THE VALIDITY, INTERPRETATION AND USE OF SCHOOL VALUE-ADDED MEASURES

By
Thomas Perry

A thesis submitted to the
University of Birmingham
for the degree of
Doctor of Philosophy

School of Education
College of Social Sciences
University of Birmingham
February 2016

Abstract

This thesis examines the validity of school value-added measures and the validity of arguments for their interpretation and use. The opening chapters review the development of school value-added measures, existing evidence on their properties and validity and their current use in research, policy and practice.

The empirical results are based on four studies using English National Pupil Database data and a large, nationally-representative dataset of teacher-assessed attainment data for English pupils aged from 7 to 13. The findings all relate to the properties of school value-added measures and the seriousness of a number of threats to their validity. The four empirical studies examine the following issues: observable bias and error, inter-method reliability when compared to estimates from a quasi-experimental regression discontinuity design, stability of school value-added scores and of specific cohorts over time, and consistency of school value-added scores within cohorts and between different school cohorts at a single point in time.

The closing chapters discuss the validity of value-added measures in general and in relation to the areas of use identified. Individually and collectively, the results advance understanding of numerous threats to validity and have substantial implications for the use of value-added measures in research, policy and practice.

To my beloved family

Elaine, Evelyn and 'Bump'

Acknowledgements

My supervisors, Peter Davies and Stephen Gorard, have been immensely influential in shaping the ideas in this thesis and my development as a researcher more generally. I truly appreciate all their generous support and valuable advice. I consider myself incredibly lucky to have had two brilliant supervisors guiding my research.

I would also like to thank my friends and family whose patience, interest and constant support have kept me going throughout my studies. I am especially grateful to my wife Elaine and daughter Evelyn who have managed to keep things in perspective and a smile on my face; my friends, Becky Morris and Keir Irwin Rogers, for always showing an interest in my research; and my parents and parents-in-law for always being there to support me.

Finally, I would like to thank the ESRC who, in funding my research, have given me an opportunity to explore my interests and pursue a career seeking answers to important questions.

Table of Contents

List of Figures

List of Tables

List of Appendices

1. Introduction and Summary

1.1 Chapter Introduction

This chapter introduces the core problem examined in this thesis, the issues surrounding it, how this thesis aims to contribute to understanding in these areas and the organisation of the thesis.

1.2 Thesis Introduction

1.2.1 Summary of Thesis Topic and Contribution

This research concerns the validity of value-added measures of school effectiveness. Value-added models are used extensively in educational research, accountability systems and policy-making to estimate school performance. Value-added evidence is used to inform and underpin myriad research findings, policy decisions and high-stakes school performance judgements. It is cause for concern then that there are theoretical and empirical grounds to doubt whether school value-added scores provide valid and unbiased measures of the causal effect of schools on their pupils (Coe and Fitz-Gibbon, 1998, Gorard, 2010, Marsh et al., 2011). Moreover, there are difficulties even specifying what would constitute measurement validity in relation to value-added measures given that school effectiveness is operationalised as a latent, unobservable property of schools which is 'revealed' by the value-added statistical procedure itself (Gorard et al., 2012, p.3). Interpretations of the evidence generated using a value-added method are underpinned by many assumptions about what constitutes error, bias and effect within the data and the extant evidence base does not provide a definitive test of these assumptions. It is within this context of undetermined validity that this study submits its core research question: Are school value-added measures valid measures of school effectiveness? Currently, researchers have strikingly different views on this question and there are numerous points of contention within debates about the validity of school value-added measures and the value-added method more generally.

There are two fundamental arguments running through this thesis which provide an answer to this core research question: First, in the absence of a definitive test of these assumptions, examining the validity of value-added involves drawing on numerous sources of evidence pertaining to bias, error, stability and consistency within the measure. By examining these sources of evidence it is possible to identify specific validity problems and create approximate bounds on what is reasonable to conclude about the validity of particular value-added measures and value-added measures in general. This thesis presents original evidence organised within four empirical studies which update, extend and advance what is known about the properties and validity of value-added measures. The findings have a number of serious implications for the use of value-added in general and the English official school value-added measure in particular and form part of the answer to the core research question.

Even with the best available evidence, such as that reviewed and presented, there is still considerable scope for differing interpretations of the available evidence. As a result, and drawing on recent work in measurement validity (notably Kane, 2013), the second fundamental argument in this thesis is that it is valuable for debates about the validity of value-added to take place in relation to specific interpretations and uses, in addition to drawing on more general validity evidence. Drawing conclusions about the validity of value-added in general ignores important distinctions which are needed about the specific data, value-added measure, its interpretation and its use. To this end, Chapter 3 reviews the current use of value-added across educational effectiveness research, English policy and English practice. This allows the final chapters to discuss results and reach conclusions about the validity of school value-added in relation to various applications and the relevant differences between them.

The key contribution of this thesis to this debate - its ‘value-added’ – is that it brings together, updates, advances, synthesises and evaluates a large range of evidence and many fundamental methodological ideas within what is a large area of enquiry. This is an ambitious undertaking which involves applying the recent methodological advances in the field as well as scrutinising its most basic foundations. This range and depth of study is what is held to be required to advance what is a complex, longstanding and unresolved issue.

1.3 Introduction to the Area of Study

1.3.1 Scope and Focus

This section specifies and justifies the focus and scope which has been chosen for this particular study within the broader topic of value-added.

The most important distinction concerns the level of analysis. Value-added is a methodological approach to isolating and estimating the effect of one factor independently of other known influences. Its initial applications to education were to estimate the effectiveness of schools (Saunders, 1999). The methodology, however, can be adapted to estimate the effect of other levels within the education system such as classes, teachers, departments and local area authorities. In different national and policy contexts, the unit of analysis which is emphasised varies. The UK government, for example, has historically seen the school as the key level at which performance is judged publically, using school 'league tables' and school value-added measures (Acquah, 2013). In contrast, the USA has seen numerous states take teachers as the key unit of analysis, creating a demand for teacher-level value-added measures (Harris, 2009). Use of value-added in research tends to place less emphasis on school level variation due to the prevailing view that teacher effects outweigh school effects (Muijs et al., 2014).

This study focuses on school-level value-added measures. This decision was taken primarily because, in the English system, the school is often taken as the key organisational unit at which many decisions are taken by policy-makers, parents and schools themselves (Chapman et al., 2011). The use of value-added in the English accountability context and the use of the official value-added measures by school leaders and governors is arguably the most consequential application of value-added measures in England. There are also more pragmatic reasons for focusing on the school level. Secondary data sources at pupil and school levels are more readily available in the UK through the National Pupil Database (NPD) as well as other sources. This allows the use of relatively high-quality data which are used by policy makers and by schools themselves.

The second decision made regarding the scope and focus of the research is where to situate the research on a continuum from technical-theoretical properties of school value-added all the way to the practical-empirical aspects. This study aims for some interim position with sufficient scope to address fundamental theoretical questions but to do so in relation to empirical evidence and specific practical contexts. Despite its breadth, this interim position excludes concerns at either end of this spectrum. Specifically, while all key mathematical models are formally specified and discussed, this thesis does not attempt an extended treatment of the theoretical mathematical foundations of the statistics involved in value-added measurement. At the more practical end, this study does not present new evidence on the public or professional usage of value-added and the wider consequences of its use. Yet, as described above, this thesis argues that it is of value to consider the validity arguments for the interpretation and use (Kane, 2013) of value-added all the way from consideration of the specific dataset and model to the interpretations and uses made on the basis of value-added evidence. Without further study of how value-added measures are used and interpreted by users in practice some of the details of how and if appropriate interpretation and use is achieved in practice cannot be addressed here.

A third and final factor focusing this research is the countries from which evidence is drawn. In relation to policy, the focus is predominately on the English education system. While international evidence is brought in during the review of the literature and many of findings and issues which are discussed have international relevance, as only English data are used, there will be some limit to the degree with which the findings can be directly applied to a more international context. With regards to the use of value-added in research, the focus is not confined to the English context and instead uses two journals as a reference point for educational effectiveness research: namely, *School Effectiveness and School Improvement* and *Journal of Research on Educational Effectiveness.* While extensive literature searches encompassed relevant publications from other sources, these journals are held to be indicative of the research use of value-added and are therefore a point of reference for this study.

1.3.2 Value-Added and the Question of its Validity

School effects are not a readily-observable, manifest quality of schools. Schools work to achieve numerous broad educational aims in a diverse range of contexts, dealing with particular circumstances, varying resources and with vastly differing pupil intakes. Differences between schools' intakes such as their prior attainment levels or in contextual factors such as economic disadvantage are problematic due to the strong association between these and pupils' subsequent achievement (Teddlie and Reynolds, 2000). Faced with this complexity, drawing conclusions about school performance from ‘raw’ attainment scores is highly problematic and is likely to reveal more about intakes than school performance. Early school effectiveness researchers set out to discover whether some schools could be shown to be ‘better’ than others. This apparently modest task is in fact surprisingly difficult to achieve. The general approach was to identify apparently similar pupils attending different schools and then compare their outcomes. If pupils with the same baseline results and similar characteristics were found to have markedly different outcomes, this would suggest that there are differences between how effective schools are. Early school effectiveness studies such as Rutter (1983) found that appreciable differences between schools remained even after taking major differences in intake into account and so concluded that there were differences between the effectiveness of schools.

The approach taken by studies to produce estimates is to construct a statistical expectation for each pupil depending on their previous performance and measured characteristics. Through this expectation, the actual pupil performance can be compared to statistically-similar pupils. A (positive or negative) value-added score is then created from the difference between this expectation and the pupil’s actual performance. The school mean of these pupil scores is calculated (subsequently or concurrently within a multi-level model) in order to create a measure of school performance (Ray, 2006). This captures the average difference in performance between the pupils at the school and that of statistically-similar pupils nationally. The assumption is generally made that this unexplained difference is causally attributable to the performance of the school because other known extraneous factors have been controlled when creating the statistically expected performance (Teddlie and Reynolds, 2000). Moreover, as the unexplained difference is thought to be attributable to the school (subject to measurement error), it is considered possible to learn about the properties of the school effect from studying this residual variation, as is discussed below.

The validity of the value-added measure as a measure of school performance, to a large extent, hinges on how valid this causal assumption is (Coe and Fitz-Gibbon, 1998). As Marsh et al. (2011, p.283) comment, the most basic assumption behind the prevailing approaches in the field is that statistical “models appropriately control for pre-existing differences so that [value-added] estimates reflect the effects of the teacher or school being evaluated and not the effects of prior schools, prior teachers or other pre-existing differences". Without this assumption, much of the existing evidence base can only be considered descriptive (with no attribution of causation) and practical conclusions become more unclear and uncertain as a result (Marsh et al., 2011). Without experimental evidence it is difficult to reach a strong position on the question of causality (Fitz-Gibbon, 1997). This problem has long been recognised by educational effectiveness researchers. For instance, Rutter (1983, p.12) noted the limitations of non-experimental evidence, but argued that research comparing pupil intakes, school characteristics and pupil outcomes across schools and over time provided 'strong circumstantial evidence' of differing school effectiveness.

This brings us to a key issue addressed in this thesis: that of justification. If value-added measures and associated estimates of school effects were valid (or meaningless), how would we know? Is this a question that can be resolved through 'circumstantial evidence'? In large part, the difficulty resolving this question stems from the nature of the value-added method: the school effect is cast as a latent property which is 'revealed' through the value-added analysis itself (Gorard et al., 2012). Key threats to validity such as the confounding effects of unobserved variables and measurement error are, by their nature, difficult to rule out. Many non-school factors which influence learning go unmeasured (Dearden et al., 2011b) and may even be unmeasurable, at least in a practical context (Tymms, 1996). All this means that it is very difficult to know the true causes for the differences captured in value-added evidence and, crucially, to what extent these differences are causally attributable to schools.

The problem of misplaced causal attribution (i.e. mistaking error or bias for school effect) is one educational effectiveness researchers have guarded against by stressing the need to consider the stability, consistency and statistical significance of effects. For instance, in response to criticism of the field in Gorard (2010), Muijs et al. (2011, pp.3-4) make the following points:

> "...it is only possible to distinguish groups of schools where pupil progress is significantly better or worse than predicted on a given outcome measure.
>
> Researchers have repeatedly emphasised that school effectiveness is a relative and retrospective concept that is both outcome- and time-dependent, and that as a consequence there is a need to study consistency, stability and differential effectiveness covering variations in different outcomes, including departmental effects for secondary schools, and trends over time and for different groups of pupils (Luyten and Sammons, 2010, Creemers et al., 2010)"
>
> (Muijs et al., 2011, pp.3-4)

Similarly, Teddlie and Reynolds (2000, p.116) state that 'if the results from difference measures of school effectiveness are consistent, then the researcher may conclude that the school is effective (or ineffective) with some confidence… [Otherwise] the researcher faces a dilemma in interpretation of the overall effectiveness status of the school.'

Yet, a large area of school effectiveness research into 'methodological issues' (Teddlie and Reynolds, 2000, p.49) examines such variation, not as a way of establishing confidence, but in order to further understand the properties of the school *effect* itself. So, for example, one can examine whether schools are differentially effective with different groups of pupils, or across different subject areas (Sammons et al., 1996). By the turn of the century, such study had led to the consensus that school effectiveness should be considered 'multi-faceted' with Thomas (2001, p.285) suggesting at least 4 dimensions are apparent: outcomes, pupil groups, cohorts and curriculum stages. More recent research continues to emphasise the inherent complexity of educational effects (Chapman et al., 2015). But returning to the problem of justification, at what point does instability or inconsistency reduce confidence (a word not used in its statistical sense) in the effect rather than merely revealing its properties? Critics point out that ascribing such 'complexity' (Sammons, 1996, p.143) to the school effect and accommodating this within similarly complex multi-level models is 'to assume from the outset that which the modelling is supposed to be seeking or testing' (Gorard, 2010, p.756). A common justification supporting the assumption that differences reflect effect rather than error given by prominent educational effectiveness researchers it that ‘there is so much independent agreement on the size of school effects, their scientific properties, the factors responsible for them, and the ways they can be utilised for school improvement’ (Reynolds et al., 2012, p.12). This response essentially appeals to the consistency of findings and their face validity (Isaacs et al., 2013) as justification for their validity, or what Rutter (1983, p.12) called ‘circumstantial evidence’ (see above).

Although the face validity justification is weak, in the absence of more robust tests (see below), it is entertained in this thesis which presents empirical evidence examining the stability and consistency of value-added measures, in addition to evidence on bias and error. Maintaining that inconsistency between pupil performances within a school is effect rather than error inevitably leads to a shift in our understanding of the nature of the school effect itself: Inconsistency must either stem from error or from differential school effectiveness. Similarly, instability must stem from changes in school performance or from error. So if one is to maintain the assumption that what is observed is effect rather than error, it is also necessary to accept the properties of the school effect which are suggested through examination of its stability and consistency. If these cannot be accepted, the face validity justification does not hold. In this way, the presentation of stability and consistency evidence in this thesis must either contribute to our understanding of school effects or give a better position from which to judge whether value-added methods are valid at face.

