	Quality	Change in the extent to which application tasks expanded upon the material that was taught in the lessons	Positive	0.238
55	Instructional	Quality	Change in the face-validity of classroom assessments	Positive	0.238
55	Instructional	Quality	Change in the influence of classroom assessments on students' learning	Positive	0.238
58	School intake	n/a	Change in the overall percentage of absence at the school	Negative	0.224
59	Instructional	Focus	Change in the range of assessment methods used by teachers during classroom assessments	Positive	0.224
60	Policies	Differentiation	Change in the emphasis that was placed on evaluating the under-performing aspects the schools' instructional provisions	Positive	0.221
60	Policies	Differentiation	Change in the emphasis that was placed on evaluating the underperforming aspects of the SLE	Positive	0.221
62	Policies	Stage	Change in whether the modifications to the policies for providing learning opportunities that were based upon evaluation data (formative use of evaluation data)	Positive	0.219
63	Policies	Quality	Change in the clarity of the policies for providing learning opportunities	Positive	0.215
64	Exam entry	n/a	Change in the percentage of Year 11 pupils entered into Progress 8	Positive	0.21
65	Exam entry	n/a	Change in the percentage of Year 11 entering Baccalaureate Humanities	Positive	0.206
66	Exam entry	n/a	Change in the percentage of Year 11 entering Baccalaureate Maths	Positive	0.203
67	Instructional	Differentiation	Change in teachers' ability to adapt their strategies for establishing on-task behaviour to meet individual students' needs	Positive	0.2
68	Policies	n/a	Change in the instruction time dedicated to other EBacc subjects	Positive	0.192
69	Policies	Differentiation	Change in the extent to which teachers were encouraged to differentiate their use of the 8 instructional behaviours	Positive	0.185
70	Policies	Differentiation	Change in the level of differentiation in the quantity of instruction policies	Positive	0.183
71	Policies	Quality	Change in the clarity of the policies on teachers' instructional behaviours	Negative	0.174
72	Policies	Stage	Change in the frequency with which the school evaluates the school teaching policies	Positive	0.172
72	Policies	Stage	Change in the frequency with which the school evaluates the SLE	Positive	0.172
74	Exam entry	n/a	Change in the percentage of Year 11 entering Baccalaureate Science	Positive	0.171
75	Policies	Frequency	Change in frequency with which the school collects data on the school teaching policies	Positive	0.166
76	Instructional	n/a	Consistency in the proportion of lesson time that was used for	Negative	0.161

			teaching		
76	Instructional	n/a	Consistency in the quality of teachers' instruction	Negative	0.161
78	School intake	n/a	Change in the percentage of Year 11 with SEN but no Statement or EHC plan	Negative	0.157
79	School intake	n/a	Change in the percentage of Progress 8 entrants that were non-mobile	Positive	0.157
80	Instructional	Focus	Change in the number of objectives behind structuring tasks	Positive	0.155
80	Instructional	Focus	Change in the number of objectives behind application tasks	Positive	0.155
82	Policies	Quality	Change in the clarity of the SLE policies	Positive	0.152
83	Instructional	n/a	Change in teachers' coverage of the school curriculum	Positive	0.151
84	Instructional	Quality	Change in the influence that orientation tasks had on students' learning	Positive	0.147
84	Instructional	Quality	Change in the clarity of structuring tasks	Positive	0.147
86	Policies	Stage	Average number of years between modifications of the policies on teachers instructional behaviours**	Positive	0.142
87	Instructional	Frequency	Change in the frequency of questioning	Negative	0.139
87	Policies	n/a	Change in the instruction time dedicated to Mathematics	Positive	0.139
87	Policies	n/a	Change in the instruction time dedicated to English Language and English Literature	Positive	0.139
90	Policies	n/a	Change in the instruction time dedicated to Level 3 qualifications	Positive	0.137
91	Instructional	Frequency	Change in the frequency of structuring tasks	Negative	0.135
92	Policies	Focus	Change in the extent to which the policies on the provision of learning opportunities dictated teachers' and students' actions	Positive	0.127
93	Policies	Focus	Change in the level of feedback generated by the evaluation of the SLE policies	Positive	0.126
94	Instructional	Frequency	Change in the frequency of teacher-modelling tasks	Positive	0.112
95	Policies	Differentiation	Change in the level of differentiation in the policies that govern the provision of students' learning opportunities	Positive	0.112
96	Policies	Frequency	Change in the coverage of the schools' policies on quantity of instruction (4 policy areas)	Negative	0.105
97	Instructional	Frequency	Change in the frequency of classroom disruptions	Negative	0.099
98	Instructional	Focus	Change in the number of times that teachers introduced more than one strategy for solving a problem	Positive	0.098
99	Instructional	Focus	Change in whether orientation tasks typically referred to a series, a whole, or part of the lesson	Negative	0.091
100	Exam entry	n/a	Change in the percentage of Year 11 entering Baccalaureate English	Positive	0.089
101	Policies	Stage	Change in whether there was a formalised procedure for evaluating the mechanisms that were used to assess the school teaching policies	Positive	0.089
101	Policies	Stage	Change in whether there was a formalised procedure for evaluating the mechanisms that were used to assess the school's policies for creating an effective learning environment	Positive	0.089
103	Instructional	Frequency	Change in the frequency of student-student interactions	Positive	0.076
103	Instructional	Focus	Change in the number of objectives behind orientation tasks	Positive	0.076
103	Instructional	Focus	Change in the number of circumstances that problem-solving strategies could be applied to	Positive	0.076
106	Instructional	Stage	Change in the stages of lessons in which student-student interactions took place	Positive	0.073
107	Instructional	Focus	Change in the extent to which teachers attempted to address the issue behind disruptions	Positive	0.067
108	Policies	Stage	Number of years that the policies on teachers' instructional behaviours had been implemented**	Negative	0.062
109	Exam entry	n/a	Change in the average number of GCSE entries per pupil (including equivalencies)	Positive	0.062
110	Policies	Frequency	Change in the coverage of the schools' learning opportunity policies (9 areas assessed)	Negative	0.055
111	Policies	Quality	Change in the amount of instruction time that was provided to students by the school policies	Positive	0.05
112	Instructional	Quality	Change in the amount of constructive feedback that was given to students during/after classroom assessments	Positive	0.047
113	Policies	Quality	Change in the level of influence that the policies on teachers' instructional behaviours had on teachers and students behaviour	Positive	0.044
114	Instructional	Quality	Change in the appropriateness of question difficulty	Positive	0.044
115	Instructional	Stage	Change in the stages of academic year that orientation tasks took place	Negative	0.038
116	Exam entry	n/a	Change in the number of students entered for Level 3 qualifications (AS levels)	Positive	0.034
117	Policies	Stage	Average number of years between modifications of the quantity of instruction policies**	Negative	0.033
118	Policies	Quality	Change in the reliability of the mechanisms/processes that evaluate the SLE	Negative	0.032
119	Instructional	Differentiation	Change in teachers' ability to adapt the allocation of lesson time around students' individual needs	Positive	0.03

120	Policies	Quality	Change in the strength of the relationship between the evaluations of the school teaching policies and students' learning	Positive	0.027
121	Policies	Quality	Change in whether evaluation data that was used to inform decisions about the SLE	Negative	0.024
121	Policies	Quality	Change in the extent to which evaluations of the SLE assessed the factors they were intended to assess	Negative	0.024
123	Policies	Quality	Change in the alignment between the school curriculum and the content assessed at KS4	Positive	0.023
124	Instructional	Focus	Change in the proportion of disruptions that were due to previously unresolved issues	Positive	0.022
125	Exam entry	n/a	Change in the average number of GCSE entries per pupil (not including equivalencies)	Positive	0.021
126	Instructional	Focus	Change in the number of objectives behind questioning tasks	Negative	0.021
127	Policies	Quality	Change in the alignment between the policies on teachers' instructional behaviours and the academic literature	Positive	0.019
128	Policies	Frequency	Change in the coverage of the schools' instructional behaviour policies (8 areas assessed)	Positive	0.019
129	Instructional	Frequency	Change in the proportion of lesson time that was used for teaching	Positive	0.019
130	Instructional	Differentiation	Change in teachers' ability to adapt questioning tasks to meet students' individual needs	Positive	0.018
131	Instructional	Focus	Change in whether questioning tasks referred to a series, the whole or part of the lesson	Positive	0.017
131	Instructional	Focus	Change in whether applications tasks referred to a series, a whole or part of the lesson	Positive	0.017
133	Policies	Quality	Change in the level of support provided to teachers and/or students to implement the SLE policies	Positive	0.017
134	School intake	n/a	Change in the Percentage of Year 11 pupils that spoke English as an additional language (EAL)	Negative	0.016
135	Policies	Quality	Change in the level of support provided to teachers and/or students to implement the policies on teachers' instructional behaviours	Positive	0.015
136	Instructional	Frequency	Change in the frequency of orientation tasks	Positive	0.015
137	Instructional	Stage	Change in the stages of the lesson in which teacher-student interactions took place	Positive	0.015
138	Policies	Quality	Change in the extent to which the benefits of monitoring the school teaching policy outweighed the drawbacks	Positive	0.014
139	Policies	Focus	Change in the extent to which the SLE policies dictated teachers' and students' actions.	Positive	0.012
140	Instructional	n/a	Consistency in teachers' coverage of the school curriculum	Negative	0.012
141	School intake	n/a	Change in the percentage of Year 11 pupils that had SEN and a Statement or EHC plan	Negative	0.008
142	Instructional	Quality	Change in the extent to which teachers sustained their interaction with the original respondent during questioning by rephrasing queries and giving clues	Positive	0.007
143	Policies	Focus	Change in the number of aspects of the SLE policies that were evaluated. (6 areas total)	Positive	0.007
144	School intake	n/a	Change in the percentage of Progress 8 entrants that spoke English as an additional language (EAL)	Positive	0.007
145	Instructional	Focus	Change in the proportion of teacher-student interactions that were task-related	Negative	0.004
145	Instructional	Focus	Change in the proportion of student-student interactions that were task-related	Negative	0.004
147	Policies	Focus	Change in the extent to which the quantity of instruction policies dictated teachers' and students' actions	Negative	0.004
148	Instructional	Quality	Change in the clarity of orientation tasks	Positive	0.004
148	Instructional	Differentiation	Change in teachers' ability to adapt structuring tasks to meet students' individual needs	Positive	0.004
150	Policies	Frequency	Change in frequency with which the school collects data on the SLE	Positive	0.004
150	Policies	Frequency	Change in number of sources of information that the evaluations of the SLE policies drew upon	Positive	0.004
152	Policies	Stage	Average number of years between modifications of the policies for providing learning opportunities (Duplication of 2018 rating)	Positive	0.003
153	Instructional	Stage	Change in the extent to which teachers' orientation tasks consistently took on board students' perspective	Positive	0.002
153	Instructional	Stage	Change in the proportion of teacher-modelling tasks which introduce problem-solving strategy after the problem	Positive	0.002
153	Instructional	Differentiation	Change in teachers' ability to adapt orientation tasks to meet students' individual needs	Positive	0.002
156	Policies	Frequency	Change in the coverage of the SLE polices (5 areas assessed)	Positive	0.002
157	Policies	Focus	Change in the number of aspects of the school teaching policies that were evaluated. (6 policy areas)	Negative	0.001
158	Instructional	Frequency	Change in the frequency of open-ended questions	Negative	0.001
159	School intake	n/a	Change in the percentage of Progress 8 entrants that were disadvantaged	Negative	0.001