What can be concluded, however, if the face validity justification is upheld or underdetermined by the evidence? Unfortunately, there are no direct alternatives to value-added to address the particular problem it seeks to solve (context-independent comparison of school performances) and so it is difficult to validate value-added scores using an alternative approach (Gorard et al., 2012). There are few options: it is unfeasible to estimate school effects experimentally due to the ethical aspect of random allocation of pupils to schools (Goldstein and Spiegelhalter, 1996), although relevant studies in teacher-level randomisation (e.g. Nye et al., 2004) are considered in Chapter 3. Design-based approaches to tackling these confounding variables and estimating their significance and impact (such as those created by ‘natural experiments’ like policy changes) are rare and are not workable more widely as policy tools. It is difficult to even find suitable alternative measures which can be used as a source of comparison. Other sources of school performance information (e.g. school inspection reports), for instance, have severe limitations as sources of confirmation that value-added is capturing differences in school performance as is claimed.

The lack of a definitive test of validity and differing methodological assumptions have led researchers and users to reach markedly different conclusions regarding the robustness and validity of value added evidence and the conclusions warranted from it. As Gorard (2010, p.746) points out, profound differences in interpretation arise even on the most fundamental issues despite interpretations being ‘based on pretty much the same evidence’. To understand the problem examined in this thesis, it is essential to recognise that it is partly a problem of interpretation of evidence. On the one hand, researchers are using value-added measures to identify more or less effective schools in order to better understand the (school and non-school) factors which influence rates of pupil progress (e.g. Strand, 2014a, Sammons, 2014, Chapman and Muijs, 2013). On the other hand, on the basis of numerous methodological studies and papers, Gorard (2011c, p.26) concludes that (contextualised) value-added is ‘meaningless for any practical purpose’ and Marsh et al. (2011, p.286) describe value-added as being ‘based on some problematic statistical assumptions’, having large standard errors, as ‘not particularly reliable or stable over time’, ‘not particularly useful for formative purposes of improving effectiveness’ and yielding ‘fragile’ causal inferences.

An important part of the task of evaluating the validity of value-added, therefore, is to understand how these profound differences of view arise. Do these highly conflicting positions reflect differences in the nature of value-added evidence considered, differences in use, differences in interpretations, methodological assumptions or some other factor? It is remarkable that, despite engaging in a series of debates - in person and through numerous publications - Gorard and prominent educational effectiveness researchers have essentially agreed to disagree and mutually accuse one another of making basic statistical errors (Reynolds et al., 2012). As has been noted, these differences arise from inspection of the same evidence. It does not seem likely that further empirical evidence (such as that presented in this thesis) will be sufficient to resolve these differences to any significant degree. As a result, part of the answer to the question of validity is held to be theoretical and philosophical in nature. It is therefore of great value to examine and evaluate the methodological assumptions surrounding the use of value-added and explore the source of these differences. This philosophical aspect of the problem is reflected in chapters such as Chapter 2, which considers the development and design of value-added measures, and Chapter 4 which scrutinises the logic of value-added through reviewing and evaluating how different researchers have interpreted value-added evidence. These aspects of the thesis complement the empirical evidence which is reviewed in Chapter 4 and the original evidence presented in Chapter 6. This empirical evidence advances understanding of the properties and threats to validity in value-added evidence but it requires a thorough consideration of the difficulties of interpretation in order to make use of it to enhance our understanding of the validity of value-added.

In summary, the debates outlined above suggest that the validity of value-added measures is a current, valuable and open question for study. Much of the available evidence reviewed has been evidence of reliability or consistency rather than direct validity evidence. This advances our understanding but is not sufficient to address the question of validity. As a result, as well as presenting new empirical evidence, this research brings both philosophical and technical understanding to bear on the question of the validity of value-added and in doing so is able to offer an original and productive contribution to the debate.

1.4 Summary of Chapters

1.4.1 Chapter Aims and Overview

A summary of the main purpose of each chapter, the key points covered and the key content is given in Table 1.4.1a, below:

Table 1.4.1a – Chapter Content Summary

Chapter 2 – The Aims and Designs of Value-Added Measures

Chapter 2 examines the motivations for the development of value-added measures. It serves as an introduction to the value-added method and the distinctions between value-added measures and value-added evidence drawn on in Chapter 3 and 4.

- Introduces what value-added measures are in general terms, detailing the origins of the term *value-added* in education and distinguishes this from other uses.
- Describes a range of statistical approaches which could be characterised as using the value-added method, giving details of various outputs which can be obtained to produce what is referred to in the thesis as ‘value-added evidence’.
- Highlights several variations on and extensions to the value-added models introduced.

Chapter 3 – Value-Added in Research, Policy and Practice

This section reviews the current use of value-added in educational effectiveness research, English policy and English practice. This draws on the technical details presented in Chapter 2 to discuss the application of the value-added method in each area of use. The chapter provides the foundations for future chapters such as the discussion chapter which explores the implications of the results presented in this thesis in relation to how it is currently used.

- Conducts a survey of the use of the value-added method between January 2013 and mid-April 2015 in two educational effectiveness journals.
- Gives examples of the research use of value-added, drawing on the information given in Chapter 2 and the methodological survey of educational effectiveness research (above).
- Describes the development of an English value-added measure as a national performance measure and the political climate in which this took place. Including an introduction to the forthcoming English ‘Progress’ value-added measures.
- Describes the use of value-added in school practice in England, considering attitudes to data and its use and what is considered to be best practice for using value-added evidence.

Chapter 4 – The Validity of School Value-Added Measures

This chapter is the key thesis literature review. It reviews the existing evidence base to which the results presented in this thesis are designed to contribute. This is organised into two main sections:

- First, a review of methodological study of school value-added, organised around a number of threats to validity and methodological issues which impinge on validity.
- Second, a section which examines the interpretation of value-added evidence. Key debates around interpretation and uncertainty are reviewed and evaluated.

There is a final section in this chapter which is somewhat separate to main review sections before it and relates to alternative measures of school effectiveness. This section introduces the regression discontinuity design and reviews its use for estimating school effects. This section provides the groundwork for one of the empirical studies in this thesis which compares estimates of school effectiveness produced using a value-added design with estimates from the quasi-experimental regression discontinuity design

Chapter 5 – Methods

The methods chapter provides crucial information for the empirical results in the next chapter. It is designed as an overview for the four studies rather than a detailed description of the analyses to follow. Detailed description of the analytical approaches are contained within the results chapter along with the actual results; descriptive statistics and equations for the statistical models used are placed in an appendix unless these are the main object of analysis. The key content within the methods chapter is as follows:

- A statement of the core research question, followed by the primary research questions in each study (see below) designed to address it.
- Description of the analytical approach taken within and across the four studies
- An introduction to the key data sources used in this thesis.
- An overview of each of the four empirical studies, further details of the research questions and of the specific data and sample used to address them.

The results chapter follows a question and answer format. The research questions are organised into four studies (see below). Each research question is presented along with a) an explanation of how the question relates to existing evidence (Chapter 4) and the core research question, b) details of the analytical steps taken to address the question and c) the results of the empirical analysis undertaken to address the research question. The four studies are as follows:

- *Study 1 - Bias and Error*, which examines sources of observable bias and error, primarily using the official English value-added measure;
- *Study 2 - Inter-Method Reliability*, which compares estimates produced using value-added with estimates produced using a regression discontinuity design;
- *Study 3* - S*tability over Time*, where the stability of English value-added measures over a number of years is examined; and
- *Study 4 - Cohort Consistency*, which examines the consistency of estimates for different cohorts within a school at a single point in time.

Each of these studies is designed to update, extend and/or advance what is known about the properties and validity of value-added measures. Some of these analyses replicate previous results, some address specific issues raised in the review of validity evidence by extending or adapting previous studies and others adopt a relatively novel approach, conducting original analyses or examining issues where there is very little existing evidence.

All the results are all based on analysis of two sources of data:

- First, extracts from the National Pupil Database. These extracts cover 2004 to 2014 and contain data at pupil-level and school-level across the pupil age range for most but not all of these years.
- The second source of data is from a large Department for Education commissioned study known as ‘Making Good Progress’. This nationally representative dataset contains teacher-assessed data for pupils aged 7 to 13. This second dataset is used where measures of pupil performance between English Key Stage years are required.

Chapter 7 – Discussion

The discussion chapter has two main parts, each addressing a specific aim. The first aim is to discuss evidence pertaining to the core research question: 'Are school value-added measures valid measures of school effectiveness?' This involves synthesising the results presented in the Chapter 6 and discussing how the findings fit within the pre-existing evidence that was reviewed in Chapter 4.

The second aim of the discussion chapter is to consider the results alongside the various areas of use reviewed in Chapter 3. This draws on a conception of measurement validity in relation to interpretation and use (Kane, 2013); this conception is adapted to the specific context of value-added by outlining a theoretical framework of important considerations for interpreting value-added evidence. Then, a series of sections discuss specific uses within policy, practice and research, according to the extent to which the findings have implications for these. In each area, specific issues are discussed, drawing on the theoretical framework advanced in the chapter. This section also draws on the discussions of interpretation and uncertainty in Chapter 4, considering their implications for various areas of use.

Chapter 8 – Summary and Conclusions

This final chapter summarises the problems examined and the contribution of the thesis to the issues and questions identified. This includes the following:

- A summary of the main question addressed in the thesis and the approach taken to addressing it.
- Discussion of the limitations of the study and the questions which remain unanswered. This considers limitations stemming from the scope and focus of the study as well as limitations of the methods and data used.
- A list of the 'headline' empirical results from the 4 empirical studies and a summary answer to the core research question on the validity of school value-added measures.
- A number of recommendations and comments, first in general and then for the specific areas of use which have been reviewed.
- Concluding remarks about value-added and the issues examined in the thesis.

2. The Aims and Designs of Value-Added Measures

2.1 Chapter Introduction

This chapter describes how value-added measures came about, the motivations for their development and what they are in in both conceptual and technical terms. In doing this, the chapter provides foundational material for the following two review chapters.

The first section, *Value-Added and the School Effect (Section 2.2)*, describes the essential logic of value-added and the origins and problems associated with its particular operational interpretation and the term itself. This serves as a non-technical introduction to the concept and an initial exploration of the issues raised by this particular construal of school performance. The final section, *Value-Added Model Designs (Section 2.3)*, introduces a range of statistical approaches based on the value-added method which can produce value-added measures and other outputs which are referred to here as 'value-added evidence'.

2.2 Value-Added and the School Effect

2.2.1 The Value and Problems of Educational Measurement

Measurement is a cornerstone of physical science. The ability to measure temperature, length, weight, volume and many other physical phenomena are some of the most important achievements in human scientific progress (see for example Chang, 2004). Similarly, social scientists, politicians, managers and employees of social institutions have sought to measure important aspects of the social world. There is a widespread view that measurement has an important role in the operation and improvement of social institutions and programmes:

> "...in the past year I have been struck again and again by how important measurement is to improving the human condition. You can achieve amazing progress if you set a clear goal and find a measure that will drive progress toward that goal..."
>
> (Gates Foundation, 2013, para. 4)

Social measurement is not without its difficulties. Social phenomena of interest are often social or psychological constructs and may not be directly observable in the way many objects of study in physical science are. Some of the most significant concerns of social policy – poverty, academic achievement, wellbeing – are difficult to define and capture in a verifiable, objective measure. A common difficulty is that social data almost invariably have some degree of measurement error, missing cases, unobserved influences, unexplained variation and issues with defining and observing the object of measurement. Educational data are no exception to this (Gorard, 2010, Koretz, 2008) and the value and use of external tests of achievement has been a longstanding and contentious issue (Stobart, 2008).

Despite these difficulties, measures and indicators are key to the functioning of many organisations, allowing resources to be directed where needed and progress towards goals evaluated. Data are generally held to be of value to individual professionals to inform and improve their practice and are widely used in England (Kelly et al., 2010).

2.2.2 The Aim of Value-Added

A problem with the use of unadjusted examination scores – 'raw scores' - as a measure of school performance is that examination results are thought to be the joint outcome of numerous, interacting factors, of which the school is just one. The main 'inputs' into the 'production' of learning, the pupils themselves, are shaped by many factors which the school cannot, or can only partially, influence such as socio-economic background (e.g. Easen and Bolden, 2005), family effects (Rasbash et al., 2010), prior attainment (Ray, 2006), genetic influences (Haworth et al., 2011), pupil relative age (Crawford et al., 2007), parental engagement (Harris and Goodall, 2008), pupil motivations and self-beliefs (Stankov and Lee, 2014) and myriad other factors impacting on life experiences of pupils (see for examples Sammons, 2014).

When considering school effectiveness, these factors are commonly known as 'non-school factors' (e.g. Saunders, 1999) or as 'external' factors (e.g. Meyer, 1997). These non-school factors bias raw examination results for use as a measure of school performance. The question is whether the effect of the school can be isolated in some way from these non-school factors. If achieved, the result could be considered a measure of school performance, independent of the intake and the external circumstances of the school and therefore a fair basis for comparison of school performance. If the influences of non-school factors are not sufficiently eliminated from the measure, however, scores are likely to reflect differences in intake (Morris, 2015) rather than school performance.

Value-added measures are a particular approach to solving this problem. Their primary aim is to compare school performance independently of context and so identify more or less effective schools and practices despite differences in pupil intake (Teddlie and Reynolds, 2000). In other words, the aim of value-added is to 'level the playing field' when comparing school performances (Nor, 2014, p.77). This is a difficult task and requires several assumptions and simplifications which many would consider problematic. Nevertheless, the value-added method is a plausible attempt at capturing the independent effect of schools and there is a clear benefit should this be achieved. A measure which manages to isolate the effect of schools from other non-school factors has a special claim to the promotion of educational goals: The measure would be directly linked with measured outcomes (unlike lesson observations) and it would allow fair comparison of the performance of schools (unlike raw examination scores), allowing high and low performers in a given context to be identified and direct resources and actions accordingly. This twofold claim is central to the potential value of school effectiveness measures as policy tools. The value-added measure is also of vital importance for school effectiveness researchers in identifying the policies and characteristics of schools which are associated with higher or lower pupil value-added performance.

2.2.3 Origins of the Term 'Value-Added'

The term 'value-added' originates in economics (Goldstein and Spiegelhalter, 1996) where it is used for several purposes such as to refer to the value (sale price) of a product minus the cost of the raw materials and other inputs or as in the value-added tax (VAT), which is the percentage of a good's base value to be added in tax (Saunders, 1999). Given these economic uses of the term, an intuitive notion of value-added in an educational context would be that it captures the effect of a school on educational outcomes over and above the effect of non-school factors. Unlike the economic examples, however, there is no satisfactory equivalent to the raw material costs which can be used as a baseline and subtracted from the total. Pupils are not rendered homogenous by subtracting their initial level of performance: unlike economic raw materials, pupils have (and continue to have) agency. Pupil's progress is therefore attributable to the ongoing effects of pupil characteristics, behaviours (e.g. effort) and other external factors (such as parental involvement) as well as the impact of schools (OECD, 2008).

Levels of attainment are consistently found to be associated with subsequent rates of progress, with pupils with higher initial attainment tending to make more progress (Ready, 2013), a tendency recognised and built into the initial designs for the English National Curriculum (TGAT, 1988). This means that a simple subtraction of baseline attainment will not be sufficient: the association between initial attainment and subsequent progress would predictably disadvantage schools whose pupils had lower average attainment levels on entry. This problem is not so evident in many of the economic uses of the term value-added, where there is clear ordering to and separation of the effects of various factors. This means that subtracting pre- from post-measures can meaningfully capture the effect of later factors. Although, even in the case of economic value-added, changes in external market conditions or factor prices can complicate the analysis.