160	Instructional	Differentiation	Change teachers' ability to adapt their strategies for dealing with classroom disruptions to individual students' needs	Positive	0.001
161	Instructional	Differentiation	Change in teachers' ability to adapt classroom assessments and feedback around students' individual needs	Positive	0.000
162	Policies	Quality	Change in whether evaluation data was used to inform decisions about school teaching practice	Positive	0.000
163	Instructional	Quality	Change in the clarity of questioning	Positive	0.000
164	Instructional	Stage	Change in the stages of lessons in which classroom disruptions took place	Positive	0.000
165	Policies	Focus	Change in the extent to which the policies on teachers' instructional behaviour dictated teachers' and students' actions.	Negative	0.000
166	Policies	Frequency	Change in number of sources of information that the evaluations of the school teaching policies drew upon	Negative	0.000
167	Policies	Stage	Number of years that the SLE policies have been implemented**	Negative	0.000
n/a	Instructional	Focus	Change in whether structuring tasks typically referred to a series, a whole, or part of the lesson	N/A*	0
n/a	Instructional	Stage	Change in the stages of academic year that structuring tasks consistently took place	N/A*	0
n/a	Instructional	Stage	Change in the stages of lessons in which application tasks consistently took place	N/A*	0
n/a	Instructional	Stage	Change in the stages of academic year that application tasks consistently took place	N/A*	0
n/a	Instructional	Stage	Change in the stages of academic year that teacher-student interactions consistently took place	N/A*	0
n/a	Instructional	Stage	Change in the stages of academic year that student-student interactions consistently took place	N/A*	0
n/a	Instructional	Stage	Change in the stages of academic year that classroom disruptions took place	N/A*	0
n/a	Policies	Quality	Change in the strength of the relationship between the evaluations of SLE and students' learning	N/A*	0
n/a	Policies	Quality	Change in the extent to which the benefits of monitoring the SLE outweighed the drawbacks	N/A*	0

* The direction of these relationships could not be assessed due to a lack of variation in the independent variable. ** Variable does not assess change and is instead a duplicate of the 2018 variable. *** Shading in Column 5 signifies that the direction of an association was not consistent with our expectations.

References

Acquah, D. (2013). *School Accountability in England: Past, Present and Future*. AQA Centre for Education Research and Policy. https://www.researchgate.net/publication/323611355_School_Accountability_in_England_Past_Present_and_Future

Adnett, N. and Davies, P. (2002). *Markets for Schooling: An Economic Analysis.* Routledge.

AERA (2015). AERA Statement on Use of Value-Added Models (VAM) for the Evaluation of Educators and Educator Preparation Programs. *Educational Researcher*, 44(8), 448-452. https://doi.org/10.3102/0013189X15618385

Agresti, A. and Franklin, C. (2014). *Statistics: The Art and Science of Learning from Data, 3rd Edition.* Pearson Education Limited.

Aitkin, M. and Longford, N. (1986). Statistical Modelling Issues in School Effectiveness studies. *Journal of the Royal Statistical Society, Series A (General)*, 149(1), 1-43. DOI: 10.2307/2981882

Alexander, K. L. and Eckland, B. K. (1975). Contextual Effects in High School Attainment Process. *American Sociological Review*, 40(3), 402-416. DOI:10.2307/2094466

Alker, H. R. (1969). A Typology of Ecological Fallacies. In M. Dogan and S. Rakkan (Eds.), *Quantitative Ecological Analysis in the Social Sciences* (pp. 69-86). MIT Press.

Allen, R. and Burgess, S. (2013). Evaluating the Provision of School Performance Information for School Choice. *Economics of Education Review,* 34, 175-190. https://doi.org/10.1016/j.econedurev.2013.02.001

Alwin, D. F. and Otto, L. B. (1977). High School Context Effects on Aspiration. *Sociology of Education*, 50(4), 259-273. https://doi.org/10.2307/2112499

Amrein-Beardsley, A. (2014). *Rethinking Value-Added Models in Education: Critical Perspectives on Tests and Assessment-Based Accountability.* Routledge.

Anastasi, A. (1984). Aptitude and Achievement Tests: The curious Case of the Indestructible Strawperson. In S. Place (Ed.), *Social and Technical Issues in Testing: Implications for Test Construction and Usage* (pp. 129-140). Erlbaum.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for Student Background in Value-Added Assessment of Teachers. *Journal of Educational and Behavioral Statistics*, 29(1), 37–65. https://doi.org/10.3102/10769986029001037

Barr, A. S., Bechdolt, B. V., Coxe, W. W., Gage, N. L., Orleans, J. S., Remmers, H. H. and Ryans, D. G. (1952). Report of the Committee on the Criteria of Teacher Effectiveness. *Review of Educational Research*, 22(3), 238-263. https://doi.org/10.3102/00346543022003238

Barr, A. S., Bechdolt, B. V., Coxe, W. W., Gage, N. L., Orleans, J. S., Remmers, H. H. and Ryans, D. G. (1953). Second Report of the Committee on Criteria of Teacher Effectiveness. *Journal of Educational Research*, 46(9), 641-658. https://doi.org/10.1080/00220671.1953.10882066

Bentler, P. M. (1989). EQS: Structural Equations Program Manual. Los Angeles: BMDP Statistical Software.

Berk, R. A. and Freedman, D. A. (2003). Statistical Assumptions as Empirical Commitments. In T.G. Bloomber and S. Cohen (Eds.) *Law, Punishment, and Social Control: Essays in Honor of Sheldon* Messinger, (2nd Edition, pp.235-254). Aldine de Gruyter.

Bloom, B. S. (1968). Learning for Mastery. *Evaluation Comment*, 1(2), 1-12.

Bondi, L. (1991). Attainment at Primary Schools: An Analysis of Variation Between Schools. *British Educational Research Journal,* 17(3), 203-217. https://doi.org/10.1080/0141192910170301

Boonen, T., Speybroeck, S., Bilde, J., Lamote, C., Damme, J., Onghena, P., Damme, J. V. and Onghena, P., (2014). Does It Matter Who Your Schoolmates Are? An Investigation of the Association between School Composition, School Processes and Mathematics Achievement in the Early Years of Primary Education. *British Educational Research Journal*, 40(3), 441–66. https://doi.org/10.1002/berj.3090

Bosker, R. J. and Scheerens, J. (1989). Issues in the Interpretation of the Results of School Effectiveness Research. *International Journal of Educational Research,* 13(7), 741-751. https://doi.org/10.1016/0883-0355(89)90025-6

Bosker, R. J. and Scheerens, J. (1994). Alternative Models of School Effectiveness Put To The Test. *International Journal of Educational Research*, 21(2), 159-181. https://doi.org/10.1016/0883-0355(94)90030-2

Bradbury, A. (2011). Equity, Ethnicity and the Hidden Dangers of 'Contextual' Measures of School Performance. *Race Ethnicity and Education*, 14(3), 277-291. DOI:10.1080/13613324.2010.543388

Brookover, W. B., Beady, C., Flood, P. K., Schweitzer, J. G., and Wisenbaker, J. (1979). *Schools, Social Systems and Student Achievement: Schools Can Make a Difference*. Praeger.

Brookover, W. B., Schweitzer, J. G., Schneider, J. M., Beady, C. H., Flood, P. K. and Wisenbaker, J. M. (1978). Elementary School Social Climate and School Achievement. *American Educational Research Journal,* 15(2), 301-318. https://doi.org/10.3102/00028312015002301

Brophy, J. (1992). Probing the Subtleties of Subject Matter Teaching. *Educational Leadership,* 49(7), 3-8.

Brophy, J. and Good, T. L. (1986). Teacher Behaviour and Student Achievement. In M.C. Wittrock (Ed.), *Handbook of Research on Teaching* (3rd edition, pp. 328-375). New York: Macmillan

Brown, B. W. and Saks, D. H. (1986). Measuring the Effects of Instructional Time on Student Learning: Evidence from the Beginning Teacher Evaluation Study. *American Journal of Education*, 94(4), 480-500. DOI:10.1086/443863

Brualdi, A. (1999). *Traditional and Modern Conceptions of Validity*, ERIC Clearinghouse on Assessment and Evaluation Washington DC. Retrieved April, 02, 2020, from https://www.ericdigests.org/2000-3/validity.htm.

Bryk, A. S. and Raudenbush, S. W. (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage.

Buchanan, J. (1972) *Theory of Public Choice: Political Applications of* Economics (Volume 1). University of Michigan Press

Buchanan, J. (1975) *The Limits of Liberty: Between Anarchy and Leviathan*. University of Chicago Press.