The independent effect of the school in an absolute, economic sense (henceforth, 'absolute value-added') requires 1) a pre- and post-measure of performance on the same scale and 2) a way to separate the effect of the school from all non-school factors. There are many practical difficulties associated with the first requirement of obtaining a measure with sufficient continuity across different school year groups (see Cahan and Elbaz, 2000), especially when the measures span large age ranges. In England, for example, the key measures of performance at Key Stage 2 (age 11) and Key Stage 4 (age 16), while both are suitable measures of academic achievement, they are not on the same continuous scale. Subtracting the former from the latter to measure progress would be meaningless.

Regarding the latter requirement, there are two options: first, a design-based approach which uses a suitable comparison group of pupils who do not attend (any) school against which the performance of the comparable pupils who did attend can be judged. This would allow the absolute value-added of a school to be calculated. Ideally, this would involve a random allocation of pupils to schools. Clearly this is not feasible (Luyten et al., 2005). The second option is the model-based approach using statistical controls, where the gain score between pre- and post-measures of performance is regressed on measures of the differences in pupil intakes, in principle isolating the independent effect of the school. For this analysis to estimate an absolute school effect, comparable pupils not attending school at all would be required so it is possible to estimate and control the effect of maturity and learning outside of school. Again, this latter requirement is clearly not feasible.

In sum, the absence of estimates from pupils who do not attend school and the lack of performance measures with meaningful continuity across time have meant that, in practice, it has generally proved unfeasible to calculate absolute value-added (but see Chapter 4, Section 4.4). We now move to consider how value-added *has* been understood and measured in education before commenting on how this departs from the general understanding of the term described above.

2.2.4 Adapting Value-Added to the Problem of School Effectiveness

While there are various understandings and formulations of value-added in education, the underpinning question which all have in common is, "How can pupil/student progress be measured in such a way as to throw light on the performance of institutions?" (Saunders, 1999, p.239). This, Saunders argues, is the "key to understanding the methodological principles of value added." This aim does not strictly require solutions to the two practical problems in creating absolute value-added above. As a result, value-added in education has come to refer to a measure of *relative performance* rather than either the absolute value-added of schools or a measure of average absolute progress (Kelly and Downey, 2011b).

The aim of the relative value-added measures used in education is to compare pupils' performances on a statistically 'like-for-like' basis (SCAA, 1994, p.6). This comparison almost always takes at least pupils' level of prior attainment into account as this variable captures large differences between pupils, typically accounting for around half of all variance in pupil performance (Teddlie and Reynolds, 2000, Thomas, 2001). Although the statistical models become more complex, the performance of pupils can be compared in this way to the mean performance for other pupils across a large range of statistical characteristics such as the level of prior attainment, gender, ethnicity, income, maternal educational level and so forth, assuming appropriate measures of these non-school factors can be obtained. The performance of statistically 'like-for-like' pupils (as captured within a statistical equation) can be used as a benchmark against which to compare pupils' actual performances. Calculating the difference between pupils' actual performances and this expected level of performance creates a measure of relative performance (i.e. performance relative to other statistically similar pupils). This relative performance is now known as 'value-added'.

It is important to emphasise that the school effect is not directly measured by value-added. The school effect is conceived as a latent variable revealed by statistically removing the predictable influence of all other explicable non-school factors. Strictly, value-added scores are merely capturing statistically unexplained differences in pupil performance. The causal attribution to the school rests on the assumption that all other appreciable non-school factors (i.e. omitted variable biases) have been ruled out (Marsh et al., 2011). Of course, there will be some level of measurement error and influence from unmeasured factors and so the estimated school effect can only ever be considered an approximation (Visscher, 2001).

It is worth commenting on the confusion caused by the shift in meaning between economic, absolute value-added (above) and this relative value-added measure used in education. Referring to the latter is somewhat a misnomer and a source of confusion for those not familiar with how the measure is produced (Goldstein, 1997, Coe and Fitz-Gibbon, 1998, Luyten et al., 2005). This confusion continues with the forthcoming English value-added measures (see Chapter 3) as these are named 'Progress' measures, which is similarly problematic. What is referred to as value-added in education is relative and has a mean of approximately zero. It is not a measure of either progress or absolute value-added. The more progress a pupil with a given prior attainment makes, the higher his or her *relative* progress, so, in this sense, the measure can serve as a measure of progress. Nevertheless, this loose terminology combined with low understanding of the value-added method (Kelly and Downey, 2010) is highly likely to lead to misunderstandings. The potential for misunderstanding is implicitly recognised in the practice of the English Department for Education (DfE) of adding an arbitrary amount to the school value-added scores (100 at primary level or 1000 at secondary level). This is done due to concerns that a negative value-added score would imply that pupils have made negative progress (Ray, 2006), rather than performed less well relative to statistically similar pupils, as is the case. Rather than this being entirely superficial, the loose use of the term value-added to refer to relative value-added is likely to exacerbate the difficulties users have in understanding the measure. The name obscures what the measure is and makes it difficult, even intuitively, to understand, for examples, how one pupil can make more progress than another yet receive a *lower* value-added ‘Progress’ score or how two pupils with identical mathematics attainment at both Key Stage 1 (age 7) and Key Stage 2 (age 11) could get a different value-added score. Nor does a casual conflation of value-added and progress make it clear that if all pupils improved their performance by the same amount, school value-added scores or the size of the ‘school effect’ would be unchanged. The measure is zero sum, a fact which limits its usefulness in monitoring improvement for the whole system over time.

A clearer name for the measure might be ‘relative performance’, ‘average relative performance’, ‘adjusted academic performance’ (Coe and Fitz-Gibbon, 1998, p.433) or ‘adjusted comparison’ (Goldstein, 1997, p.383), where the crucial aspect to convey is that performance is evaluated through comparison and captures performance relative to the average score of other (statistically) similar pupils, subject to the measures taken into account.

2.3 Value-Added Model Designs

2.3.1 Introduction

This section outlines different specifications of value-added models. Several broad groupings are discussed on a conceptual level to introduce their key differences as well as their common approach. Model specifications and technical notes are included in Appendix A (Sections A1 – A3).

The Value-Added Method

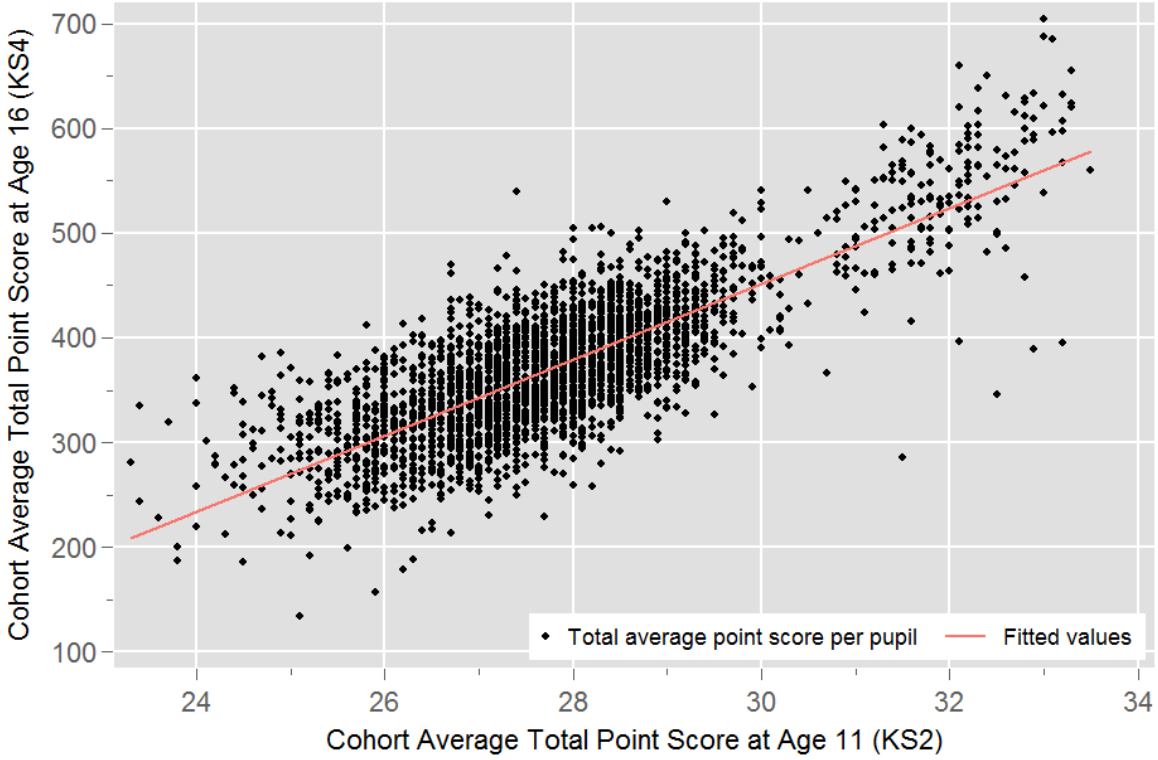
Rather than being a single measure based on an identical design, value-added is a general method which uses a range of different statistical models for different purposes. What these all have in common is that they aim to solve the problem of isolating school effects from those of non-school factors as described above. Conceptually, all models are based on a production function approach used in the economics of education (Ladd and Walsh, 2002, Brewer and McEwan, 2010). This approach conceives education as a mechanism in which a number of factors of production ‘produce’ educational outcomes which are typically measured by examination scores. The ‘raw’ performance scores are used as a dependent variable in an econometric model which is regressed on a number of factors which are regarded as beyond the control of the school (i.e. non-school factors). In this way, the remaining unexplained variation, the ‘value-added’ can be considered independently of the impact of the factors which have been taken into account. Exactly which factors need to be controlled depends on the purpose of the measure: If the measure is designed to measure school performance all non-school factors should be accounted for, if the measure is intended to inform parental choice, controlling for pupil composition is not appropriate since parents will be interested in the contribution of peers to the performance of other children attending the school (Raudenbush and Willms, 1995). Designs differ depending on the specific dataset and the purposes for which the value-added method is used. Moreover, any given value-added model can be used to produce numerous different outputs such as school value-added scores, estimates of the effect of school practices or estimates of the relative importance of the various factors which influence pupil performance.

2.3.2 School-Level Models

Due to the difficulties of obtaining pupil-level data, many early attempts to produce value-added measures used school-level data (Woodhouse and Goldstein, 1988). Despite their uncommon use, school-level models are described here to introduce the process of producing value-added estimates and as a contrast to the models covered in the subsequent sections.

School-level models use average final examination results as the outcome measure of performance and regress these scores on available measures of non-school factors such as prior attainment and socio-economic status, also recorded as school averages. As Woodhouse and Goldstein (1988, p.301) explain, “a ‘residual deviation score’ [is] then assigned to each school, being the difference between the school’s actual examination score and that predicted by the regression equation”. This process is most clearly shown visually. Figure 2.4.2a, below, plots the 2014 school-level Key Stage 2 (KS2, age 11) and Key Stage 4 (KS4, age 16) scores:

Figure 2.4.2a –2014 KS2 Scores Against KS4 Scores at School-Level (n=3015)



There is a clear relationship between prior attainment at KS2 and final attainment at KS4, highlighting the problem with using the KS4 examination results alone to judge school performance. Without taking this relationship into account, the schools taking the most able intakes will be identified as being the best performers. To address this, the value-added method fits a regression line to these data to estimate the 'expected' KS4 performance for any given KS2 score (shown as a fitted value line on Figure 2.4.2a). In this example, a linear relationship is fitted. The equations underlying this are included in Appendix A1.

This regression line can be used as a baseline for comparison of school performance: schools above the line have performed higher than might be expected given their prior attainment and schools below have performed relatively poorly. Schools are therefore judged by the difference between the regression line and their actual performance. To illustrate this, the performance of School A is shown on Figure 2.4.2b, below. In this case, the residual difference, r_A , is negative and School A would receive a negative value-added score.

Figure 2.4.2b –2014 KS2 Scores against KS4 Scores at School-Level (n=3015) with added marking showing the (negative) residual (r_A) for School A

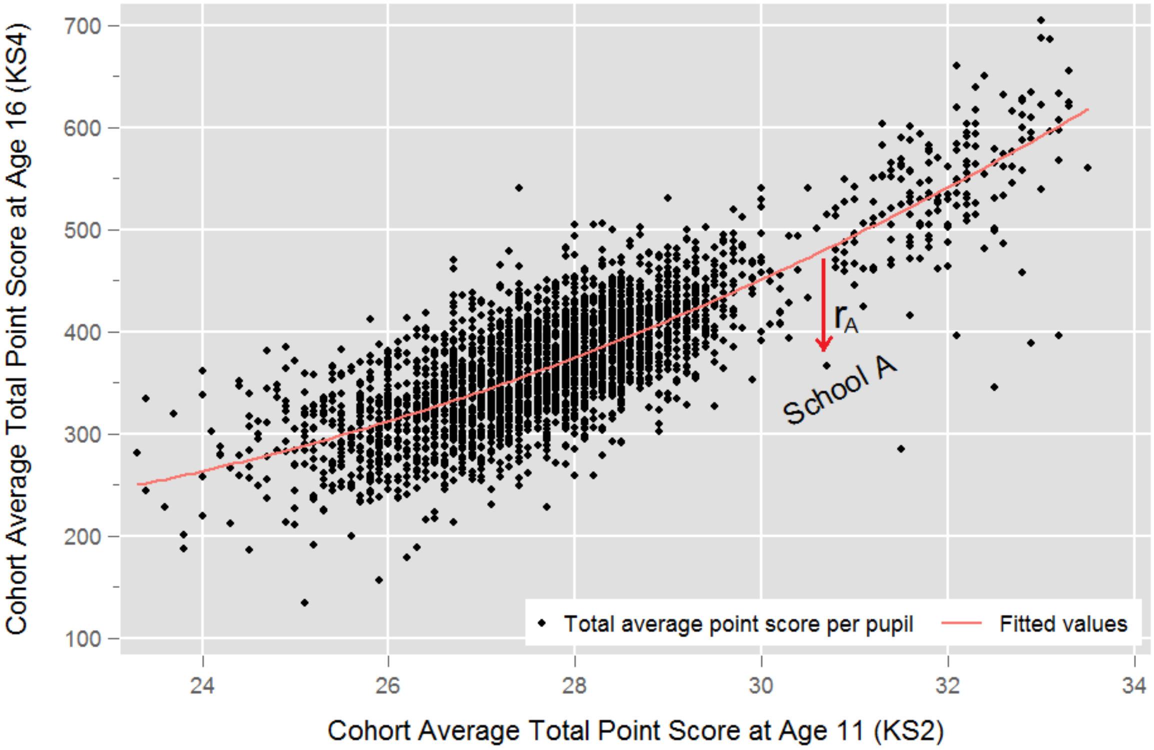


Figure 2.4.2b is also adapted from Figure 2.4.2a to show a non-linear (curved) regression line. Allowing the functional form of the relationship to vary in this way often improves model fit, better capturing the relationship between prior and final attainment. The official English measure, for example, includes both prior attainment and prior attainment squared to take non-linearity into account (DfE, 2013a) (see Appendix A1 for the underlying regression equation for Figure 2.4.2b).