Burgess, S. and Thomson, D. (2013a). *Key Stage 4 Accountability: Progress Measure and Intervention Trigger*. University of Bristol. http://www.bristol.ac.uk/media-library/sites/cubec/migrated/documents/report11.pdf

Burgess, S. and Thomson, D. (2013b). *Key Stage 4 Accountability: Progress Measure and Intervention Trigger. Technical Annex: Techniques for producing an unbiased national pupil progress line*. University of Bristol. http://www.bristol.ac.uk/media-library/sites/cubec/migrated/documents/technical-annex11.pdf

Cahan, S., and Cohen, N. (1989). Age Versus Schooling Effects On Intelligence Development. *Child Development*, 60(5), 1239-1249. https://doi.org/10.2307/1130797

Cahan, S., and Davis, D. (1987). A Between-Grades Approach to the Investigation of the Absolute Effects of Schooling on Achievement. *American Educational Research Journal,* 24(1), 1-13. https://doi.org/10.3102/00028312024001001

Cahan, S. and Elbaz, J. G. (2000). The Measurement of School Effectiveness. *Studies in Educational Evaluation,* 26(2), 127-42. https://doi.org/10.1016/S0191-491X(00)00012-2

Cameron, K. S., & Whetten, D. A. (1983). *Organizational Effectiveness. A Comparison of Multiple Models*. Academic Press.

Camilli, G. (1996). Standard Errors in Educational Assessment: A Policy Analysis Perspective. *Education Policy Analysis Archives*, 4(4), 1-18. DOI:10.14507/epaa.v4n4.1996

Carrol, J. B. (1963) A Model of School Learning, *Teachers College Record,* 64(8), 723-733.

Ceci, S.J. (1991). How Much Does Schooling Influence General Intelligence and its Cognitive Components? A Reassessment of the Evidence. *Developmental Psychology*, 27(5), 703-722. https://doi.org/10.1037/0012-1649.27.5.703

Ceci, S.J. & Williams, W.M. (1997). Schooling, Intelligence and Income. *American Psychologist*, 52(10), 1051-1058. https://doi.org/10.1037/0003-066X.52.10.1051

Chapman, C. P., Armstrong, P., Harris, A., Muijs, D. R., Reynolds, D. and Sammons, P. (2011) *School Effectiveness and Improvement Research, Policy and Practice: Challenging the orthodoxy?* Routledge.

Chapman, C., Muijs, D., Reynolds, D., Sammons, P. and Teddlie, C. (2015) *The Routledge International Handbook of Educational Effectiveness and Improvement: Research, Policy, and Practice.* Routledge.

Cheng, Y. C. (1993). Profiles of Organizational Culture and Effective Schools. School Effectiveness and School Improvement, 4(2), 85-110. doi:10.1080/0924345930040201

Cheng, Y. C. (1996). *School Effectiveness and School-Based Management: A Mechanism for Development.* Falmer Press.

Chetty, R., Friedman, J. N. and Rockoff, J. E. (2011). *The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood.* National Bureau of Economic Research.

Chetty, R., Friedman, J. N. and Rockoff, J. E. (2014a). Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates. *American Economic Review,* 104(9), 2593-2632. DOI:10.1257/aer.104.9.2593

Chetty, R., Friedman, J. N. and Rockoff, J. E. (2014b). Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *American Economic Review,* 104(9), 2633-2679. DOI:10.1257/aer.104.9.2633

Choi, J. I. and Hannafin, M. (1995). Situated Cognition and Learning Environments: Roles, Structures and Implications for Design. *Educational Technology Research and Development*, 43(2), 53-69. DOI:10.1007/BF02300472

Coe, R. and Fitz-Gibbon, C. T. (1998). School Effectiveness Research: Criticisms and Recommendations. *Oxford Review of Education,* 24(4), 421-438. DOI:10.1080/0305498980240401

Coleman, J. S., Campbell, E. Q., Hobson, C. F., McPartland, J., Mood, A. M., Weinfeld, F. D. and York, R. L. (1966). *Equality of Educational Opportunity.* US Government Printing Office.

Collins, A., Brown, J. S. and Newman, S. E. (1989). Cognitive Apprenticeship: Teaching the Crafts of Reading, Writing and Mathematics. In L. B. Resnick (Ed.), *Knowing, Learning and Instruction* (pp.453-495). Lawrence Erlbaum.

Cook, T.D., & Campbell, D.T. (1979). *Quasi-Experimentation: Design Analysis Issues for Field Settings*. Rand-McNally.

Couldry, N., (2010) *Why Voice Matters: Culture and Politics After Neoliberalism*. SAGE Publications Ltd.

Creemers, B. P. M. (1994) *The Effective Classroom*, Cassell.

Creemers, B. P. M. (2002). From School Effectiveness and School Improvement to Effective School Improvement: Background, Theoretical Analysis, and Outline of the Empirical Study. *Educational Research and Evaluation*, 8(4), 343-362, DOI: 10.1076/ edre.8.4.343.8814

Creemers, B. P. M. (2007). Combining different ways of learning and teaching in a dynamic model of educational effectiveness. *Journal of Basic Education*, 17(1), 1-39. DOI:10.1.1.555.8031

Creemers, B. P. M. and Kyriakides, L. (2008). *The Dynamics of Educational Effectiveness: A Contribution to Policy, Practice and Theory in Contemporary Schools.* Routledge

Creemers, B. P., Kyriakides, L. and Sammons, P. (2010). *Methodological advances in educational effectiveness research.* Routledge.

Cronbach, L.J. (1990). *Essentials of Psychological Testing* (5th ed.). Harper & Row.

Daniel, F. M., Lockwood, J. R., Koretz, D. M. & Hamilton, L. S. (2003). *Evaluating Value-Added Models for Teacher Accountability.* Rand Corporation. https://www.rand.org/content/dam/rand/pubs/monographs/2004/RAND_MG158.pdf

Darling-Hammond, L. (2000). Teacher Quality and Student Achievement: A Review of State Policy Evidence. *Educational Policy Analysis Archives*, 8(1), 1-44. https://doi.org/10.14507/epaa.v8n1.2000

Data Educator (2018, April 20). Progress 8 and ECDL. Data Educator. https://dataeducator.wordpress.com/2018/04/20/progress-8-and-ecdl/

Dearden, L., Micklewright, J. and Vignoles, A. (2011a). The Effectiveness of English Secondary Schools for Pupils of Different Ability Levels. *Fiscal Studies,* 32(2), 25-244. https://doi.org/10.1111/j.1475-5890.2011.00134.x

Dearden, L., Miranda, A. and Rabe-Hesketh, S. (2011b) Measuring School Value Added with Administrative Data: The Problem of Missing Variables. *Fiscal Studies,* 32(2), pp. 263-278. https://doi.org/10.1111/j.1475-5890.2011.00136.x

De Jong, R., Westerholf, K. J. and Kruiter, J. H. (2004). Empirical evidence of a comprehensive model of school effectiveness: a multilevel study in mathematics in the 1st year of junior general education in the Netherlands. *School Effectiveness and School Improvement*, 15(1), 3-31. DOI: 10.1076/sesi.15.1.3.27490

Delors, J. (1996). *Learning: The Treasure Within: Report to UNSESCO of the International Commission for Education.* UNESCO. https://www.eccnetwork.net/sites/default/files/media/file/109590engo.pdf

Dettmers, S., Trautwein, U. and Ludtke, O. (2009). The Relationship Between Homework Time and Achievement is Not Universal: Evidence from Multilevel Analyses in 40 countries. *School Effectiveness and School Improvement,* 20(4), 375-405. DOI:10.1080/09243450902904601

DfE (2010a). A Technical Guide to Contextual Value Added (including English & maths) Key Stage 2 to 4 2010 model. London: Department for Education. Retrieved from https://webarchive.nationalarchives.gov.uk/20130503112607/http://www.education.gov.uk/schools/performance/archive/schools_10/documents.shtml

DfE (2010b). Test and Examination Point Scores used in the 2010 School and College Performance Tables. London: Department for Education. Retrieved from https://webarchive.nationalarchives.gov.uk/20130503112607/http://www.education.gov.uk/schools/performance/archive/schools_10/documents.shtml

DfE (2010c). The Importance of Teaching: The Schools White Paper 2010. London: Department for Education. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/175429/CM-7980.pdf

DfE (2011). How do pupils progress during Key Stages 2 and 3? (Research Report. DFE-RR096). London: Department for Education. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/182413/DFE-RR096.pdf.

DfE (2018). Permanent and fixed period exclusions in England: 2016 t0 2017. London: Department for Education. Retrieved from https://www.gov.uk/government/statistics/permanent-and-fixed-period-exclusions-in-england-2016-to-2017

DfE (2019). School and College Performance Tables: Statement of Intent. London: Department for Education. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/819611/Statement_of_Intent_for_2019.pdf

DfE (2020, February). Secondary Accountability Measures Guide for Maintained Secondary Schools, Academies and Free Schools. London: Department for Education. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/872997/Secondary_accountability_measures_guidance_February_2020_3.pdf

Dieterle, S,, Guarino, C. M., Reckase, M. D. and M Wooldridge, J. M. (2015). How Do Principals Assign Students to Teachers? Finding Evidence in Administrative Data and the Implications for Value Added. *Journal of Policy Analysis and Management*, 34 (1), 32–58. https://doi.org/10.1002/pam.21781

Domas, S. J. and Tiedman, D. V. (1950). Teacher Competence: An Annotated Bibliography, *The Journal of Experimental Education*. 19(2), 101-218. DOI:10.1080/00220973.1950.11010421

Downey, D. B., Von Hippel, P. T., & Broh, B. (2004). Are Schools the Great Equalizer? *American Sociological Review*, 69(5), 613–635. https://doi.org/10.1177/000312240406900501

Doyle, W. (1977) Paradigms for Research on Teacher Effectiveness. *Review of Research in Education*, 5(1), 163-198. https://doi.org/10.3102/0091732X005001163

Driessen, G. & Sleegers, P. (2000). Consistency of Teaching Approach and Student Achievement: An Empirical Test. *School Effectiveness and School Improvement*, 11(1), 57-79. DOI: 10.1076/0924-3453(200003)11:1;1-A;FT057

Dumay, X., Coe, R. & Anumendem, D. N. (2014). Stability Over Time of Different Methods of Estimating School Performance. *School Effectiveness and School Improvement*, 25(1), 64-82. doi:10.1080/09243453.2012.759599

Dunne, R. and Wragg, E. R. (1994). *Effective Teaching.* Routledge.

Edmonds, R. R. (1979). Effective Schools for the Urban Poor. *Educational leadership*, 37(1), 15-11.

EduBase (2018). NPD Key Stage 4 Performance Tables - England (Revised Version) [Dataset] . Accessed through https://get-information-schools.service.gov.uk/Downloads

Elberts, R. W. and Stone, J. A. (1988). Student Achievement in Public Schools: Do Principles Make a Differences? *Economics Educational Review*, 7(3), 291-299. https://doi.org/10.1016/0272-7757(88)90002-7

Entwisle, D. R., Alexander, K. L., & Olson, L. S. (1997). *Children, Schools, and Inequality.* Westview.