These figures illustrate the basic approach used in the value-added method: A statistical relationship is fitted between the outcome (KS4 scores) and a non-school factor (KS2 scores) and then the residual difference between these is used as an estimate of school performance. These simple models can be extended to consider the influence of other non-school factors. Models including controls for pupil background are generally referred to as ‘contextualised value-added’ (CVA) measures to distinguish them from measures taking only prior attainment into account. The statistical models can also be extended to include interactions between the variables to capture more complex relationships. Strand (2014b), for example, uses a model with a large number of interaction terms to see if ethnicity and socio-economic status interact in their relationship with performance. In all of these extensions, although the equations get longer, the basic logic remains the same.

2.3.3 Pupil-Level Value-Added Models

There are several limitations of using school-level models. One of these is that they preclude examination of pupil-level differences or differences within school. As well as limiting the possibilities for research, use of school-level results potentially masks substantial differences in performance in relation to different pupil groups. With the creation of the National Pupil Database and widespread use of performance data systems within schools in England (Kelly et al., 2010), pupil-level data are now commonly available for analysis and can easily be aggregated to school-level when this is required. As a result, the debate about the seriousness of the limitations in school-level models is no longer current and pupil-level data are used. Further details regarding the limitations of school-level models are given in Appendix A.

There are several statistical approaches to estimating relationships between non-school factors and measured school outcomes. This section describes a) the common regression based ‘ordinary least squares (OLS)’ approach and b) the technique used for the forthcoming English Progress 8 measure (Burgess and Thomson, 2013a). Further approaches are examined in the technical document accompanying the report for the Progress 8 measure (Burgess and Thomson, 2013b).

Model Specification 1- OLS

One specification of a pupil level model uses ordinary least squares (OLS) multiple regression. This applies the school-level approach described above, to pupil-level data, where each data point represents a pupil rather than a school average figure. To get school-level value-added scores, the school mean of these pupil-level residuals can be calculated. The underlying equations for this are given in Appendix A2.

As noted in the previous section, models are typically extended to include a greater range of non-school factors in addition to prior attainment to further isolate the school effect from other confounding factors (see Appendix A2 for technical details). There are many non-school factors which have been consistently associated with school performance over several decades of educational effectiveness research (Teddlie and Reynolds, 2000). Prior attainment is the most important of these: As well as reflecting a direct effect of previous performance supporting future learning, prior attainment acts as a ‘black box’ which reflects a large number of unobserved factors which have brought about the differences in previous performances. There are several other factors which also tend to be associated with rates of progress over time, over and above those feeding into prior attainment (Teddlie and Reynolds, 2000). For example, the official 2007 contextualised value-added (CVA) measure accounted for associations between performance and deprivation, local area deprivation, in care status, special educational needs status, pupil mobility, gender, age within year, English language status, ethnic group and school average prior attainment (Evans, 2008).

Adding extra contextual variables explains more of the overall differences between pupils' performances and so reduces the size of the residual variation which is used as evidence of value-added. Kelly and Downey (2011b, p.64) estimated that a KS2-KS4 value-added (VA) model controlling for prior attainment accounts for 49% of the pupil performance variance while a CVA model accounts for about 57%. Accounting for a greater range of contextual variables makes greater demands of the available data (Kelly and Downey, 2010) and it is often difficult to control for non-school factors without also attenuating the value-added scores (Visscher, 2001) (see Chapter 4).

Model Specification 2 – Progress 8

The forthcoming English 'Progress 8 measure' (due 2016) uses a different estimation approach to that used in the regression-based models above. The approach described in the last section used an equation to fit a regression line to estimate the relationship between prior and final attainment. In contrast, the Progress 8 measure calculates the mean pupil KS4 point score for every possible pupil KS2 score for all pupils in the national cohort. The underlying measures of overall performance used for this are the KS2 average point score and the KS4 Attainment 8 measure (not discussed further here). Mapping of KS2 scores to KS4 scores in this way is analogous to fitting a regression line, above, but will produce an irregular (as opposed to linear or curvilinear) expectation line. In practice, the result closely resembles the estimate that would be produced using a non-linear regression line (Burgess and Thomson, 2013b). See Appendix A2 for a technical note on how to best map the KS2 scores to the KS4 scores.

The big advantage of this approach is that it is very simple and the correspondence between KS2 and expected KS4 scores can be provided in simple table or graph. Moreover, the resulting estimates and the proportion of the variance explained using the method is almost identical to approaches based on far more complex techniques (Burgess and Thomson, 2013b). One limitation is that the model cannot be extended to include a greater range of control variables (see above) without losing this simplicity, although subsequent adjustment is possible.

Outputs

Both the OLS and Progress 8 methods can be used to produce pupil-level value-added scores from the difference between the actual performance and the expected performance generated from each fitting method. If school-level value-added is required (e.g. for the English school performance tables), the school means of pupil-level results can be calculated. It is also possible to work with pupil-level value-added scores if this is of interest. For example, the mean value of groups of pupils within the school such as ethnic groups or classes can be found or the overall distribution be viewed graphically or summarised using key statistics.

Researchers are likely to include other educational effectiveness factors to examine their relationship (causal or otherwise) with performance. The simple Progress 8 approach does not allow for this kind of research – or at least without taking subsequent steps with the results. The OLS model, however, can be extended to examine factors which are associated with school performance. Factors can be included at various levels of analysis using the teacher-, school- or regional-level means, subject to their availability and concerns of the research in question. The inclusion of variables for purposes of study will generate estimates for variable coefficients within the model. These are an important output of the model in the research context, where the aim is often to understand the association between educational factors and performance.

2.3.4 Multi-level Value-Added Models

By the late 80s, educational effectiveness researchers were calling attention to methodological issues with the use of pupil-level models (Aitkin and Longford, 1986). These technical problems are detailed in Appendix A3. This section concentrates on explicating the difference between pupil-level and multi-level models on a less technical level.

As described above, it is possible to include school mean scores or other variables relating to larger organisational units (e.g. a local authority) within the OLS pupil-level model (e.g. Muñoz-Chereau and Thomas, 2015). In this sense, pupil-level models can be ‘multi-level’ and consider factors at school or other levels. This is not, however, the sense in which the term multi-level is conventionally used. The pupil-level models above are single-level models in the sense that they are ‘blind’ to the hierarchical nature of the data and cannot take group membership into account during the estimation process. This has two main consequences:

First, as was touched on when discussing school-level models, this limits the analytical possibilities of the model and prevents relationships within the data varying at the level of the school. If, for example, one wanted to examine whether a school’s effectiveness varied across the ability range, this would be difficult to ascertain using pupil-level data for large numbers of schools (see example in Appendix A3). Moreover, factor relationships may be different at different levels of analysis (see technical note on school-level models in Appendix A1). Analysis at only one level, therefore, has the potential to yield misleading results about effectiveness factors (Snijders and Bosker, 2011). Addressing these problems in pupil-level models is often possible to some degree through subsequent analysis of the residuals, use of school-level variables and the inclusion of interaction terms, but this is generally unfeasible for larger samples or more complex models.

The second limitation of using the pupil-level models relates to non-independence of observations violating the assumptions underpinning statistical tests within the model (Aitkin and Longford, 1986). In short, the problem is that a pupil-level model that is unable to account for group membership and treats each pupil as independent. This is to assume that two pupils from the same school are not expected to be any more similar than two pupils from different schools. When one allows for correlation between pupil-level errors within schools, larger standard errors are produced and so the stringency of statistical tests tend to be higher in a multi-level framework (Snijders and Bosker, 2011, Goldstein, 1997). The effect on the fixed effects (see below) estimates tends to be fairly small (Snijders and Bosker, 2011).

Critics have questioned whether multi-level models are any improvement on simpler methods in practice (Gorard, 2007). Moreover, Chapter 4, Section 4.3.3, examines the suitability and value of statistical techniques, raising several serious limitations. Nonetheless, whether multi-level models are as essential and valuable as is claimed or not, they have been overwhelmingly preferred in educational effectiveness research.

Outputs

Multi-level models offer two classes of output: First, coefficients on the fixed effects of the model show relationships between the dependent variable (pupil performance) and the independent variables included in the model such as pupil prior attainment or pupil background variables. As fixed effects, these hold (on average) across the sample. Educational effectiveness researchers are interested in examining the effect of additional effectiveness factors which have been measured and included in the model to examine the strength of relationships between these and performance. Isac et al. (2013, p.29), for example, use 'an educational effectiveness approach' to estimate the effect of a range of school factors (e.g. exposure to political and social issues information) and non-school factors (e.g. socio-economic status) on various student outcomes related to citizenship education (e.g. civic knowledge).

The second group of outputs which one might obtain from multi-level models are the 'random effects'. Note that the use of causal language in 'effect' might be misleading given that these are model residuals. Simple random effects include the residuals partitioned into the school and pupil-levels in the model, the former being school value-added. Examples of the use of random effects include the creation of school effects as value-added scores in the English performance tables and studies such as Noyes (2013) which examined school effects on mathematics performance and on post-16 participation. Random effects can also be estimated for school-level differences in the fixed effect coefficients. These can be understood as interaction effects between specific schools and the factor in question (see Appendix A3).

Other output which may be of interest are the standard errors of the estimates or other statistics associated with inferential statistical methods. Educational effectiveness researchers draw conclusions about effectiveness factors based in part on these statistical tests (see Chapter 4). Another noteworthy output of multilevel models which has been examined in a large number of studies (Luyten, 2003) is to partition the residual variance between various levels (e.g. pupil, class, school) in the multilevel model to see the proportion of variance situated at each level. Sometimes the term *school effect* is used in this sense, as the percentage of variance situated at school level (either including or excluding variance at lower levels such as teacher-level).

2.3.5 Growth Models

One final group of models which fall within the value-added method discussed here, albeit more briefly than for previous models, are ‘growth’ models with longitudinal outcomes. The models which have been discussed in the previous sections above have been longitudinal in the sense that they typically include a measure of prior attainment. Despite this, these have taken performance at a single time period as the outcome. In contrast, ‘growth models’ measure an outcome for two or more periods and attempt to measure a school effect as a trajectory over time (van der Werf et al., 2008, Guldemond and Bosker, 2009). A school’s effect is measured in terms of the slope of its pupils’ growth trajectory. This shifts the meaning of the school effect somewhat (see Chapter 7 for discussion). Growth models create a function which tracks changes in performance over time and can vary according to the form of the growth trajectory (i.e. linear, logarithmic, quadratic) (see for examples Guldemond and Bosker, 2009, p.260).

This is an important strand of educational effectiveness research which stresses the value of looking at how an outcome measure changes over time and the growth (rather than the status) of performance (Teddlie and Reynolds, 2000, Creemers et al., 2010). Growth models can become incredibly complex when looking at numerous variables, at numerous levels, over several time periods (see for example McCaffrey et al., 2004). Growth models can be calculated within a multi-level model framework (similar to that shown above but using time-dependent independent variables) or within the framework of structural equation modelling (Creemers et al., 2010). Methods used in this area of research are on first inspection very different to those described above. Despite these differences and complexities, the important similarity for present purposes is that the school effect is conceived as a latent variable (OECD, 2008). In other words, after controlling for a number of non-school factors, the remaining differences in the growth of performance are either attributed to the effectiveness of the schools or educational effectiveness factors when correlations are found.

2.3.6 Further Variations and Alternative Applications

This section draws attention to several extensions and variations on the models which have been described above. These all have some bearing on the validity of value-added measures but, to a large degree, can be considered separately from the core concerns.

One noteworthy extension is practice in multilevel value-added models of applying a ‘shrinkage’ factor to the value-added estimates. Residuals from all of the value-added models above tend to be more widely spread for smaller cohorts (Leckie and Goldstein, 2011, Gorard et al., 2012). This creates the problem that most of the worst and best performing schools tend to be for smaller schools. To combat this, residual, or ‘Bayesian’, shrinkage is performed as part of (or following) the multilevel modelling process (Snijders and Bosker, 2011). This reduces the value-added scores by some percentage inversely proportional to the size of the cohort on which the scores are based (Kelly and Downey, 2010, DfE, 2013a). Scores for very small cohorts are markedly shrunk towards the mean while the effect becomes negligible for much larger schools.

Other noteworthy model extensions include ‘multiple-membership’ models which take pupil mobility between schools into account and ‘cross-classified’ models where pupils are members of multiple groups, such as neighbourhoods of residence or the school attended at an earlier stage of schooling (see Goldstein, 1997, Goldstein et al., 2007). These and other variations extend the core models above in order to address inadequacies in the data and/or more fully capture the complexities in the statistical relationships being studied. By taking steps such as allowing parameters to vary by group, including known disturbance factors or adding weighting to the estimates, the analyst seeks to move closer to a true reflection of reality.

3. Value-Added in Research, Policy and Practice

3.1 Chapter Introduction

This section reviews the current use of value-added in educational effectiveness research, English policy and English practice. This draws on the technical details presented in Chapter 2 concerning the purposes, designs and outputs of value-added in order to discuss the application of the value-added method in each area of use. This chapter is organised into four main sections. The first two consider the use of value-added in educational effectiveness research (Sections 3.2 and 3.3), the last two consider the use of value-added in policy (Section 3.4) and practice (Section 3.5), respectively.

This chapter provides important foundations for the discussion and conclusion chapters which explore the implications of the new and pre-existing evidence presented in this thesis for the validity of value-added in relation to different areas of use.

3.2 Survey of Educational Effectiveness Research Use of Value-Added

3.2.1 Methodological Survey

Chapter 2 (Section 2.3) outlined a broad conception of the value-added method which had many different formulations and several possible outputs. These were linked by way of their common correlational, production function approach to identifying educational effectiveness factors. Elsewhere, this has been described as an econometric approach (Marsh et al., 2011). Identifying school effects and producing school value-added measures is one possible application of the value-added method. Section 3.2.1 reviews the scale and purposes of various uses of value-added in current educational effectiveness research (EER). This is done through a survey of the two highest ranked educational effectiveness journals (Scimago Lab, 2016), taken to be representative of current research methodology in EER: namely, *School Effectiveness and School Improvement* (SESI) and the *Journal of Research on Educational Effectiveness* (JREE). While neither EER nor the use of the value-added method in research is confined solely to these journals, it was preferred to review use in a core area (which may be expected to be following best practice) in detail than to review use across all educational research at a lower level of detail. The two journals reviewed are held to share similar goals and make use of fairly similar methods to EER published elsewhere in terms of variety but maybe not entirely in terms of proportion. Note that the term *EER* and other terms such as educational effectiveness and improvement (EEI) tend to be used in self-reference by researchers associated with the SESI journal and with the two handbooks reviewing the field (Teddlie and Reynolds, 2000, Chapman et al., 2015). However, for present purposes, *EER* is used in a general sense to refer to all sub-groups in what is an increasingly inter-connected and international field (see Chapman et al., 2015, pp.1-4, for an overview).

The following survey was of all papers published in SESI and JREE from January 2013 to when the survey was conducted in mid-April 2015. All papers were collected, read and the key methodological details of the publications were recorded through a simple sorting process (see below). In total, 9 issues of SESI and 10 issues of JREE had been published in this period. Only articles were considered and all content listed as commentaries, corrigendum, miscellany, introductions, editorials or notes were excluded from the survey. In total, 100 articles were surveyed: 44 from JREE, 56 from SESI. These articles were sorted into 7 categories according to the main methods used in the paper. The categories were designed to shed light on the use of value-added rather than give general readers an overview of methods used in the field. Two distinctions introduced in Chapter 2 were used to create the four categories for the correlational methods related to the value-added method: first, whether the study concerns longitudinal outcomes, as described in the growth models section (Chapter 2, Section 2.3.5) or cross-sectional outcomes, as elsewhere (Chapter 2, Section 2.3.2 to 2.3.4). Second, a distinction was made between studies which were primarily interested in the fixed effects from statistical models and those primarily concerned with the latent random effects or residuals (see Chapter 2, Section 2.3.4). Studies which fall outside of correlational methodologies were grouped into three groups: the first group was for experimental and quasi-experimental approaches. These are the key methodological alternative to value-added analysis, controlling for pupil differences by design rather than statistical analysis. The second group contained studies dealing with qualitative evidence or the implementation and refinement of research instruments (e.g.

through factor analysis or item response analyses). Third, there was a group for studies which reflected on the field or knowledge-base through theory, review or simulation studies. The second and third groups were far more general and are not considered in detail because these methods were not the main concern of the survey.

Several of the 100 articles showed more or less equal concern with more than one type of analysis and, as a result, are placed in two categories, bringing the total to 107. For simplicity, however, all figures are referred to as representing single papers. The results are given in Table 3.2.1a, below:

Table 3.2.1a – Methods used in Educational Effectiveness Research

Method	**SESI**	**JREE**	**Total**
Cross-sectional outcome, correlational, random effects focus	4	1	5
Cross-sectional outcome, correlational, fixed effects focus	14	0	14
Longitudinal outcome, correlational, random effects focus	7	1	8
Longitudinal outcome, correlational, fixed effects focus	16	5	21
Experimental or quasi-experimental	1	20	21
Qualitative analysis, implementation or methodological study	10	1	11
Theory, simulation or review	10	17	27
Total	62	45	107

Overall, about half of all papers presented evidence which could be broadly described as correlational. About a fifth of papers presented experimental evidence and just over a quarter reflected on the field through theory, simulation or review. Value-added models focusing on random effects in a single time period, as is the case in English accountability systems, were quite rare, comprising about 1 in 20 studies. Of these five studies (see below for further details), only two are concerned with school-level estimates and none seek to reach conclusions about specific schools, instead looking only at the overall magnitude of effects, their properties and factors correlated with better outcomes. In contrast with policy (see Section 3.4, below), educational effectiveness research shows little concern with value-added evidence for individual schools. It does, however, use the value-added method extensively to identify educational effectiveness factors from within the residual differences between schools after statistical controls (i.e. from their value-added). Another point of contrast between policy and research relates to whether cross-sectional or longitudinal outcomes are considered: this survey suggests that studies concerned with longitudinal outcome measures slightly outnumber those looking at cross-sectional outcomes in EER. The key English policy use of value-added is the production of the official value-added measures which concern outcomes for single years. These are two notable distinctions between policy and research use of value-added which have bearing on a central theme of this thesis that the validity of value-added evidence should be judged in relation to usage (Kane, 2013). Other distinctions which are valuable are detailed in an original theoretical framework presented in the discussion chapter in order to draw implications of the results for various areas of use.

Table 3.2.1a above also reveals stark differences between SESI and JREE. Of 62 articles reviewed from SESI, only one took an experimental approach (Antoniou and Kyriakides, 2011) compared to 41 whose primary mode of analysis could be described as correlational and so related to the value-added method. In contrast, 20 of 45 articles surveyed from JREE were experimental or quasi-experimental; only 7 adopted a correlation approach. This shows a clear difference in the methodological preferences of each journal and the communities which contribute to them. Something common to both journals is that there are a greater number of studies concerned with fixed effects rather than random effects. This suggests that the majority of current educational effectiveness research is focused on performance of particular educational interventions or approaches, or teacher-level effectiveness rather than studies of the properties of school or teacher effects.

To examine the focus of value-added studies in greater detail, a further analysis was conducted of the 45 papers recorded in the first four (value-added) categories (38 of these papers were in SESI, 7 in JREE.). Table 3.2.1b sorts the findings of these 45 papers into several groups according to their study design and main findings. As before, papers which presented evidence across more than one category were ‘double-counted’. For reference, the full table showing how each individual paper was sorted is given in the appendices (Appendix B1).

Table 3.2.1b – Overview of the Concerns of Empirical Educational Effectiveness Studies

General Concern	Specific Concern	No. of Instances Across 45 papers
Fixed Effects	School/Leader Practices or Characteristics	20
	Teacher Practices or Characteristics	12
	Pupil Characteristics	11
	System/other Practices or Characteristics	7
Random Effects	School Effects	7
	Teacher Effects	4
Other Concerns	Methodology	8
	Specified Interventions or Systemic Features	7

These results are in line with the broader analyses earlier in this section, emphasising two main points: First, there is a greater focus on the fixed effects of regression models than on random teacher or school effects. Second, there are differences in the unit of analysis for research use of value-added; 20 out of the 45 correlational studies, for example, sought to draw conclusions at the level of schools or leadership. The unit of analysis has implications for considering validity (see discussion chapter). It is also worth noting that there was an overlap between the random effects focus and the methodological studies; this is not apparent in the table but is described in greater detail in the next section. Many of the studies of school effects were methodological studies rather than attempts to draw conclusions about the properties of schools or school effects in their own right.

The results of the methodological study in this section underpin three main conclusions: first, the VA method is an important component of educational effectiveness research, although this is largely concentrated in the SESI journal. Second, there are large differences between the research use and policy use of VA. Third, there are many differences even within research applications of the value-added method, with a variety of foci at different levels of analysis. These are broad conclusions and it is important to note that many important distinctions and complexities have been lost through this summary. This includes details of the implementation of the methods and of how the evidence is linked to studies' conclusions. Moreover, it is likely that other authors may have categorised articles somewhat differently and surveys spanning different time periods would alter the overall distribution. These caveats are not thought to undermine the broad conclusions drawn above.

3.3 The Use of Value-Added in Educational Effectiveness Research

3.3.1 Introduction

This section is largely focused on examining the differences in how researchers have understood the limitations of value-added evidence and therefore the interpretations reached from the evidence presented. Several examples of studies are detailed below. These are chosen to exemplify principles and problems of value-added. Although they were not chosen to be strictly representative of all 49 studies, they do largely reflect the full sample of studies. The methods and type of findings in each paper is the main concern of this review, rather than the specific findings. The review is organised using the correlational groups related to the value-added method identified in Table 3.2.1b, above. This review allocates each paper to one of three sections: a) studies of fixed effects for educational effectiveness factors at all levels (n=49); b) studies of school-level random effects, the major focus of this thesis (n=7); and c) studies of teacher or class level random effects (n=4).

3.3.2 Use of the Value-added Method for the Identification of Educational Effectiveness Factors

Of the 100 studies identified in the educational effectiveness research survey detailed above, 20 studies used this approach to identify school-level effectiveness factors and 29 examined factors at other levels. This first group of studies examined the association between measured educational effectiveness factors and pupil performance after controlling for other (non-school) pupil characteristics. These studies do not seek to reach estimates of the performance of individual schools or the distribution of school differences, using value-added scores as an outcome in their own right. Their methodology does, however, rest on the assumption that, once non-school factors are controlled for, the variance in outcomes predominantly reflect (or at least contain) differences attributable to variable school effectiveness.

The first example discussed here is Chapman and Muijs (2013) who examined the practice of federating (i.e. partnering) schools. They obtained a sample of federated schools and used a statistical matching approach to find statistically-similar non-federated schools to use during subsequent comparisons using multi-level modelling techniques (see last chapter). The title, *'Does school-to-school collaboration promote school improvement? A* *study of the impact of school federations on student outcomes'*, is explicitly causal (Chapman and Muijs, 2013, p.351) and this claim is repeated in the text of the paper. They found that 'federation is positively related to performance in the years following federation' (p. 382) and concluded that 'federations can have a positive impact on student outcomes and federation impact is strongest where the aim of the federation is to raise educational standards by federating higher and lower attaining schools' (p385). Chapman and Muijs (2013) noted, however, that their conclusion should be treated with caution and observe the difficulties of reaching causal conclusions from correlation evidence. They point out that 'the possibility that differences found reflect non-measured variables cannot be fully discounted' (p. 358).

Next we look at de Bilde et al. (2013) who also examined the influence of school types, this time comparing the results of alternative and traditional mainstream schools in a longitudinal study. Growth curve analysis is used to model the rates of change between the 3rd year of kindergarten until the 3rd grade (pupil ages ranged from 53 to 87 months) on two measured outcomes: enjoyment and independent participation. Their results showed that there was no difference in enjoyment between alternative and mainstream schools and that equivalent pupils in alternative schools were actually rated lower in terms of independent participation by their teachers. When it came to drawing conclusions from this, de Bilde et al. (2013, p.229) explicitly ruled out the possibility of causal interpretations stating, "…although sometimes we referred to the term effect, the correlational nature of the data does not allow for causal interpretations."

Another example is Melhuish et al. (2013), who looked at the effects of preschools. Similar to the studies above, the effect of several preschool types (and non-preschool) are examined after controlling for a range of non-school factors including family socio-economic status, birth weight, developmental problems and home learning environment. They concluded that certain types of preschool provision, especially of a higher-quality, have a positive impact on pupil performance and call for expansion of high-quality preschool provision. The question of causal attribution was raised in relation to the observational nature of the study. In this case the authors made a judgement on the extent to which the results are likely to be confounded by unmeasured differences and the limitations of the measures used. This study had a large range of control variables which proved decisive in favour of their causal conclusion.

Similar studies are conducted to examine teacher-level effectiveness factors. Vanlaar et al. (2013), for example, measured various aspects of classroom practice and examined how these relate to reading comprehension. They used a multi-level model with a repeated measures design, 'controlling for student characteristics' and estimating differential as well as average effects of classroom practices. This is another example of a study in which the limitations and value of the value-added design are clearly noted:

> To truly determine whether there is a different effect of certain class-level variables on low- and high-risk students, an experimental design with a control group would result in more certainty. This study was nonetheless useful as it relied on the use of a longitudinal design and a large sample to indicate which variables may be worth studying through such an experimental design.
>
> (Vanlaar et al., 2013, pp.423-424)

These studies are all examples of what could be characterised as following an 'educational effectiveness approach' (Isac et al., 2013, p.29). There are differences in the exact statistical models used, how the statistical controls are implemented (e.g. Chapman and Muijs, 2013, who included a separate statistical matching process), whether outcomes are tracked in longitudinal data, as well as many other differences in detail and emphasis. Nevertheless, these all share a common approach and the authors have had to consider the limitations associated with this, choosing how to interpret and present the results. There are two key general threats to the strength of the conclusions reached in these papers which stem from limitations of the value-added method: First, the issue of whether or not the measured variables have done a sufficient job of ruling out confounding factors (i.e. are unbiased). It is possible that the estimates are biased by unmeasured non-school factors which have not been controlled for (omitted variable bias) and likely that schools or teachers adopting certain practices (such as choosing to federate) are somehow different to other schools, leading to a form of selection bias. The second major limitation is whether the relationships identified can be interpreted as causal. Drawing causal conclusions from correlational evidence is a well-known problem (Shadish et al., 2002). Some of the studies reviewed have made strong recommendations for practice based on value-added evidence, others have positioned it more as a step within a larger research process which is capable of identifying factors warranting further study.

3.3.4 School-Level Random Effects

Of the 100 papers published between 2013 and April 2015, 45 were correlational and in 7 of these, the main findings related to school-level random effects, (i.e. school value-added). Of particular relevance are a group of methodological studies into value-added measures. These studies are reviewed in detail in the next chapter along with other important methodological studies which were published outside of SESI or JREE. As a result, this section simply summarises their focus and conclusion, as follows: First, Dumay et al. (2013) examined the stability over time of value-added measures using different designs. They conclude that the low level of stability found 'poses a significant challenge to the conventionally accepted view that we can make a generalized evaluation of how effective a school is, based on cross-sectional data from a single cohort' (Dumay et al., 2013, pp.78-79). Another methodological study examining the stability of value-added estimates over time is Ferrão and Couto (2013) in which value-added estimates were produced for cohorts across three years and 9 grades in compulsory education in Portugal. Ferrão and Couto (2013, p.186) concluded that "the findings reveal a systematic pattern of educational units' performance is more than just randomness." At this – rather low – threshold for value, they conclude that Portugal should include a VA indicator into its system of evaluation. The two other methodological studies covered in this review are Televantou et al. (2015) and Lenkeit (2013). The first of these examined the impact of measurement error on value-added estimates, finding that traditional approaches to estimating school effects are positively biased, giving rise to 'phantom' compositional effects associated with school average achievement, for instance (see Chapter 4, Section 4.2.2). The second study compared value-added measures with estimates produced using growth models and cross-sectional estimates – so-called 'contextualised attainment models'. The latter were found to be 'adequate substitutes' (Lenkeit, 2013, p.39) for measures including prior attainment, although their predictive power is lower and correlations in some cases only moderate.

What can be seen from these studies is that bias, stability, the strength of causal inferences and the specification of value-added models are all current concerns in educational effectiveness research. Another key point is that, of the 7 EER research papers which have been reviewed which focus on school-level value-added scores, 4 are methodological studies examining the properties of value-added scores rather than attempting to draw conclusions using them. Only 3 (or 3%) of the studies specifically base conclusions on school-level value-added scores. These are as follows:

First, Noyes (2013) looked at school effects on mathematics performance using the National Pupil Database as well as the effect on subsequent outcomes such as participation at higher levels of mathematics. Noyes compared official school contextualised value-added (CVA) with CVA scores for mathematics only, created for the study. The latter are also used to examine how effective schools are in encouraging pupils to progress to more advanced study. Schools are found to have a 'very real effect' on mathematics progress to age 16 and on post-16 participation but that there was little correlation between these two (Noyes, 2013, p.101). The overall school CVA scores and mathematics CVA scores are found to show a 'considerable degree of variation' (p. 95), suggesting differential effectiveness between school departments.

The second study basing findings on school-level random effects is Isac et al. (2013) who compared outcomes across schools in different countries. They studied school and system effects on outcomes relating to citizenship education such as civic knowledge, civic attitudes and intended civic behaviour. They found only very small differences in attitudinal non-academic outcomes between schools but found some differences between systems and between schools in relation to civic knowledge. Like many studies reviewed in the previous section, they note the difficulties of using a correlational design and recognise that this prevents causal conclusions being drawn.

A final study considered here is Sammons et al. (2012). This paper investigated the impact of early disadvantage and the role of schools in ameliorating this. A multilevel structural equation model (which they note is a form of value-added statistical analysis) is used to track associations between self-regulation, academic attainment, early disadvantage and academic effectiveness over time, where a CVA measure is used to estimate the academic effectiveness of the schools. They found lasting effects of early disadvantage on regulatory skills throughout primary school. They also found associations between their measure of academic effectiveness and the academic attainment and self-regulation of children experiencing early disadvantage. Sammons et al. (2012, pp.15-16) drew a causal conclusion, stating that 'what this paper has shown is that more academically effective primary schools can make a significant difference to the academic attainment and self-regulation of children who experienced more disadvantages early on in life (before age 5 years)'

These three studies show that value-added is in active use within educational effectiveness research. They also continue to demonstrate differences in emphasis and caution in causal interpretations of the results: some authors explicitly rule out causal interpretations, while others present evidence as showing the 'effects' of schools and educational effectiveness factors, recognising but not being deterred by the threats to validity noted above.