EPI (2017). The Introduction of Progress 8. London: Department for Education. Retrieved from https://epi.org.uk/publications-and-research/analysis-introduction-progress-8/

Evans, H. (2008). *Value-added in English schools.* Department for Children, Schools, Families. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.178.9363&rep=rep1&type=pdf

Ferrão, M. E. and Goldstein, H. (2009). Adjusting for Measurement Error in the Value Added Model: Evidence from Portugal. *Quality & Quantity,* 43(6), 951-963. DOI:10.1007/s11135-008-9171-1

Fitz-Gibbon, C. (1996). Monitoring School Effectiveness: Simplicity and Complexivity. In Gray, J., Reynolds, D., Fitz-Gibbon, C. and D. Jesson (Eds.), *Merging Traditions: The Future of Research on School Effectiveness and School Improvement.* Cassell

Fitz-Gibbon, C. (1997). *The Value Added National Project: Final Report: Feasibility Studies for a National System of Value-added Indicators*. SCAA. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/605795/0297_CarolTaylorFitz-Gabbon__Feasibility_Study_Nat_System_VA_Indicators.pdf

Foley, B. and Goldstein, H. (2012). *Measuring success: League tables in the public sector*. British Academy.

Frost, C. and Thompson, S. G. (2000). Correcting for Regression Dilution Bias: Comparison of Methods for a Single Predictor Variable. *Journal of the Royal Statistical Society*, Series A (Statistics in Society), 163(2), 173–89. DOI:10.1111/1467-985X.00164

Gage, N. L. (1963). *Handbook of Research on Teaching*. Rand McNally.

Gibbons, S. and Telhaj, S. (2012). *Peer Effects: Evidence from Secondary School Transition in England*. Forschungsinstitut Zur Zukunft Der Arbeit. http://ftp.iza.org/dp6455.pdf.

Glass, G. V. (2014). A Response to Gorard. *The Psychology of Education Review,* 38(1), 12-13.

Goldstein, H. (1997). Methods in School Effectiveness Research. *School Effectiveness and School Improvement*, 8(4), 369–395. DOI:10.1080/0924345970080401

Goldstein, H. (2008). Evidence and Education Policy – Some Reflections and Allegations.*Cambridge Journal of Education*, 38(3), 393-400. DOI: 10.1080/03057640802299643

Goldstein, H. (2011). *Multilevel statistical models* (4th edn). Wiley.

Goldstein, H. and Leckie, G. (2008). School League Tables: What Can They Really Tell Us?, *Significance,* 5(2), 67-69. DOI:10.1111/j.1740-9713.2008.00289.x

Goldstein, H. and Noden, P. (2004) A Response to Gorard on Social Segregation. *Oxford Review of Education,* 30(3), 441-442. https://doi.org/10.1080/0305498042000260539

Goldstein, H. and Spiegelhalter, D. J. (1996). League Tables and their Limitations: Statistical Issues in Comparisons of Institutional Performance. *Journal of the Royal Statistical Society*, Series A (Statistics in Society), 159(3), 385-443. DOI:10.2307/2983325

Goldstein, H. & Thomas, S. (1996). Using Examination Results as Indicators of School Performance. *Journal of the Royal Statistics Society,* Series A (Statistics in Society), 159(1), 149-163. DOI:10.2307/2983475

Goldstein, H. and Woodhouse, G. (2000). School Effectiveness Research and Educational Policy. *Oxford Review of Education,* 26(3-4), 353-363. DOI:10.1080/3054980020001873

Goldstein, H., Daphne, K. and Anthony, R. (2008). Modelling Measurement Errors and Category Misclassifications in Multilevel Models. *Statistical Modelling,* 8(3), 243–261.

Goldstein, H., Rasbash, J., Yang, M., Woodhouse, G., Pan, H., Nuttall, D. & Thomas, A. (1993). A Multilevel Analysis of School Examination Results. *Oxford Review of Education Review*, 19(4), 425-433.

Gorard, S. (2006a). Is There a School Mix Effect? *Educational Review,* 58(1), 87-94. DOI:10.1080/00131910500352739

Gorard, S. (2006b). Value-added is of little value. *Journal of Educational Policy,* 21(2), 235-243. DOI:10.1080/02680930500500435

Gorard, S. (2007). The Dubious Benefits of Multi-Level Modeling. *International Journal of Research & Method in Education,* 30(2), 221-236. DOI: 10.1080/17437270701383560

Gorard, S. (2008a). The Value-Added of Primary Schools: What is it Really Measuring? *Educational Review,* 60(2), 179-185. DOI:10.1080/00131910801934185

Gorard, S. (2008b) *Quantitative Research in Education: Volumes 1 to 3*. Sage

Gorard, S. (2010a). Serious Doubts about School Effectiveness. *British Educational Research Journal,* 36(5), 745-766. https://doi.org/10.1080/01411920903144251

Gorard, S. (2010b). All Evidence is Equal: The Flaw in Statistical Reasoning. *Oxford Review of Education*, 36(1), 63-77. DOI:10.1080/03054980903518928

Gorard, S. (2011a) *Comments on 'The value of educational effectiveness research'*. BERA Conference: BERA. https://research.birmingham.ac.uk/portal/en/publications/comments-on-the-value-of-educational-effectiveness-research(049c1113-9b7a-4978-b370-98cdc55d3e3a).html

Gorard, S. (2011b). Doubts About School Effectiveness Exacerbated by Attempted Justification. *Research Intelligence*, 114, 24-26.

Gorard, S. (2011c). Now You See It, Now You Don't: School Effectiveness as Conjuring? *Research in Education,* 86(1), 39-45. https://doi.org/10.7227/RIE.86.4

Gorard, S. (2012a). The Increasing Availability of Official Datasets: Methods, Limitations and Opportunities for Studies of Education. *British Journal of Educational Studies,* 60(1), 77-92. https://doi.org/10.1080/00071005.2011.650946

Gorard, S. (2012b). Who Is Eligible for Free School Meals? Characterising Free School Meals as a Measure of Disadvantage in England. *British Educational Research Journal,* 38(6), 1003-1017. https://doi.org/10.1080/01411926.2011.608118

Gorard, S. (2014) The Widespread Abuse of Statistics by Researchers: What is the Problem and What is the Ethical Way Forward? *The Psychology of Education Review,* 38(1), 3-10.

Gorard, S. (2015). Rethinking 'Quantitative' Methods and the Development of New Researchers. *Review of Education,* 3(1), 72-96. https://doi.org/10.1002/rev3.3041

Gorard, S., & Cheng SC (2011). Pupil Clustering in English Secondary Schools: One Pattern or Several? *International Journal of Research and Method in Education*, 34(3), 327-339. https://doi.org/10.1080/1743727X.2011.609548

Gorard, S. & Smith, E. (2004). What is 'Underachievement'? *School Leadership and Management*, 24 (2), 205-225. DOI:10.1080/1363243041000695831

Gorard, S., Hordosy, R. and Siddiqui, N. (2013). How Unstable are 'School Effects' Assessed by a Value-added Technique? *International Education Studies,* 6(1), 1-9. DOI: 10.5539/ies.v6n1p1

Gray, J. (2004). School Effectiveness and the 'Other Outcomes' of Secondary Schooling: A Reassessment of Three Decades of British Research. *Improving Schools,* 7(2), 185-198. https://doi.org/10.1177/1365480204047348

Gray, J., Jesson, D., Goldstein, H., Hedger, K. & Rasbash, J. (1995). A Multi-Level Analysis Improvement: Changes in Schools' Performance Over Time. *School Effectiveness and School Improvement*, 6(2), pp.97-114. DOI: 10.1080/0924345950060201

Greany, T. (2017). Collaboration, Partnerships and System Leadership Across Schools. In P. Earley and T. Greany (Eds.), *School Leadership and Education System Reform* (pp. 56-65). Bloomsbury Publishing.

Guldemond and Bosker (2009). School Effects on Students' Progress - A Dynamic Perspective. *School Effectiveness and school improvement*, 20(2), 255-268. https://doi.org/10.1080/09243450902883938

Hanushek, E. A. (1986). The Economics of Schooling: Production and Efficiency in Public Schools, *Journal of Economic Literature*, 24(3), 1141-1177.

Hanushek, E. A. (1989). The Impact of Differential Expenditures on Student Performance. *Educational Research*, 66(3), 397-409. https://doi.org/10.3102/0013189X018004045

Harker, R. and Tymms, P. (2004). The Effects of Student Composition on School Outcomes. *School Effectiveness and School Improvement,* 15(2), 177-199. https://doi.org/10.1076/sesi.15.2.177.30432

Harlen, W. (2005). Trusting Teachers' Judgement: Research Evidence of the Reliability and Validity of Teachers' Assessment used for Summative Purposes. *Research Papers in Education,* 20(3), 245-270. https://doi.org/10.1080/02671520500193744

Harris, A. (2001). Building the Capacity for School Improvement. *School Leadership and Management*, 21(3), 261-270. DOI:10.1080/13632430120074419

Harris, D. N. (2009). Would Accountability Based on Teacher Value Added be Smart Policy? An Examination of the Statistical Properties and Policy Alternatives, *Education,* 4(4), 319-350. https://doi.org/10.1162/edfp.2009.4.4.319

Harrison, A. (2013, September 29). Michael Gove Acts to Block 'Damaging' Early GCSE Entry. BBC News. https://www.bbc.co.uk/news/uk-24326087.

Hattie, J. (2009). *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement.* Routledge.

Hauser, R. M. (1971). *Socioeconomic Background and Educational Performance*. American Sociological Association

Hauser, R. M., Sewell, W.H. and Alwin, D. F. (1976). High school effects of achievement. In Sewell, W. H., Hauser, R. M. and Featherman, D. L. (eds), *Schooling and Achievement in American society* (pp. 309-341). Academic Press.

Hayek, F. (1944). *The Road to Serfdom.* Routledge Press

Hayek, F. (1945). The use of knowledge in society. *The American Economic Review*, 35(4), 519-530.

Heck, R. A. and Thomas, S. L. (2000). *An Introduction to Multilevel Modeling Techniques*. Lawrence Erlbaum.