3.3.5 Teacher-Level Random Effects

The final set of studies reviewed from the 45 correlational studies identified earlier are those which focus on teacher-level random effects (i.e. teacher value-added). As with the school-level papers, there were studies which had a methodological focus but, nonetheless, show that teacher-level estimates are an important concern. As in the last section, methodological studies are not reviewed and are left until the next chapter. These studies were as follows: Gustafsson (2013), identified threats to validity in observational data and 'discusses designs and analytic approaches which protect against them'. They found that all three designs tested found positive effects of the time spent on homework on student achievement. Second is Guarino et al. (2013) who examined causal inference in longitudinal data, looking at key threats to validity. They conclude that 'important features needed to establish causal inference' are 'likely neglected' (p. 164). Finally is Isenberg et al. (2015) who examined data issues involved in matching teachers to subjects in administrative data for teacher value-added estimates; about 1 in 6 teachers are found to be linked to a subject which he or she did not teach. As at school-level, these papers demonstrate that examining validity of school value-added estimates as a causal estimate of school effectiveness is a current and valuable concern.

Examples of educational effectiveness research drawing conclusions at teacher level include Konstantopoulos and Sun (2013) who looked at association between teacher effects and class size, finding few systematic effects but some evidence that teacher effects seem to be more variable in smaller classes; Johnson et al. (2014) who examined the sensitivity of teacher value-added models when including student and peer background characteristics, finding substantial differences between models; and Boonen et al. (2013b, p.126) who studied teacher effects and the impact of teacher characteristics on mathematics, reading and spelling achievement in the 1st grade, finding 'modest to strong' teacher effects depending on the subject area.

3.3.6 Review of Wider Research Use of Value-Added

Value-added is also used outside the EER field. This section concludes by briefly discussing research from several areas of study in which value-added evidence has been used or examined.

A common area of research is the use of value-added in monitoring and accountability regimes internationally (OECD, 2008, Chapman et al., 2011, Visscher and Coe, 2002). There has been much recent research on teacher-level value-added measures and alternative measures used in accountability systems in the United States (e.g. (Goldhaber et al., 2013, Walsh and Isenberg, 2015). Morganstein and Wasserstein (2014) observed and outlined the key issues in this discussion surrounding the use of value-added to make high-stakes decisions about teachers and schools. Researchers have attempted to make sense of the validity of the teacher-level scores by looking at correlations with the evaluations of school principals (Harris et al., 2014) and other tests of the assumptions required for causal interpretation and beneficial use (Condie et al., 2014, Goldhaber, 2015). There are similar efforts to scrutinise the policy and accountability use of value-added in many other countries (Timmermans et al., 2014, Gorard et al., 2012, Leckie, 2013, Manzi et al., 2014, Ferrão, 2012).

Value-added is also used to varying degrees across many educational research studies. There are recent, high-quality studies which have drawn on value-added evidence at different levels in different ways including Gutman and Vorhaus (2012), who used pupil value-added as an outcome measure to estimate the impact of behavioural and wellbeing measures on educational outcomes; Gershenson and Langbein (2015), who used school value-added data to estimate the effect of primary school size on academic achievement; Strand (2014a), who made use of school value-added data to identify more effective schools as well as pupil-level data to analyse educational performance differences associated with gender and with ethnic and socio-economic groups; and Coe et al. (2014) who made use of teacher value-added evidence alongside other sources of evidence (such as classroom observation data) to review what is known about effective teaching. Value-added evidence is one source of evidence focused on by Coe et al. (2014, p.4) due to its ‘moderate validity’ in signalling teacher effectiveness.

These four studies illustrate different ways in which value-added evidence has been produced and used. There are two implications for this thesis of the extensive and varying use of value-added in research. First, it underlines the value in considering what bearing the validity of the measure has on different uses. Second, this means that evidence presented in this thesis has implications for research beyond the field of educational effectiveness research, to the degree that the problems identified are present in other uses of the value-added method.

3.4 The Use of Value-Added in English Educational Policy and Governance

3.4.1 Introduction

This section reviews the origins of value-added as a national school performance indicator and its current use in a policy context. Section 3.4.2 presents a historical narrative review describing changes in policy thinking over several decades and how value-added has come to play a vital role in the English education system. The review makes selected links to academic thinking where supportive of the policy narrative. The historical narrative is followed by sections which review recent reforms (Section 3.4.3) and the current policy use of value-added (Section 3.4.4), including consideration of the new Progress measures.

3.4.2 Developing Value-Added as a National School Performance Indicator

The Changing Role of the State – 1970s

For most of the twentieth century, the prevailing view was that differences between schools of the same type were marginal, with schools doing little more than reflecting the individual ability and social circumstances of their pupils (Rutter and Maughan, 2002). In this context, although the state certainly sought to ensure a certain standard of provision and monitored school performance accordingly, the need for school performance measures which identified particularly effective or ineffective schools was not pressing. Without a national examination system which was standardised and criterion-referenced, both relative and absolute standards of school performance would have been difficult to discern in any case. The relatively small role that differences in school performance had in explaining differences in educational outcomes was well grounded in the educational and social research at this time. The view is well captured in the title of Bernstein's (1970) essay, 'Education cannot compensate for society' and large studies such as the Coleman report (Coleman, 1968) and Jencks (1972) seemed to confirm the same basic result: that the vast majority of differences in the educational outcomes of individuals can be attributed to factors other than the quality of the school.

Within a period of about 30 years, the idea that schools could be more or less effective and that this difference matters became the accepted view among researchers (Teddlie and Reynolds, 2000), policy-makers and the public. In the academic sphere, in response to these studies and the assumption which followed their findings that 'schools make no difference', school effectiveness researchers set out to discover whether a 'school effect' could be identified (Teddlie and Reynolds, 2000, p.3). Seminal school effectiveness studies such as Rutter et al's (1979) *Fifteen thousand hours: Secondary schools and their effects on children* and Mortimore et al's (1988) *School matters: The junior years* formed the reply to the view that schools made no difference.

A similar transformation in policy thinking was also taking place. A clear example of this is the 1976 Ruskin College speech delivered by the then Prime Minister, James Callaghan, which is widely acknowledged to mark a shift towards more active government involvement in education and an increasing concern with standards (Chitty, 2009). The origins of the current examination and accountability systems can clearly be traced back to the Ruskin speech (Ball, 2008). Nonetheless, at this point performance measures were primarily seen as a technology for understanding the system rather one for directly steering or shaping it. This view can also be seen in original statements from the Assessment and Performance Unit (APU), established in 1975. The APU was tasked with developing and promoting methods to monitor national standards and identify under-achievement, through use of sampling techniques. As its first director, Brian Kay HMI, explained:

> A national monitoring scheme would involve neither decisions nor judgments about individual pupils, teachers or schools - and the direct effects therefore might be expected to be small, especially when viewed in the light of the rarity with which any one school might be involved.
>
> (Kay, 1976, p.111)

The Changing Role of the State – 1979 to 1997

The gradual shift in attitudes in the 1970s, can be seen as preparing the ground for the policies enacted by the Conservative governments of the 80s and 90s. The wider context for these reforms was the political philosophy of the Conservative administrations in this period and how this informed their reform of education and other key public services. A key strand of thinking, often referred to as neoliberalism (e.g. in Gunter and Fitzgerald, 2015), held that markets and quasi-markets were solutions to improving public services. Conservative governments spanning from 1979 to 1997 privatised previously state-owned organisations and worked market mechanisms into the operation of remaining public sector organisations, creating 'quasi-markets' (Le Grand, 1991, Bridges and McLaughlin, 1994). These mechanisms went alongside and in some cases replaced older bureaucratic forms of organisation and notions of public service and professionalism (Ball, 2008). At the school level, ideas and methods from management and business were increasingly imported into education (Pring, 2012, Acquah, 2013). Words such as 'targets', 'performance indicators', 'audits', 'delivery', 'workforce', 'inputs', 'outputs' and 'efficiency' became part of the educational nomenclature (Pring, 2012, p.748) and there was increasing pressure to raise standards in education. The basic rationale behind these reforms was the intention to create a system in which 'successful decision-making is rewarded by an automatic, hands-off mechanism' which is 'largely independent of the direct influence of the LEA [local educational authority] hierarchy' (Le Grand, 1991, p.125).

For present purposes, what is important to note is that this market approach implicitly assumes that standards vary between schools, that agents within this system are able to identify differences in standards and are in a position to improve them. Given that school effectiveness researchers had developed methodological tools for estimating school performance, there were clear potential links which could be made between school effectiveness research going on during this period and these policy developments (Reynolds et al., 1996). At this time, however, the development of indicators of performance which could compare the performance of schools and local educational authorities was still ongoing (Woodhouse and Goldstein, 1988). This practical obstacle began to shift with the 1988 Education Reform Act (ERA) which introduced a National Curriculum and proposed a system of Key Stages and criterion-referenced testing at the end of each of these Key Stages (Gillard, 2011). ERA drew heavily on a report from the National Curriculum Task Group on Assessment and Testing (TGAT, 1988) to inform the guidelines relating to assessment (Whetton, 2009). The methodological problem of how to fairly measure school performance was raised by TGAT within the context of the policy use of the resulting national curriculum test scores:

> "Finally, there is a fear that results will be published in league tables of scores, leading to ill-informed and unfair comparisons between schools… Judgements about the quality of a school should not be confined to the extent to which the targets are actually reached. They should also take into account the educational value added – that is, the progress it might have been reasonable to expect a school in such circumstances to secure among its pupils, bearing in mind the pattern of attainment at intake and variations in children's rates of development."
>
> (TGAT, 1988, para. 18)

Despite this concern, league tables based on 'raw' examination results were published from 1992 by major UK newspapers using data made available by the Department for Education (West and Pennell, 2000). Along similar lines of thinking, a key actor to emerge within the reforms in the early 90s was the new schools' inspectorate, The Office for Standards in Education (Ofsted). Ofsted was largely created from the existing resources and personnel of Her Majesty's Inspectorate (HMI) which it replaced (Lee and Fitz, 1997, Smith, 2000). Before Ofsted, HMI's role had been the provision of advice to government and schools. In the metamorphosis into Ofsted, the inspectorate was turned 'inside out' (Smith, 2000), allowing parents, governors and the general public to be privy to the performance information originally reported only in private to government and schools. This shift is in line with the overall philosophy of introducing market-based mechanisms rather than being merely incidental: neo-classical economic theory stresses the importance of information in optimising the consumption decisions of consumers and so the publication of information from Ofsted and in the form of league tables can be seen as a response to potential problems of incomplete and asymmetric information within the economic mechanism which casts parents as consumers optimising school choices (Adnett and Davies, 2003).

In both Ofsted reports and league tables, information was now being provided outside of a professional context for purposes of accountability and parental choice. This requires forms of information which could encapsulate valid judgements about key aspects of schools' performance through standardised performance indicators, Ofsted ratings and criterion-based reports. This is in contrast to providing non-standardised data and feedback to users who are expected to have wider experience and professional understanding in which to contextualise the information. Central to all of this is a notion of differential school performance and the view that knowledge of standards at individual institutions was crucial to the maintenance and improvement of the system. In one sense, then as now, this public-policy position agrees with the prevailing view in educational effectiveness research: that standards differ, these differences matter and, therefore, identifying best practice and promoting it in some way is an important undertaking. Despite this alignment, the value-added estimates of school performance in school effectiveness research contrasted with the 'raw score' measures used by policymakers to measure performance. In other words, the system was being set-up around the idea of differential school effectiveness yet policy-makers were yet to have developed a credible national approach to measuring these differences.

The Policy Adoption of Value-Added

The need to produce league tables which reliably guide judgements about school performance brings the problem of measuring school effectiveness, the concern of this thesis, clearly into view. In the research context, during the 80s and 90s, value-added measures had been overwhelmingly accepted as essential for meaningful comparison of school performance by school effectiveness researchers (Teddlie and Reynolds, 2000). The policy context lagged behind this. Kenneth Baker, the Secretary of State for Education and

Science from 1986-1989, and the department were certainly aware of the research perspective following a number of seminars to explain value-added measures at the DES and a meeting between Baker and ‘half a dozen of the leading UK specialists in school effectiveness and value added methods’ (Smith, 2000, p.339). Nevertheless, presumably because of the technical and practical requirements raised in these discussions (Smith, 2000) and the view that adjusting raw figures made them ‘cooked’ (i.e. distorted and untrustworthy) (Saunders and Rudd, 1999, p.3), the league tables produced in the 1990s did not include a value-added measure and were criticised by some researchers as being misleading and flawed as a result (Mortimore et al., 1994, Thomas, 1998).

This problem, however, was increasingly being recognised by policy makers and the inspectorate (Sammons et al., 1993, Mortimore et al., 1994). Interest from Ofsted led to a substantial report into how value-added measures could be used by the inspectorate to put performances in context and make ‘like with like’ comparisons (Sammons et al., 1994). As Mortimore et al. (1994, p.322) point out, “the OFSTED specification recognised that a 'value added' approach, which employed baseline measures of pupils' prior attainment, would provide the most appropriate basis for evaluating school performance.” Value-added was also playing a key role in other key aspects of Ofsted’s development at this time. Ofsted was keen to build on the professional understandings which had been developed in the days of HMI by drawing on the school effectiveness knowledge base to identify the key determinants of school effectiveness. This led to a major report to Ofsted produced by leading school effectiveness researchers (Sammons, 1995). The reliance of educational effectiveness research on correlational methods using a value-added analysis means that Ofsted’s fusion of HMI expertise and the findings of school effectiveness research to produce inspection frameworks was resting to a substantial degree on the questionable validity of value-added evidence (Coe and Fitz-Gibbon, 1998, p.422). In these reports school effectiveness researchers stressed that findings should not be applied ‘mechanically and without reference to a school’s particular context’ (Sammons, 1995, p.5).

By the mid-90s the government was looking into the possibility of developing an official measure of value-added for use in accountability and monitoring systems. In the mid-1990s official reports, notably the School Curriculum and Assessment Authority (SCAA) report in 1994 (SCAA, 1994), considered policy options on value-added. The SCAA report was followed by a substantial study, The Value Added National Project, commissioned from the Centre for Evaluation and Monitoring at The University of Durham (Ray, 2006). The final report of this study was published in 1997 (Fitz-Gibbon, 1997). The report recommended that value-added measures for internal school use should be developed as soon as possible and identified features of the infrastructure and further considerations required for use as a public accountability tool. By the end of the 1990s, the stage was set for official value-added measures to be introduced for policy and accountability uses. The significance of school effectiveness research in these developments is made clear by Reynolds et al. (1996, p.134), who, reviewing the field in the mid-1990s, comment on the growing interest in school effectiveness research, how interactions with policy have fuelled this interest and the explicit intention of the Labour Government to explicitly base its reforms 'upon the insights of effectiveness knowledge'.

The Introduction of Official School Value-added Measures

The 'New Labour' government first took office in 1997, led by Tony Blair. Ostensibly, Blair offered a 'Third Way' between the market-based neoliberal approach of the Conservatives and the older statist approach (Ball, 2008). Instead of assuming that market forces such as parental choice would gradually improve standards alone, the onus was on the government to decisively act to bring higher standards about. The 1997 white paper, *Excellence in Schools*, was emphatic that there would be 'zero tolerance of underperformance… Schools which have been found to be failing will have to improve, make a fresh start, or close' (p12). League tables were to be reformed to be more useful, 'showing the rate of progress pupils have made as well as their absolute levels of achievement' (p6). The local education authorities were to regularly monitor schools 'on the basis of objective performance information' to ensure they were 'setting challenging targets and performing successfully' (p28).

With the level of priority given to education and large increases in the education budget, the political stakes were high. To make the best use of resources, intervention was to be inversely proportional to success, a principle introduced by the new 'Standards and Effectiveness Unit' headed by Michael Barber, the Chief Adviser to the Secretary of State for Education between 1997 and 2001 and, not coincidentally, a co-author of the review of the school effectiveness field and its increasing links to policy cited earlier (Reynolds et al., 1996). Despite this link, many school effectiveness researchers felt that the New Labour government, used school effectiveness research selectively as a ‘legitimation of its preconceived policies’, ‘frustrated with the caution, caveats and complexities of academics’ and ‘pursuing an agenda driven as much by political ideology and economics as by educational priorities’ (Chapman et al., 2015, pp.394-395). The appeal of value-added measures, and their potential to yield context-free, generalisable measures of school performance within this political agenda is clear:

> “The claim of measurable and generalisable outcomes helps simplify the task of the politician and the civil servant seeking to impose clarity on a ‘messy’ policy domain and… a polity that [has] grown impatient with careful research analyses of complex educational and social issues.”
>
> (Slee and Weiner, 2001, pp.86, cited in Chapman, 2015)

One gets a sense of the intellectual and political tone of the Standards and Effectiveness Unit’s approach and narrative by reading Barber’s publications while at the unit (Barber, 2001) and afterwards (Barber, 2004). In the former example, Barber positions educational reform as an ‘urgent’, ‘immense’ and ‘global’ challenge (Barber, 2001, p.217). After reaching this imperative, Barber claims the British government ‘know[s] how to create successful education systems’ and that their approach fits into ‘an emerging template for educational reform’ (Barber, 2001, p.217). Further specifics of an important component of the overall approach and its rationale are given in Barber’s *The Virtue of Accountability: System Redesign, Inspection, and Incentives in the Era of Informed Professionalism* (Barber, 2004):

> At its simplest, realising the benefits of accountability requires that people know what their goals are, that progress towards those goals is measured and that success is rewarded and failure addressed. That this process delivers better results should hardly come as a surprise. It is common sense.
>
> (Barber, 2004, p.8)

Barber continues to set out the thinking behind a performance management framework heavily based on accountability, comparative data; devolved responsibility; high standards; rewards, assistance and consequences. This performance management system is impacted by market forces and several forms of collaboration and capacity building features - some old, some new. As Barber asserts, “this process of performance management, drawing on business models, works well.” (Barber, 2004, p.15). This all fitted into what Barber calls an ‘informed prescription’ approach (p. 31). He cites the National Literacy and Numeracy Strategies as successful examples of a government actively enforcing a detailed, evidence-based accountability framework. Barber’s aspiration, however, was to move towards ‘informed professional judgement’ which, ‘require[s] persistent analysis of the data and the adoption of practice on the basis of evidence” (Barber, 2004, pp.30 & 32, respectively) (see Section 3.5 on current data use in practice).

Development of Policy Value-Added 2002-2010

Following major reports into the establishment of a national school value-added measure (SCAA, 1994, Fitz-Gibbon, 1997), the first value-added measure was introduced in 2002 based on a simple ‘median method’ (Ray, 2006). The measure was phased in at key stages within primary and secondary education between 2002 and 2004 as data became available (Ray, 2006). Initial models did not use ‘fine-grained’ prior attainment scores due to the low reliability of the early data; instead, there was a ‘compromise between marks and levels’ (Ray, 2006, p.40) in which three sub-levels were used for each national curriculum level. The quality of the attainment measures is a likely cause of the difficulties found by Gorard (2006c) relating to the correlations between intake prior attainment and value-added performance (see Chapter 4).

There have been several distinct methods used to produce value-added models since their introduction. CVA measures for use in policy were initially developed in the early 2000s (Reynolds et al., 2012, Evans, 2008), although simpler measures were preferred initially. At this time the DfE (then known as the Department for Education and Skills, DfES) was consulting academics and schools about how to improve the measure and the possibility of including contextual variables or taking different modelling approaches. In 2002, the Value Added Methodological Advisory Group was set up at the DfES for this purpose. EER researchers had had little input into the original development of the CVA measure but were involved in the advisory group where they, according to Reynolds et al. (2012, p.13), ‘consistently pointed out the limitations of existing CVA measures at meetings.’ Evans (2008, p.6) reports (on behalf of the DfE) that ‘although there was no consensus of opinion on the detail, there was support from most [academics] for the development of more complex models that used the new data.’ This is despite the major initial report finding that the estimates produced by more complex models were almost identical (Fitz-Gibbon, 1997). Piloting of CVA at different key stages took place between 2004 and 2006 (Evans, 2008). The KS2-4 CVA measure, for example, was introduced in 2006, after piloting in 2005 (Evans, 2008, Kelly and Downey, 2010). The KS4 CVA measure ran from 2005-2010 (including the first pilot year). It included a large number of contextual variables to account for non-school factors. Over 50 variables were included in the CVA model with variables or clusters of variables attempting to capture the effects of prior attainment, deprivation, special educational needs, mobility (recent school movers), gender, age within year, English as an additional language status, ethnic group, interactions between ethnicity and deprivation, school-level average prior attainment and its standard deviation (Evans, 2008, p.11). These were all captured within a complex multi-level framework (Kelly and Downey, 2010) (see Chapter 2).

3.4.3 Reforms 2010-2015

In 2010, the new educational secretary, Michael Gove, set out the coalition government's agenda for education in the Schools White Paper, *The Importance of Teaching* (DfE, 2010). A key strand of this policy agenda relates to school performance data and its use and publication. The government believed that “comparisons between different schools and local authority areas will drive higher performance and better value for money” (p. 13). These outcomes all fit within the overall theory of improvement described above which holds that the availability, accessibility, quality and quantity of information available on school performance promote system improvement. To encourage data-driven improvement, the DfE was to “set out more prominently in performance tables how well pupils progress.” (DfE, 2010, p.68). An indicator of the increased emphasis on data is the quantity being made available: in 2012 the DfE published “400 per cent more data about secondary schools [than in 2010]”, an increase from 46 columns of data to 230 (DfE, 2012).

Another important change in relation to the validity of the measure is the discontinuation of the CVA measure in favour of a VA measure (2011-2015). The VA measure (still current at the time of writing) uses a similar multilevel methodology but excludes all factors other than prior attainment. This move was ostensibly a political decision made by the government who felt that it was 'wrong in principle' to take characteristics other than prior attainment into account when comparing pupil performances, asserting that this entrenches low aspirations for children because of their background (DfE, 2010, p.68). Similar arguments were advanced by academics (Bradbury, 2011), albeit without providing evidence that inclusion of contextual factors did lower expectations and indeed that lower expectations have led to lower performance. The impact of this policy decision is analysed in the results chapter and discussed in Chapter 7. This is a current concern given that accounting for only prior attainment within the official VA measure is set to continue with the introduction of the new Progress 8 measure in 2016 (DfE, 2013c) (see below and see Chapter 2, Section 2.3.3).

In 2014, Nicky Morgan took over as Secretary of State for Education, continuing in a very similar vein to Michael Gove, albeit with what was viewed by many as a less combative style (Guardian, 2014). In January 2015 Morgan gave a speech titled 'Why Knowledge Matters' (WKM) which reflects on the past 5 years of reforms since 2010 as well as discussing current and future policies (Morgan, 2015). Morgan's account, starkly and simplistically depicts education as being in crisis before the (2010) reforms. Morgan's WKM speech continues Gove's general policy narrative instantiated in the *Importance of Teaching* (DfE, 2010) and makes use of many of the same rhetorical approaches (see Lumby and Muijs, 2014). Morgan's speech suggests that the government now not only holds a clear opinion on education but also is a key agent in managing and shaping state education. Although performance measures are not discussed at length in this particular speech, Morgan does briefly discuss the current reforms which were to 'transform' the way schools are held to account, referring explicitly to the new Progress 8 measure as a way of measuring the progress of all pupils. This is seen as an improvement on older headline measures which distorted behaviour around key performance thresholds (Davies et al., 2005, West, 2010). Although the decision to continue to exclude contextual factors from the new measure is not explicitly discussed, the issue is eluded to through a number of assertions that there 'is no automatic link between disadvantage and poor attainment' (section 7, para. 2). Morgan asserts that this is demonstrated by an example of a school with good performance despite a disadvantaged intake. These two ideas – that the new measures capture the progress of all pupils and encapsulate high expectations for all – are supported in Morgan's speech and elsewhere as key features of the new Progress measures, to which we now turn.

3.4.4 The New Generation of Progress Measures

Due in 2016, the new Progress measures (to be called 'Progress 8' for the KS2-4 measure) are in some ways a continuation of previous policy. As detailed in Chapter 2 (Section 2.3.3), Progress 8 is much simpler than previous measures. Pupil performances at KS4 are compared on the basis of a single variable: an average point score (APS) of English and mathematics performance at KS2. Progress 8 uses the mean APS performance of pupils with the same attainment at KS2 as the baseline for value-added comparisons at KS4. Using a single prior attainment variable and ignoring contextual factors eliminates the need to use complex statistical models to compare results across numerous variables. This has been found to produce almost identical estimates to those which would be produced by the current VA method (Burgess and Thomson, 2013b). Moreover, simplicity is likely to be advantageous in some respects: it is more likely than CVA to be stable over time (Allen and Burgess, 2011, Dumay et al., 2013) and has the potential to be more widely understood than previous measures which were considerably more complex (Kelly and Downey, 2010).

Like the current measure, Progress 8 ignores contextual factors. This was explicitly stipulated as a requirement for its design (Burgess and Thomson, 2013a). Consequently, the measure is knowingly disadvantageous to schools with a disproportionate number of pupils whose characteristics are associated with lower performance such as those classified as being in poverty. This places a limit on the extent to which the measure can be said to be 'unbiased' and 'fair' (Burgess and Thomson, 2013a, p.7). The implications of bias are especially concerning because the measure is also planned to be used as a basis for a 'floor standard' of performance, identifying low performing schools which require intervention. In their report on the new Progress 8 measure, Burgess and Thomson (2013a) show that, due to the exclusion of contextual factors, schools falling below the floor standard are likely to serve localities with high rates of poverty. The combination of this bias with a floor standard of performance will mean that, 'schools in disadvantaged areas may face continuous intervention' (Burgess and Thomson, 2013a, p.17). Progress 8 is also likely to be biased in relation to other contextual factors such as those included in previous CVA models (Evans, 2008). Pupil-level estimates of biases provided by Burgess and Thomson (2013a) are consistent with these but school-level estimates are not given. The magnitude of biases in the school results is not clear from pupil-level results alone as it will depend on how pupils with different levels of performance are distributed across schools. Also, as Burgess and Thomson's report pertains to the secondary performance measure, no primary-level data were presented. These issues are addressed in the first empirical section in this thesis where school-level estimates of bias in the current VA scores at secondary and primary level are presented.

One final noteworthy aspect of the planned Progress 8 measure for future years are the plans to set the expected level of performance in advance. This approach would make use of the associations between KS2 and KS4 performances from previous cohorts (an *ex ante* model) rather than waiting for the actual cohort data to be in before making performance comparisons, as is currently practised. This would have the advantage that schools would know the baseline against which their pupils' actual scores will be evaluated in advance; although whether this will be workable or will have any beneficial effect on school or pupil behaviour is debatable. A final decision on *ex ante* models will be made in 2017 after considering the decision of Ofqual, the examinations regulator, regarding standard setting in the reformed GCSEs (DfE, 2014a).

3.5 The Impact and Use of School Value-Added in School Practice

3.5.1 Introduction

This final section looks at how value-added measures are used in practice in English schools. Without further empirical evidence, it is not possible to give an entirely clear picture of how data and value-added data in particular are used in schools. Instead, the intention is to sketch out the common uses and how these differ from use by policy makers and researchers.

3.5.2 Use of and Attitudes towards Data in English Schools

Data of various kinds are in widespread use in English schools: around 85% of the school staff who are respondents in Kelly et al. (2010) reported using data regularly and 95%

reported using pupil performance data in a practical way to inform teaching and management. Research suggests that use of data is increasing and that this is an international phenomenon, with data feedback systems being set up in many countries (Verhaeghe et al., 2015). English schools have access to performance data through multiple channels including information published in public performance tables, services such as RAISEonline and the Fischer Family Trust (FFT) data, information services provided by local education authorities, internal data produced by teachers and school leaders and many other sources like those provided by the Centre for Evaluation and Monitoring (CEM) based at Durham University. More information on these data sources is given below, after discussing data use in general.

All of these data are provided to and sought by schools with a view to raising standards (Kelly and Downey, 2011b). Currently there is not extensive empirical evidence which demonstrates that data improves performance (Demie, 2013) but there are several studies which suggest that it might (Carlson et al., 2011, McNaughton et al., 2012). Kelly et al. (2010, pp.28-29) identified the main uses of data in four broad categories: informing whole school evaluation and public accountability; informing target setting; tracking and monitoring the progress of individuals and groups of pupils; and question-level analyses for individual subjects. These are very similar uses to those discussed in previous research in this area such as Kirkup et al. (2005) and subsequent formulations such as in Demie (2013).

The available research suggests a complex picture of mixed professional attitudes towards data in which the extent of use and the attitudes towards data vary considerably between and within schools. Kelly et al. (2010) found considerable differences in data use within schools between different members of staff. In general, data use was found to be proportional to the extent of managerial responsibility within the school, with senior managers reporting the most frequent use of data and classroom teachers reporting the least (Kelly et al., 2010, p.35). One exception to the relationship between seniority and data use was that many schools appointed a data manager and so it was senior leaders, rather than head teachers, who reported the greatest use of data (Downey and Kelly, 2013). Other factors linked with the extent of data use in other studies include the school staff’s level of data skills and the time and support available (Verhaeghe et al., 2010, Schildkamp et al., 2015).