Hedges, L. V., Laine, R. D. and Greenwald, R. (1994). Does Money Matter? A Meta-Analysis of Studies of the Effects of Differential School Inputs on Student Achievement (An exchange: Part 1). *Educational Researcher*, 23(3), 5-14. https://doi.org/10.2307/1177220

Henderson, V., Mieszkowski, P., and Sauvageau, Y. (1978). Peer Group Effects and Educational Production Functions. *Journal of Public Economics,* 10(1), 97-106. https://doi.org/10.1016/0047-2727(78)90007-5

Hill, P. and Rowe, K. (1996). Multilevel Modelling in School Effectiveness Research. *School Effectiveness and School Improvement,* 7(1), 1-34. https://doi.org/10.1080/0924345960070101

House of Commons (2005). *June 2005 Debates* (Column WA151). Retrieved from https://publications.parliament.uk/pa/ld200506/ldhansrd/vo050620/text/50620w02.htm

House of Commons (2008). *Testing and Assessment; Third Report of Session 2007-08.* http://www.publications.parliament.uk/pa/cm200708/cmselect/cmchilsch/169/169.pdf

House of Commons (2009) *School Accountability; First Report of Session 2009-10.* http://www.publications.parliament.uk/pa/cm200910/cmselect/cmchilsch/88/88i.pdf

Hoxby, C. (2002). *The Economic Analysis of School Choice*. University of Chicago Press.

Howe, C. (2014). A Response to Gorard. *The Psychology of Education Review,* 38(1).

Hoyle, R., & Robinson, J. (2003). League tables and School Effectiveness: A Mathematical Model. Proceedings of the Royal Society of London B, 270, 113-199. http://dx.doi.org/10.1098/rspb.2002.2223

Hutchison, Dougal (2007). When Is a Compositional Effect Not a Compositional Effect? *Quality & Quantity*, 41(2), 219–232. DOI:10.1007/s11135-007-9094-2

Huttner H. J. M. and van de Eeden, P. (1995). *The Multilevel Design: A Guide with an Annotated Bibliography, 1980-1993.* Greenwood Press.

Ing, E. (2018, September 4). Vocational Qualifications, Progress 8 and 'Gaming'. Ofsted Blog: Schools, Early Years, Further Education and Skills. https://educationinspection.blog.gov.uk/2018/09/04/vocational-qualifications-progress-8-and-gaming/

Isaacs, T., Zara, C., Smith, C., Herbert, G. and Coombs, S. J. (2013). *Key Concepts in Educational Assessment.* Sage.

Jacob, B.A., L. Lefgren, and D. Sims (2010). The Persistence of Teacher-Induced Learning Gains. *The Journal of Human Resources*, 45 (4), 915-943. DOI:10.3368/jhr.45.4.915

Jencks, C., Smith, M., Acland, H., Bane, M. J., Cohen, D., Gintis, H., Heyns, B. and Michelson, S. (1972). *Inequality: A Reassessment of the Effects of Family and Schooling in America*. Basic Books.

Jesson, D. & Gray, J. (1991). Slants on Slopes: Using Multi-level Models to Investigate Differential School Effectiveness and its Impact on Pupils' Examination Results. *School Effectiveness and School Improvement*, 2(3), 230-247. DOI:10.1080/0924345910020305

Johnson, S. (2013). On the reliability of high-stakes teacher assessment, *Research Papers in Education,* 28(1), 91-105. DOI:10.1080/02671522.2012.754229

Kaliszewski, M., Fieldsend, A. and McAleavy, T. (2017). *England's Approach to School Performance Data – Lessons Learned.* Education Development Trust. https://files.eric.ed.gov/fulltext/ED586972.pdf

Kane, T. J. and Staiger, D. O. (2008). *Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation*: National Bureau of Economic Research. https://www.nber.org/papers/w14607

Kelly, A. and Downey, C. (2010). Value-Added Measures for Schools in England: Looking inside the 'Black Box' of Complex Metrics. *Educational Assessment, Evaluation and Accountability,* 22(3), 181-198. DOI:10.1007/s11092-010-9100-4

Kelly, A. and Downey, C. (2011) *Using Effectiveness Data for School Improvement: Developing and Utilising Metrics.* Routledge.

Kelly, S. & Monczunski, L. (2007). Overcoming the Volatility in School-Level Gain Scores: A New Approach to Identifying Value-Added with Cross-Sectional Data. *Educational Researcher*, 36(5), 279–287. DOI:10.3102/0013189X07306557

Knuver, A. and Brandsma, H. (1993). Cognitive and Affective Outcomes in School Effectiveness Research. *School Effectiveness and School Improvement,* 4(3), 189-204. https://doi.org/10.1080/0924345930040302

Konstantopoulos, S. (2007). *How Long Do Teacher Effects Persist?* Forschungsinstitut Zur Zukunft Der Arbeit. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1000142

Konstantopoulos, S. (2008). Do Small Classes Reduce the Achievement Gap between Low and High Achievers? Evidence from Project STAR. *The Elementary School Journal*, 108(4), 275-291. https://doi.org/10.1086/528972

Koretz, D. M. (2008). *Measuring Up: What Educational Testing Really Tells Us.* Harvard University Press.

Krueger, A.B. and Whitmore, D.M. (2001).The Effect of Attending a Small Class in the Early Grades on College Test Taking and Middle School Test Results: Evidence from Project STAR. *Economic Journal*, 111(468), 1–28. https://doi.org/10.1111/1468-0297.00586

Kurtz, M. D. (2018). Value-Added and Student Growth Percentile Models: What Drives Differences in Estimated Classroom Effects? *Statistics and Public Policy*, 5(1), 1-8, DOI: 10.1080/2330443X.2018.1438938

Kyriakides, L. (2004). Differential School Effectiveness in Relation to Sex and Social Class: Some Implications for Policy Evaluation. *Educational Research and Evaluation*, 10(2), 141-161, DOI: 10.1076/edre.10.2.141.27907

Kyriakides, L. (2005). Extending the Comprehensive Model of Educational Effectiveness by an Empirical Investigation. *School Effectiveness and School Improvement,* 16(2), 103-152. DOI:10.1080/09243450500113936

Kyriakides, L. and Charalambous, C. (2005). Using Educational Effectiveness Research to Design International Comparative Studies: Turning Limitations into New Perspectives, *Research Papers in Education*, 20(4), 391-412. DOI:10.1076/sesi.11.4.501.3560

Kyriakides, L. and Tsangaridou, L. (2008). Towards the Development of Generic and Differentiated Models of Educational Effectiveness: A Study of School and Teacher Effectiveness in Physical Education. *British Educational Research Journal*, 34(6), 807-838. DOI:10.1080/01411920802041467

Kyriakides, L., Campbell, R. J. and Christofidou, E. (2002). Generating Criteria for Measuring Teacher Effectiveness through a Self-Evaluation Approach: A Complementary Way of Measuring Teacher Effectiveness. *School Effectiveness and School Improvement*, 13(3), 291-325. DOI: 10.1076/sesi.13.3.291.3426

Kyriakides, L., Campbell, R. J. and Gagatsis, A. (2000). The Significance of the Classroom Effect in Primary Schools: An Application of Creemers' Comprehensive Model of Educational Effectiveness. *School Effectiveness and School Improvement*, 11(4), 501-529. DOI: 10.1076/sesi.11.4.501.3560

Kyriakides, L., Christoforou, C. and Charalambous, C. Y. (2013). What Matters for Student Learning Outcomes: A Meta-Analysis of Studies Exploring Factors of Effective Teaching. *Teaching and Teacher Education*, 36, 143-152. http://dx.doi.org/10.1016/i.tate.2013.07.010

Ladd, H. F. and Walsh, R. P. (2002). Implementing Value-Added Measures of School Effectiveness: Getting the Incentives Right. *Economics of Education Review,* 21(1), 1-17. DOI:10.1016/S0272-7757(00)00039-X

Lang, M. H., Teddlie C. and Oescher, J. (1992, April). *The Effect of Varying the Test Mode Had on School Effectiveness Ratings* [Conference session] AERA 2019 Convention, SanFrancisco, CA, USA. https://files.eric.ed.gov/fulltext/ED350314.pdf

Lavy, V., Silva, O. and Weinhardt, F. (2012). The Good, the Bad, and the Average: Evidence on Ability Peer Effects in Schools. *Journal of Labor Economics,* 30(2), 367-414. https://doi.org/10.1086/663592

Leckie, G. and Goldstein, H. (2009). The Limitations of Using School League Tables to Inform School Choice. *Journal of the Royal Statistical Society,* 172(4), 835-851. https://doi.org/10.1111/j.1467-985X.2009.00597.x

Leckie, G. and Goldstein, H. (2011). Understanding Uncertainty in School League Tables. *Fiscal Studies,* 32(2), 207-224. https://doi.org/10.1111/j.1475-5890.2011.00133.x

Leckie, G. and 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, DOI: 10.1002/berj.3264

Leckie, G. and 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), 518-538, DOI: 10.1002/berj.3511

Lee, V., & Bryk, A. (1989). A multilevel model of the social distribution of educational achievement. *Sociology of Education, 62(3),* 172–192. https://doi.org/10.2307/2112866

Leithwood, K. and Jantzi, D. (2006). Transformational School Leadership for Large-Scale Reform: Effects on Students, Teachers, and their Classroom Practices. *School Effectiveness and School Improvement*, 17(2), 201-227. DOI: 10.1080/09243450600565829

Leithwood, K., Leonard, L., Sharratt, L. (1998). Conditions Forstering Organizational Learning in Schools. *Educational Administration Quarterly*, 34(2), 243-276. https://doi.org/10.1177/0013161X98034002005

Levine, D. U. and Lezotte, L. W. (1990) *Unusually Effective Schools: A Review and Analysis of Research and Practice.* National Center for Effective Schools Research and Development

Lewis, C. and Tsuchida, I. (1997). Planned Educational Change in Japan: The Case of Elementary Science Instruction. *Journal of Educational Policy,* 12(5), 313-331. DOI:10.1080/0268093970120502

Linn, R. L., & Haug, C. (2002). Stability of School-Building Accountability Scores and Gains. *Educational Evaluation and Policy Analysis*, 24(1), 29–36. DOI:10.3102/01623737024001029

Lubienski, S., & Lubienski, C. (2006). School Sector and Academic Achievement: A Multi-Level Analysis of NAEP Mathematics Data. *American Educational Research Journal*, 43(4), 651-698. DOI:10.3102/00028312043004651

Luyten, H. (1994). Stability of School Effects in Dutch Secondary Education: The Impact of Variance Across Subjects and Years. *International Journal of Educational Research,* 21(2), 197-216. https://doi.org/10.1016/0883-0355(94)90032-9

Luyten, H. (2003). The Size of School Effects Compared to Teacher Effects: An Overview of the Research Literature. *School Effectiveness and School Improvement,* 14(1), 31-51. DOI:10.1076/sesi.14.1.31.13865

Luyten, H. and Sammons, P. (2010). Multilevel Modelling. In Creemers, B., Kyriakides, L. & Sammons, P. (Eds.), *Methodological Advances in Educational Effectiveness Research* (246-276). Routledge.

Luyten, H., Tymms, P. and Jones, P. (2009). Assessing School Effects Without Controlling for Prior Achievement? *School Effectiveness and School Improvement,* 20(2), 145-165. DOI:10.1080/09243450902879779

Luyten, H., Visscher, A. and Witziers, B. (2005). School Effectiveness Research: From a Review of the Criticism to Recommendations for Further Development. *School Effectiveness and School Improvement,* 16(3), 249-279. DOI:10.1080/09243450500114884

Mandeville, G. K. (1988). School Effectiveness Indices Revisited: Cross-Year Stability. *Journal of Educational Measurement*, 25(4), 349-356.

Mandeville, G. K. and Anderson, L. W. (1987). The stability of school effectiveness indices across grade levels and subject areas. *Journal of educational measurement,* 24(3), 203-216. https://doi.org/10.1111/j.1745-3984.1987.tb00275.x

Marks, G. N. (2015a). Are school-SES effects statistical artefacts? Evidence from longitudinal population data. *Oxford Review of Education,* 41(1), 122-144. https://doi.org/10.1080/03054985.2015.1006613

Marks, G. N. (2015b). The size, stability, and consistency of school effects: Evidence from Victoria. *School Effectiveness and School Improvement*, 26(3), 397–414. https://doi.org/10.1080/09243453.2014.964264

Marsh, H. W., Nagengast, B., Fletcher, J. and Televantou, I. (2011). Assessing Educational Effectiveness: Policy Implications from Diverse Areas of Research. *Fiscal Studies,* 32(2), 279-295. https://doi.org/10.1111/j.1475-5890.2011.00137.x

McCaffrey, D. F., Castellano, K. E. and Lockwood, J. R. (2015). The Impact of Measurement Error on the Accuracy of Individual and Aggregate SGP. *Educational Measurement: Issues and Practice,* 34(1), 15–21. DOI:10.1111/emip.12062.

McCaffrey, D. F. and Hamilton, L. S. (2007). *Value-Added Assessment in Practice: Lessons from the Pennsylvania Value-Added Assessment System Pilot Project.* RAND Corporation. https://www.rand.org/content/dam/rand/pubs/technical_reports/2007/RAND_TR506.pdf

McCaffrey, D. F., Lockwood, J. R. Koretz. D., Louis. T. A., Hamilton L. (2004). Models for Value-Added Modeling of Teacher Effects. *Journal of Educational and Behavioral Statistics*, 29(1), 67-101. DOI: 10.3102/10769986029001067

McCaffrey, D., Sass, T., Lockwood, J., and Mihaly, K. (2009). The Intertemporal Variability of Teacher Effect Estimates. *Education Finance and Policy*, 4(4), 572-606. http://dx.doi.org/10.1162/edfp.2009.4.4.572

McDill, E. L., Rigsby, L. C., and Meyers, E. D. Jr. (1969). Educational Climates of High Schools: Their Effect and Sources. *American Journal of Sociology*, 74(6), 567-586. https://doi.org/10.1086/224711

Messick, S. (1989a). Validity. In R. L. Linn (Ed.), Educational Measurement (3rd edition, pp. 13-104). American Council on education and Macmillan.

Messick, S. (1989b). Meaning and Values in Test Validation: The Science and Ethics of Assessment. *Educational Researcher*, 18(2), 5-11. DOI:10.3102/0013189X018002005

Messick, S. (1995). Validity of psychological assessment: Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry into Score Meaning. *American Psychologist, 50*(9), 741–749. DOI:10.1037/0003-066X.50.9.741

Messick, S. (1996a, December). *Standards-based Score Interpretation: Establishing Valid Grounds for Valid Inferences*. Proceedings of the Joint Conference on Standard Setting for Large Scale Assessments, Washington, DC, United States. https://onlinelibrary.wiley.com/doi/epdf/10.1002/j.2333-8504.1994.tb01630.x

Messick, S. (1996b). Validity of Performance Assessment. In Philips, G. (Ed.). *Technical Issues in Large-Scale Performance Assessment* (pp. 1-18). National Center for Educational Statistics.

Meyer, R. H. (1997). Value-added indicators of school performance: A primer, *Economics of Education Review.* 16(3), 283-301. https://doi.org/10.1016/S0272-7757(96)00081-7

Mintzberg. H. (1979). *The Structuring of Organizations*. Prentice Hall.

Monk, D. H. (1992). Education Productivity Research: An Update and Assessment of Its Role in Education Finance Reform. *Educational Evaluation and Policy Analysis*, 14(4), 307-332. https://doi.org/10.3102/01623737014004307

Morris, R. and Perry, T. (2016). Reframing the English Grammar Schools Debate. *Educational Review*, 69 (1), 1–24. DOI:10.1080/00131911.2016.1184132

Morris, R., Davies, N. M., Dorling, D., Richmond, R. C. and Smith, G. D. (2018). Testing the Validity of Value-Added Measures of Educational Progress with Genetic Data. *British Educational Research Journal*, 44(5), 725-747. DOI:10.1002/berj.3466

Mortimore, P., Sammons, P., Stoll, L., Lewis, D., and Ecob, R. (1988). *School Matters: The Junior Years*. Open Books.

Mortimore, P., Sammons, P., and Thomas, S. (1994). School Effectiveness and Value Added Measures. *Assessment in Education: Principles, Policy and Practice*, 1(3), 315–332. DOI: 10.1080/0969594940010307

Muijs, D. and Reynolds. D. (2000). School Effectiveness and Teacher Effectiveness in Mathematics: Some Preliminary Findings from the Evaluation of the Mathematics Enhancement Programme (Primary). *School Effectiveness and School Improvement*, 11(3), 273-303. DOI:10.1076/0924-3453(200009)11:3;1-G;FT273

Muijs, D., Kelly, T., Sammons, P., Reynolds, D. and Chapman, C. (2011). The Value of Educational Effectiveness Research: A Response to Recent Criticism. *Research Intelligence,* 114, 24-25.

Muijs, D., Kyriakides, L., van der Werf, G., Creemers, B., Timperley, H. and Earl, L. (2014) State of the Art – Teacher Effectiveness and Professional Learning. *School Effectiveness and School Improvement,* 25(2), 231-256. DOI:10.1080/09243453.2014.885451

Mulford, B. (2003). *School Leaders: Changing Roles and Impact on Teacher and School Effectiveness*. OECD. http://www.oecd.org/education/school/2635399.pdf

Muthen, B. and Khoo, S. (1998). Longitudinal Studies of Achievement Growth using Latent Variable Modeling. *Learning and Individual Differences*, 10(2), 73-101.

NAO (2003). *Making a Difference: Performance of Maintained Secondary Schools in England.* https://www.nao.org.uk/report/making-a-difference-performance-of-maintained-secondary-schools-in-england/

Nash, Roy. (2003). Is the School Composition Effect Real?: A Discussion With Evidence From the UK PISA Data. *School Effectiveness and School Improvement,* 14(4), 441–457, DOI:10.1076/sesi.14.4.441.17153.

Neale, D. (2015). Defending the Logic of Significance Testing: A Response to Gorard. *Oxford Review of Education*, 41(3), 334-345. https://doi.org/10.1080/03054985.2015.1028526

Newton, P. E. (2013). Ofqual's Reliability Programme: A Case Study Exploring the Potential to Improve Public Understanding and Confidence. *Oxford Review of Education,* 39(1), 93-113. https://doi.org/10.1080/03054985.2012.760285

Nor, M. Y. M. (2014). Potentials of Contextual Value-Added Measures in Assisting Schools Become More Effective. *International Education Studies,* 7(13), 75-91. oi:10.5539/ies.v7n13p75

Nuttall, D. L., Goldstein, H., Prosser, R. Rashbash, J. (1989). I*nternational Journal of Educational Research*, 13(7), pp.769-776.

Nye, B., Konstantopoulos, S. and Hedges, L. V. (2004). How Large Are Teacher Effects? *Educational Evaluation and Policy Analysis,* 26(3), 237-257. https://doi.org/10.3102/01623737026003237

OECD (2008) *Measuring Improvements in Learning Outcomes: Best Practices to Assess the Value-Added of Schools.* OECD Publishing.

Ofsted (2019) School Inspection Handbook. Retrieved from https://www.gov.uk/government/publications/school-inspection-handbook-eif

Olssen, M. & Peters, M. A. (2005). Neoliberalism, Higher Education and the Knowledge Economy: From the Free Market to Knowledge Capitalism. *Journal of Educational Policy*, 20(3), 313-345. DOI:10.1080/02680930500108718

Opdenakker, M. C. and Van Damme, J. (2000). Effects of Schools, Teaching Staff and Classes on Achievement and Well-Being in Secondary Education: Similarities and Differences Between School Outcomes. *School Effectiveness and School Improvement*, 11(2), 165–196. DOI:10.1076/0924-3453(200006)11:2;1-Q;FT165

Opposs, Dennis, and Qingping He. (2011). *The Reliability Programme: Final Report.* Ofqual. http://dera.ioe.ac.uk/2568/1/11-03-16-Ofqual-The-Final-Report.pdf

Oser, F. K. (1994). Moral perspective on teaching. *Review of Research in Education*, 20(1), 57-127. https://doi.org/10.3102/0091732X020001057

Palley, T. I. (2004). From Keynesianism to Neoliberalism: Shifting Paradigms in Economics. In Johnston & S. Filho (Eds.), *Neoliberalism: A Critical reader* (pp. 1-11). Pluto Press

Perry, T. (2016a). English Value-Added Measures: Examining the Limitations of School Performance Measurement. *British Educational Research Journal,* 42(6), 1056–1080. DOI:10.1002/berj.3247

Perry, T. (2016b). *The Validity, Interpretation and Use of Value-Added Measures* (unpublished PhD Thesis), University of Birmingham.

Perry, T. (2019) 'Phantom' Compositional Effects in English School Value-Added Measures: The Consequences of Random Baseline Measurement Error. *Research Papers in Education,* 34(2), 239-262. https://doi.org/10.1080/02671522.2018.1424926

Peterson, P., Wilkinson, L.C., and Hallinan, M. (eds) (1984). *The Social Context of Instruction: Group Organisation and Group Processes*. Academic Press.

Plewis, I., & Fielding, A. (2003). What is Multilevel Modelling For? A Critical Response to Gorard. *British Journal of Educational Studies*, 51(4), 408–418. https://doi.org/10.1046/j.1467-8527.2003.00246.x

Pokropek, A. (2015). Phantom Effects in Multilevel Compositional Analysis Problems and Solutions. *Sociological Methods and Research*, 44(4), 677-705. DOI:10.1177/0049124114553801

Popper, K. (2005). *The Logic of Scientific Discovery.* Routledge

Rasbash, J., Leckie, G., Pillinger, R. & Jenkins, J. (2010). Children's Educational Progress: Partitioning Family, School and Area Effects, *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 173(3), 657–682. https://doi.org/10.1111/j.1467-985X.2010.00642.x

Raudenbush, S. W. (2004). What are Value-Added Models Estimating and What Does This Imply for Statistical Practice? *Journal of Educational and Behavioral Statistics*, 29(1), 121–129. DOI:10.3102/10769986029001121

Raudenbush, S. W. and Bryk, A. S. (1986). A Hierarchical Model for Studying School Effects, S*ociology of Education*, 59(1), 1-17. DOI:10.2307/2112482

Raudenbush, S. W. and Willms, J. (1995). The Estimation of School Effects. *Journal of Educational and Behavioral Statistics,* 20(4), 307-335. https://doi.org/10.2307/1165304

Raven, J. (1991). The Wider Goals of Education: Beyond the 3 Rs, *The Educational Forum*, 55(4), 343-363. https://doi.org/10.1080/00131729109335666

Ray, A. (2006). *School Value Added Measures in England. Paper for the OECD Project on the Development of Value-Added Models in Education Systems*. Department for Education and Skills. https://webarchive.nationalarchives.gov.uk/20110219221253/https:/consumption.education.gov.uk/publications/eOrderingDownload/RW85.pdf

Ready, D. D. (2013). Associations Between Student Achievement and Student Learning Implications for Value-Added School Accountability Models. *Educational Policy,* 27(1), 92-120. https://doi.org/10.1177/0895904811429289

Reezigt , G. J., Guldemond, H. & Creemers, B. P. M. (1999). Empirical Validity for a Comprehensive Model on Educational Effectiveness. *School Effectiveness and School Improvement*, 10(2), 193-216. DOI: 10.1076/sesi.10.2.193.3503

Reynolds, D. (2008). *Schools Learning from Their Best: The Within School Variation (WSV) Project*, Nottingham: NCSL.

Reynolds, D., Chapman, C., Kelly, A., Muijs, D. and Sammons, P. (2012). Educational Effectiveness: The Development of the Discipline, The Critiques, The Defence, and The Present Debate, *Effective Education,* 3(2), 109-127. DOI:10.1080/19415532.2011.686168.

Reynolds, D., Sammons, P., De Fraine, B., Van Damme, J., Townsend, T., Teddlie, C. and Stringfield, S. (2014). Educational Effectiveness Research (EER): A state-of-the-art review. *School Effectiveness and School Improvement,* 25(2), 197-230. DOI:10.1080/09243453.2014.885450.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, Schools and Academic Achievement. *Econometrica*, 73(2), 417–458. https://doi.org/10.1111/j.1468-0262.2005.00584.x

Rosenshine, B. (1976). Classroom Instruction. In N.L. Gage (Ed.), *The Psychology of Teaching Methods: The Seventy-Fifth Yearbook of the National Society for the Study of Education* (pp. 335-371). University of Chicago Press

Rosenshine, B. (1983). Teaching Functions in Instructional Programs. *Elementary School Journal*, 83(4), 335-351. DOI:10.1086/461321

Rosenshine, B. and Stevens, R. (1986). Teaching Functions. In M.C. Wittrock (Ed.), *Handbook of Research on Teaching* (3rd edition, pp. 376-391). MacMillan.

Rothstein, J. (2010). Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement. *The Quarterly Journal of Economics,* 125(1), 175-214. https://doi.org/10.1162/qjec.2010.125.1.175

Rubin, D. B., Stuart, E. A., & Zanutto, E. L. (2004). A Potential Outcomes View of Value-Added Assessment in Education. *Journal of Educational and Behavioral Statistics*, 29(1), 103–116. DOI:10.3102/10769986029001103

Rutter, M. (1983). School Effects on Pupil Progress: Research Findings and Policy Implications, *Child development*, 54(1), 1-29. https://doi.org/10.2307/1129857

Rutter, M., Maughan, B., Mortimore, P. and Ouston, J. (1979). *Fifteen Thousand Hours: Secondary Schools and their Effects on Children.* Open Books Publishing Ltd.

Sammons, P. (1996). Complexities in the Judgement of School Effectiveness. *Educational Research and Evaluation,* 2(2), 113-149. https://doi.org/10.1080/1380361960020201

Sammons, P., Hillman, J., & Mortimore, P. (1995). *Key Characteristics of Effective Schools: A Review of School Effectiveness Research.* Ofsted. https://files.eric.ed.gov/fulltext/ED389826.pdf

Sammons, P., Mortimore, P., & Thomas, S. M. (1996). Do Schools Perform Consistently across Outcomes and Areas. In J. Gray, D. Reynolds, C. Fitz-Gibbon, & D. Jesson (Eds.), *Merging Traditions: The future of research on school effectiveness and school improvement* (pp. 3 - 29). Cassell.

Sammons, P., Nuttall, D. and Cuttance, P. (1993). Differential School Effectiveness: Results from a Reanalysis of the Inner London Education Authority's Junior School Project data. *British Educational Research Journal,* 19(4), 381-405. DOI:10.1080/0141192930190407

Sandoval-Hernandez, A. (2008). School Effectiveness Research: A Review of Criticisms and some Proposals to Address Them. *Educate*, Special Issue, 31-44.

Saunders, L. (1999). A Brief History of Educational 'Value Added': How Did We Get To Where We Are? *School Effectiveness and School Improvement,* 10(2), 233-256.

Savery, J. R. and Duffy, T.M (1995). Problem Based Learning: An Instructional Model and its Constructivist Framework. *Educational Technology*, 35(5), 31-38.

SCAA (1994). *Value Added Performance Indicators for Schools.* School curriculum and Assessment Authority.

Scheerens, J. (1992). *Effective Schooling: Research, Theory and Practice.* Cassell.

Scheerens, J. (1993). Basic School Effectiveness Research: Items for a Research Agenda. *School Effectiveness and School Improvement*, 4(1), 17-36. https://doi.org/10.1080/0924345930040102

Scheerens, J. (2013). *What is Effective Schooling? A Review of Current Thought and Practice*. International Baccalaureate Organization. https://research.utwente.nl/en/publications/what-is-effective-schooling-a-review-of-current-thought-and-pract

Scheerens, J. and Bosker, R. J. (1997). *The Foundations of Educational Effectiveness*. Pergamon Press.

Scheerens, J. and Creemers, B. P. M. (1989). Conceptualizing School Effectiveness. *International Journal of Educational Research*, 13(7), 691-706. https://doi.org/10.1016/0883-0355(89)90022-0

Schmidt, W., Jakwerth, P. and McKnight, C. C. (1998). Curriculum Sensitive Assessment: Content *Does* Make a Difference, *International Journal of Educational Research*, 29(6), 503-527. https://doi.org/10.1016/S0883-0355(98)00045-7

Seidel, T. and Shavelson, R. J. (2007). Teaching Effectiveness Research in the Past Decade: The Role of Theory and Research Design in Disentangling Meta-Analysis Results. *Review of Educational Research*, 77(4), 454-499. DOI: 10.3102/0034654307310317.

Shadish, W. R., Cook, T. D. and Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Wadsworth Cengage learning.

Shavit, Y. and Williams, R. A. (1985). Ability Grouping and Contextual Determinants of Educational Expectations in Israel. *American Sociological Review*, 50, 62-73. DOI:10.2307/2095340

Siddique, N., Boliver, V. and Gorard, S. (2019). Reliability of Longitudinal Social Surveys of Access to Higher Education: The Case of Next Steps in England, *Social Inclusion*, 7(1), 80–89. DOI: 10.17645/si.v7i1.1631

Simons, R. J., van de Linden, J. and Duffy, T. (2000). New Learning: Three Ways to Learn in a New Ballence. *In R.J. Simons, J., van de Linden and T. Duffy (Eds), New Learning* (pp. 1-20). Springer.

Sirin, S.R. (2005). Socioeconomic Status and Academic Achievement: A Meta-Analytic Review of Research, *Review of Educational Research*, 75(3), 417-453. DOI:10.3102/00346543075003417

Smith, D., Tomlinson, S., Bonnerjea, L., Hogarth, T. and Thomes, H. (1989). *The School Effect: A Study of Multi-Racial Comprehensives*. Policy Studies Institute.

Snijders, T. and Bosker, R. (2011). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modelling.* 2nd edition. London: Sage.

Spielman, A. (2017, March 10). *Ofsted to Investigate Schools 'Gaming System' to Move Up League Tables*. The Guardian. https://www.theguardian.com/education/2017/mar/10/ofsted-to-investigate-schools-gaming-system-to-move-up-league-tables

Strand, S. (2006). Comparing the Predictive Validity of Reasoning Tests and National End of Key Stage 2 Tests: Which Tests Are The 'Best'? *British Educational Research Journal,* 32(2), 209-225. https://doi.org/10.1080/01411920600569073

Strand, S. (2010). Do Some Schools Narrow the Gap? Differential School Effectiveness by Ethnicity, Gender, Poverty, and Prior Achievement. *School Effectiveness and School Improvement,* 21(3), 289-314. https://doi.org/10.1080/09243451003732651

Strand, S. (2016). Do some schools narrow the gap? Differential School Effectiveness Revisited. *Review of Education*, 4(2), 107–144. https://doi.org/10.1002/rev3.3054

Stringfield, S. C. and Slavin, R. E. (1992). A Hierarchical Longitudinal Model for Elementary School Effects. In B.R.M. Creemeres and G.J Reezigt (eds), *Evaluation of Educational Effectiveness* (pp. 35-69). ICO.

Stringfield, S., Winfield, L. and Abt Associates (1994). *Urban and Suburban/Rural Special Strategies for Educating Disadvantaged Children: First Year Report*. US Department of Education. *https://files.eric.ed.gov/fulltext/ED369854.pdf*

Styles, B. (2014) A Response to Gorard. *The Psychology of Education Review,* 38(1), 20-21.

Summers, A. A. and Wolfe, B. L. (1977) Do schools make a difference? *American Economic Review*, 67(4), 639-652.

Sutton Trust (2014). *Extra-curricular Inequalities.* https://www.suttontrust.com/our-research/enrichment-brief-private-tuition-extracurricular-activities/

Sutton Trust, Coe R., Jones K., Searle J.,Kokotsaki D., Kosnin A. M., Skinner P. (2008) *Evidence on the Effects of Selective Educational Systems. A Report for the Sutton Trust.* CEM Centre. https://www.suttontrust.com/our-research/evidence-effects-selective-educational-systems/

Teddlie, C. and Reynolds, D. (2000). *The International Handbook of School Effectiveness Research*. Farmer Press.

Tekwe, C. D., Carter, R. L., Ma, C., Algina, J., Lucas, M. E., Roth, J., Fisher, M. A. T. and Resnick, M. B. (2004). An Empirical Comparison of Statistical Models for Value-Added Assessment of School Performance. *Journal of Educational and Behavioral Statistics*, 29(1), 11–36. DOI:10.3102/10769986029001011

Televantou, Marsh, H. W., Kyriakides, L., Nagengast, B., Fletcher, J. & Malmberg, L. (2015). Phantom Effects in School Composition Research: Consequences of Failure to Control Biases Due to Measurement Error in Traditional Multilevel Models. *School Effectiveness and School Improvement*, 26(1), 75-101,DOI: 10.1080/09243453.2013.871302

Telhaj, S., Adnett, N., Davies, P., Hutton, D. and Coe, R. (2009). Increasing Within-School Competition: A Case for Department Level Performance Indicators? *Research Papers in Education,* 24(1), 45-55. https://doi.org/10.1080/02671520701809858

TGAT (1988) *National Curriculum: Task Group on Assessment and Testing, A Report.* Department of Education and Science *and the Welsh Office.* http://www.educationengland.org.uk/documents/pdfs/1988-TGAT-report.pdf

Thomas, S. (2001) Dimensions of secondary school effectiveness: Comparative analyses across regions. *School Effectiveness and School Improvement,* 12(3), 285-322. DOI:10.1076/sesi.12.3.285.3448

Thomas, S. & Mortimore, P. (1996) Comparison of Value-Added Models for Secondary-School Effectiveness. *Research Papers in Education*, 11(1), 5-33, DOI:10.1080/0267152960110103

Thomas, S., Peng, W. J., Gray, J., Thomas, S., Peng, W. J. and Gray, J. (2007). Modelling Patterns of Improvement over Time: Value Added Trends in English Secondary School Performance across Ten Cohorts. *Oxford Review of Education,* 33(3), 261-295. https://doi.org/10.1080/03054980701366116

Thomas, S., Sammons, P., Mortimore P. and Smees R. (1997a). Differential Secondary School Effectiveness: Comparing the Performance of Different Pupil Groups. *British Educational Research Journal*, 23(4), 451-469. https://doi.org/10.1080/0141192970230405

Thomas, S., Sammons, P., Mortimore, P. and Smees, R. (1997b). Stability and Consistency in Secondary Schools' Effects on Students' GCSE Outcomes over Three Years. *School Effectiveness and School Improvement,* 8(2), 169-197. https://doi.org/10.1080/0924345970080201

Thomas, S., Smees, R., MacBeath, J., Robertson, P. and Boyd, B. (2000). Valuing Pupils Views in Scottish Schools. *Educational Research and Evaluation,* 6(4), 281-316. DOI:10.1076/edre.6.4.281.6934

Thompson, J. D. (1967). Organizations in Action. McGraw Hill.

Timmermans, A. C., and Sally M. T. (2014). The Impact of Student Composition on Schools' Value-Added Performance: A Comparison of Seven Empirical Studies. *School Effectiveness and School Improvement*, 26(3), 487–498. DOI:10.1080/09243453.2014.957328

Timmermans, A. C., Doolaard, S. and de Wolf, I. (2011). Conceptual and Empirical Differences among Various Value-Added Models for Accountability. *School Effectiveness and School Improvement*, 22(4), 393-413. DOI:10.1080/09243453.2011.590704

Torres, R. T. and Preskill, H. (2001). Evaluation and Organizational Learning: Past, Present and Future. *American Journal of Evaluation*, 22(3), 387-395. DOI:10.1177/109821400102200316

Trafimow, D. and Rice, S. (2009) A Test of the Null Hypothesis Significance Testing Procedure Correlation Argument. *The Journal of General Psychology,* 136(3), 261-270. DOI:10.3200/GENP.136.3.261-270

Tymms, P. (1996) Theories, Models and Simulations: School Effectiveness at an Impasse. In J. Gray, D. Reynolds, C. Fitz-Gibbon, & D. Jesson (Eds.), *Merging Traditions: The Future of Research on School Effectiveness and School Improvement* (pp. 121-135). Cassell.

Tymms, P. , Merrell, C., Heron,T., Jones, P., Albone, S. & Henderson, B. (2008). The Importance of Districts. *School Effectiveness and School Improvement*, 19(3), 261-274, DOI:10.1080/09243450802332069

Valverde, G. A. and Schmidt, W. H. (2000). Greater Expectations: Learning from Other Nations in the Quest for 'World-Class Standards' in US School Mathematics and Science. *Journal of Curriculum Studies*, 32(5), 651-687. DOI:10.1080/00220270050116932

van der Werf. G, Opdenakker, M. & Kuyper, H. (2008). Testing a Dynamic Model of Student and School Effectiveness with a Multivariate Multilevel Latent Growth Curve Approach. School Effectiveness and School Improvement, 19(4), 447-462, DOI:10.1080/09243450802535216

van Ewijk, R., and Sleegers, P. (2010). Peer Ethnicity and Achievement: A Meta-Analysis into the Compositional Effect. *School Effectiveness and School Improvement*, 21(3), 237-265. https://doi.org/10.1080/09243451003612671

Visscher, A. J. (2001). Public School Performance Indicators: Problems and Recommendations. *Studies in Educational Evaluation,* 27(3), 199-214. https://doi.org/10.1016/S0191-491X(01)00026-8

Walberg, H.J. (1986). Syntheses of Research on Teaching. In M.C. Wittrock (ed.), *Handbook of Research on Teaching* (3rd edition, pp. 214-229). Macmillan.

Wang, M. C., Haertel, G. D. and Walberg, H. J. (1993). Towards a Knowledge Base for School Learning. *Review of Educational Research*, 63(3), 249-294. https://doi.org/10.2307/1170546

Wentzel, K. R. and Wigfield, A. (1998). Academic and Social Motivation Influences on Students' Academic Performance. *Educational Policy Review*, 10(2), 155-175. DOI:10.1023/A:1022137619834

West, A. (2010). High Stakes Testing, Accountability, Incentives and Consequences in English Schools. *Policy and Politics*, 38 (1), 23–39, DOI:10.1332/030557309X445591

West, A. & Pennell, H. (2000). Publishing School Examination Results in England: Incentives and Consequences. *Educational Studies*, 26(4), 423–436. DOI:10.1080/03055690020003629

White, P. (2014). A Response to Gorard, *The Psychology of Education Review,* 38(1), 24-28.

Wiliam, D. (2001). *Level Best? Levels of Attainment in National Curriculum Assessment*. Association of Teachers and Lecturers. https://www.dylanwiliam.org/Dylan_Wiliams_website/Papers_files/Level%20best%20-%20Levels%20of%20attainment%20in%20the%20national%20curriculum%20%28ATL%202001%29.pdf

Willms, J. D. (1985). The Balance Thesis: Contextual Effects of Ability on Pupils' O-Grade Examination Results. *Oxford Review of Education*, 11(1), 33-41, DOI:10.1080/0305498850110103

Willms, J. D. (1986). Social Class Segregation and its Relationship to Pupils Examination Results in Scotland. American Sociological Review, 51(2), 224-241. https://doi.org/10.2307/2095518

Willms, J. D. (1992) *Monitoring School Performance: A Guide for Educators*. Falmer Press.

Willms, J. D. (2003) *Monitoring School Performance: A Guide for Educators*. Routledge.

Willms, J. D. and Raudenbush, S. W. (1989). A longitudinal hierarchical linear model for estimating school effects and their stability. *Journal of Educational Measurement*, 26(3), 209-232. https://doi.org/10.1111/j.1745-3984.1989.tb00329.x

Wilson, D. (2009). Exit, Voice and Quality in the English Education Sector. *Social Policy & Administration*, 43(6), 571-584, DOI:10.1111/j.1467-9515.2009.00681.x

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. DOI:10.1080/10967490802301336

Wilson, D., Croxson, B. & Atkinson, A. (2006). What Gets Measured Gets Done: Headteachers' Responses to the English Secondary School Performance Management System, *Policy Studies*, 27(2), 153–171, DOI:10.1080/01442870600637995

Witzier, B., Bosker, J. R. and Kruger, L. M. (2003). Educational Leadership and Student Achievement: The Elusive Search for an Association. *Educational Administration Quarterly,* 39(3), 398-425. https://doi.org/10.1177/0013161X03253411

Woodhouse, G. and Goldstein, H. (1988). Educational Performance Indicators and LEA League Tables. *Oxford Review of Education*, 14(3), 301-320.

Woodhouse, G., Yang M., Goldstein H., and Rasbash J., (1996). Adjusting for Measurement Error in Multilevel Analysis. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 195(2), 201–212. https://doi.org/10.2307/2983168

Yair, G. (1997). When Classrooms Matter: Implications of Between-Classroom Variability for Educational Policy in Israel. *Assessment in Education: Principles, Policy & Practice*, 4(2), 225-248, DOI: 10.1080/0969594970040202