A key theme identified in Kelly et al. (2010) in relation to attitudes towards data uses was the distinction between external and internal sources of and purposes for data. Many teachers in Kelly and Downey (2011a, p.158) saw internal data as more useful than the ‘official’ data and there were specific instances where confidence was low in the value and quality of the data. One example of this relating to value-added concerns the Key Stage 2 SATs results. These are used at secondary level as a baseline for value-added in secondary school. Yet many schools thought that the intensive preparation involved in the key stage tests made these unreliable as baseline measures and preferred internally generated data for tracking pupil progress as a result (Kelly et al., 2010). Kelly et al. (2010) also reported considerable tensions relating to purposes for which data are used. For some respondents there was ‘considerable negative feeling’ about the use of data to “tick boxes; to be used as a stick to beat teachers and schools; to set ever-increasing targets; to encourage competition between schools; and because the government does not trust teachers to be professional” (Kelly et al., 2010, p.8). Moreover, recent reforms have given schools more control over teacher pay and government guidance encourages teachers’ impact on pupil progress to be one factor when considering rates of pay (DfE, 2013b). Value-added is being used for high-stakes purposes with real consequences in schools. This raises many tensions and can create negative feeling towards use of data (Kelly et al., 2010). As Kelly et al. (2010) described, however, there are also many examples of positive views to data and acceptance of these tensions:

> “… I think certain teachers do feel under pressure but I am not sure that that is a bad thing. I think it is a positive thing because if using data is flagging up where a teacher is consistently getting negative value-added, then that teacher needs to be aware of it, rather than just ignoring it and pretending that things are OK when they’re not.”
>
> Advanced Skills Teacher (English)
>
> (Kelly et al., 2010, p.107)

There were also many respondents who saw data as a valuable part of professional practice who perceived greater ownership of data and more positive views about its value. Interestingly, data use was cast, on one hand, as an important part of best practice for educational professionals (also see Demie, 2013), on the other, as a threat to the professionalism of teachers, where excessive trust in data erodes rather than underpins the scope for professional judgement.

A conceptual framework for characterising the range of attitudes towards data is advanced by Saunders (2000) who describes a matrix of attitudes towards data based on two dimensions: ‘hot’ to ‘cold’ and ‘literal’ to ‘provisional’. These result in four broad attitudinal positions (Saunders, 2000, pp.252-253):

Table 3.5.2a – Attitudes to the Use of Data, adapted from Saunders (2000, pp.252-253)

1. **Unengaged** (cold, literal): this position is characterised ‘by an apparent resistance to taking the initiative in making use of data. Often because the data was perceived as being ‘out there’ and not intrinsically relevant to pedagogical needs.’ Data was only used when it was imposed as an external requirement.
2. **Technist** (hot, literal): the technist attitude was highly enthusiastic towards data and relied on data for monitoring and evaluating pupils’ performance. ‘Such data was seen as problematic with regard only to its accuracy; its meaning and interpretation were largely taken at face value.’
3. **Sceptical** (cold, provisional): this attitude did not reject data per se but was ‘marked by a resistance to the literal use of data’ and the individual raising reasoned and sometimes sophisticated objections relating to the limitations of data use.
4. **Heuristic** (hot, provisional): finally, the heuristic position is characterised by a positive approach to data which accepts and values data but views it as more useful ‘for raising questions rather than making judgements’. These individuals made use of data to inform their practice while recognising that data doesn’t ‘have to be perfect i.e. totally valid and reliable, in order to be useful.’

Effective data use is depicted by Kirkup et al. (2005) and elsewhere in terms of professional dialogue which combines data expertise with pedagogical knowledge to interpret the data and use these interpretations to inform decision-making. Demie (2013), for example, stresses that data should be seen as raising questions rather than providing answers and discusses how the Local Authority data service was designed to encourage this subsequent discussion. This depiction of best practice aligns with Saunders’ ‘heuristic’ approach to data (above) where staff are positive about the possibilities but are able to treat data as ‘provisional’ and indicative rather than definitive and self-explanatory. This is of great relevance to the concerns of this thesis as it suggests that the ability to interpret data is important to its effective use and so some understanding of the validity and reliability of data is needed for valid interpretation and beneficial use. When it comes to value-added measures, this raises the extra difficulty that there may be barriers to effectively using value-added data relating to understanding of statistics, the method and some of the issues with the data, especially for more complex measures (Kelly and Downey, 2010).

3.5.3 Value-Added Evidence as Part of School Data Use

Value-added data is one of many sources of information used by schools (Saunders, 2000, Kelly and Downey, 2010, Demie, 2013). Publically-available, recent evidence on what use is being made of value-added evidence (from various sources) and how it is being interpreted is currently limited. There is a larger amount of information available when it comes to examples of, or conceptions of, best practice. Demie (2013), for example, gives a detailed description of what is thought to be effective use of value-added and other data in a report titled, '*Using Data to Raise Achievement: Good Practice in Schools*'. How representative these case studies are of typical schools is unclear: they are explicitly put forward as examples of best practice. Nevertheless these give a clear description of a conception of best practice by a local authority (Lambeth) which has been 'recognised by Ofsted' for the quality the data services supplied by its research and statistics unit (Demie, 2013, p.18):

> “The schools and governors use contextual and value-added reports to monitor progress over time and to identify factors influencing performance, to identify key areas of action, to ensure improvements and to set targets and address issues of underperforming groups of pupils. Over time the schools’ own data, the Local Authority contextual and value-added reports and RAISEonline reports have been very useful in asking a number of the following questions in [the] context of factors influencing performance in the school:
>
> - How does the school compare to other borough schools in respect of performance at entry KS1, KS2, KS3 and GCSE, by gender, free school meals, mobility rate, and terms of birth and levels of fluency in English?
> - What is the relative performance of different ethnic groups and mobile pupils in the school compared to the Local Authority and national average?
> - What is the relative performance of different ethnic groups by free school meals and gender in the school compared to the Local Authority and national average?
> - How many pupils appear to be achieving less than expected levels at the end of KS2, KS3 and GCSE tests?
> - What are the school’s strengths and weaknesses?
> - What must be done to improve?
>
> These questions are debated and discussed at staff and governors meetings as a basis for self-evaluation and raising standards in all schools. As a result the senior management team, teaching staff and governors are well informed of the performance trends of the schools.”
>
> (Demie, 2013, p.48)

Presumably, many schools nationally will fall short of this vision of rich data usage. This description is, however, held up as an example to emulate and, given the common view of the studies reported here that data use is increasing, this example seems to be an indication of the direction of travel. It is noteworthy that this description and others in Demie (2013) mentions a number of data sources, many of which make use of value-added data other than the official measure.

It is also valuable to consider the indirect impacts of value-added evidence, in particular the use of value-added measures by Ofsted, the inspectorate. Published performance data including value-added data is taken into account by Ofsted when making judgements and in planning and preparation for directing and informing inspections (Ofsted, 2015). This information is set against other sources of evidence including for example pupils' books, in-year performance information used by the school and self-generated data from parental surveys (Ofsted, 2015). The official value-added scores are one of several sources of value-added and other information which informs the inspection process. Ofsted have also created groups of 'similar schools' based on a variant of value-added analysis using prior attainment (but not other contextual variables) to match schools for purposes of comparison (Ofsted, 2013). These data are made available publically for use by parents and governors and attempts have been made to make these as accessible as possible through 'data dashboards' which give a series of graphics summarising the data (Ofsted, 2013). As with the performance tables (see Section 3.4), the publication of such data is intended to create 'bottom-up' accountability from parents and pressure from governors to maintain/improve performance (DfE, 2010). Value-added data also indirectly influence parents and the public through inspection reports given Ofsted's use of value-added evidence to inform inspection judgements.

3.5.4 Use of Value-Added Data Services and Tools

Alongside publication of official value-added measures through Ofsted and the performance tables, there have been numerous other different ways of presenting progress data to schools (Kirkup et al., 2005, Kelly and Downey, 2011b). Given how numerous these are and the many differences between them in terms of how widespread their use is and the differing emphasis and presentation of each service, this data landscape is not reviewed here in any depth. Three sources of information are discussed: RAISEonline, the Fischer Family Trust service and an example of provision from a local educational authority (Lambeth). These are chosen because empirical evidence on their use is more readily available and a clearer picture can be given (e.g. Kelly et al., 2010, Demie, 2013). One notable omission from this discussion is the services offered by the Centre for Evaluation and Monitoring (CEM) based at The University of Durham. Further information about this and other data sources used by schools can be found in Tymms and Albone (2002) and Kirkup et al. (2005).

RAISEonline is a free online system provided by the Department for Education (DfE) in England since 2007 (Evans, 2008). It offers a more extensive range of data provided than those which are provided in the performance tables. This includes a break-down of value-added by subgroups, question-level data and data management and comparison facilities (Evans, 2008). All of this can be used by schools to monitor and evaluate their own performance and base decisions such as targets of school improvement initiatives on the available data (Evans, 2008, Demie, 2013). RAISEonline provides data which are common to schools, local authorities, inspectors, dioceses, academy trusts and governors (see RAISEonline.org). One advantage of this may be that common data will lead to a common understanding of school performance and so will foster a level of agreement about school performance between these groups; however, as Kelly and Downey (2011a, p.416) note (see above), using the same data for school improvements for both accountability and school improvement creates tensions which may lead to a less 'data-friendly' climate. It is also interesting to compare this practice of providing common access to several bodies for numerous purposes to the policy of CEM, another data service, to not permit general publication of the data it provides to schools (CEM, 2015). Guidance from CEM justifies this position by contrasting use of the data for publicity purposes and using it for feedback to assist school improvement (as is the primary intention):

> Publicity material is inevitably somewhat sweeping and simplistic. On the other hand, value-added data is often problematic and complex. Feedback should be seen as a starting point for investigation, subject to complex influences, requiring local knowledge and judgement to interpret, not a simple index of the quality of a school or department. There is a real danger that even honestly presented data may be misunderstood or wrongly interpreted.
>
> (CEM, 2015)

While this caution refers to the use of data for publicity, CEM's description of value-added also poses serious questions about the publication of value-added data for other uses (such as in performance tables) and underlines the conflicts which can arise when data are used by different users for multiple conflicting purposes.

Another common source of value-added data in schools is that provided by the Fischer Family Trust (FFT). Kelly and Downey (2011b) state that over 98% of secondary schools were making ‘regular use’ of the FFTlive service (now replaced by FFT Aspire) which provides analyses in a wider range of subject areas than RAISEonline. The reports provided by FFT are designed to help to support school self-evaluation and include value-added data from contextualised value-added data and measures which allow schools to evaluate performance against their targets and other performance standards (Demie, 2013). Subject-level data which is (or can be) disaggregated and manipulated by schools is generally thought to be more useful by practitioners; policy makers, however, often want average figures and headline performance measures (Kelly and Downey, 2011a).

The final source of value-added information looked at here is data which are provided by local authorities. Many local authorities provide data packages to support school improvement, one good example of which is provided by Demie (2013) in a description of Lambeth local authority’s data services. This provision includes value-added analyses which schools can use to ‘set targets and assess how well it has educated individuals and groups of pupils’ (Demie, 2013, p.8). The council provides school profiles to all secondary and primary schools in the area which contain a comprehensive set of data designed to be readily comprehensible to schools (sub-headed, ‘Making figures speak for themselves’). This concern with comprehensibility carries over into the provision of median-based VA comparisons within school profiles which, relative to contextualised value-added measures, are designed to be easier to understand than RAISEonline data (Demie, 2013).

3.5.5 Summary: Value-Added Data in English schools

The evidence reviewed has suggested extensive use of value-added data in English schools from a range of sources. Schools are drawing on these sources and sometimes combining them. One of the case study schools in Demie (2013), for example, was drawing on FFT, RAISEonline, school-generated data and CATs (cognitive abilities test) to inform their pupils’ performance targets. Another example school was using data to allocate resources in ‘data driven interventions’ targeted at groups and individuals identified by data (Demie, 2013, p.22). As well as these examples of rich data use there are likely to be many schools who are less enthusiastic and/or proficient in their data use. One difficulty is that a level of 'data literacy' is required to use data – especially more complex analyses – effectively (Kelly and Downey, 2011b, p.153).

Data presented in Kelly et al. (2010) suggests that many staff are confident with the use of data with around 90% of the respondents reporting being confident in using data. Taking ownership of the more complex data was, however, a source of frustration for staff in some cases. This problem is explored further in Kelly and Downey (2011a) who discussed the appropriate balance between technical and professional expertise required to successfully use data. Collecting and analysing data, they observe, requires relatively more technical expertise whereas the use and interpretation requires relatively more professional expertise. Many of the ways of sharing data discussed above can be seen as an attempt to bridge this gap. Similarly, Demie (2013) discussed how the local authority's provision sought to address some of the difficulties with providing usable data to schools and the difficulties with statistical/academic language creating a barrier to understanding. In doing this, they are seeking to bring both technical and professional understanding to bear in the practical context in which value-added evidence is used.

The task of bringing both technical and professional expertise to the data to reach justified and measured interpretations can be considered in light of this chapter more generally. This is at its most challenging when it comes to performance tables. English performance tables aim to reduce school performance to a small number of performance measures and, as they are for public consumption (where little technical or professional expertise can be assumed), these measures need to be as self-explanatory as possible, requiring little or no contextual or expert knowledge to interpret. High-stakes publication in performance tables, therefore, makes the highest demands of value-added data in terms of its validity and reliability. Compare this with the other uses which have been reviewed across the chapter; many of these have been able to bring together numerous sources of data as well as a level of technical and/or professional expertise when interpreting and using value-added evidence.

Whether this additional information and expertise is required for interpreting value-added evidence depends on one's position with regards to data use: If the heuristic orientation to using value-added data described in Saunders (2000) is taken, interpretation of the meaning and use of value-added data becomes a considerable part of the overall problem. On the other hand, the 'technists' described by Saunders (2000) take value-added evidence at face value and so the only question is what to do with the information revealed; the problem is solely a technical one where the analyst must use the best methodological tools to produce the most valid estimate of performance.

Value-added evidence has many serious threats to the validity and there are considerable difficulties with the interpretation value-added and its validity (Chapter 4 examines these threats in more detail). Looking further ahead, the evidence and debates reviewed in Chapter 4 and the evidence presented in the results chapter strongly suggest that the technists' position is untenable: value-added evidence is highly problematic and the means of conveying uncertainty are inadequate to the task of communicating the level of confidence to be placed in the results and their key threats to validity. As this thesis concludes, if it is possible to base justified and educationally valuable conclusions on value-added evidence, it will be in a context of data practice which draws on a range of data sources and combines professional knowledge and technical understanding. These ideas are developed over the coming chapters. More immediate conclusions relate to the use of value-added: the present chapter has made it is very clear that there are large differences in how the value-added method is used and value-added evidence is interpreted across educational effectiveness research, English policy and English practice. In all cases it has been found that value-added plays an important role and has real consequences.

4. The Validity of Value-Added Measures

4.1 Chapter Introduction

This chapter reviews the literature on the validity of school value-added measures, the validity of the value-added method more generally and the properties of the school effect revealed through value-added analysis. The focus is the use of value-added at school-level (see Chapter 1). Research at other levels is included in selected areas of the review but discussion of other levels is intended to be illustrative rather than comprehensive.

Section 4.2 centres on methodological study of school effects and value-added, reviewing research from within and beyond the educational effectiveness research community. The section is organised around a number of threats to validity and the methodological issues which are covered over several decades of research. The core concerns relate to the validity, reliability and consistency of the value-added measure and the properties of school effects.

As was discussed in the introduction, while further empirical evidence would advance understanding, it would not be sufficient to address important aspects of the core research question pertaining to the interpretation of the measures. Early literature reviews during the opening stages of this research programme revealed apparently intractable differences of interpretation of the available empirical evidence. Consequently, researchers expressed markedly different views about the validity of value-added evidence. The second main section in this chapter (Section 4.3), therefore, examines how these differences arise, evaluates recent debates in relation to inference and justification and thereby reaches an independent position on the issue of interpretation.

The final section 4.4 is somewhat separate to the main reviews in sections 4.2 and 4.3 and it does not address the core question of the validity of school value-added. Section 4.4 provides important information for one of the empirical studies in this thesis. The study in question compares value-added estimates of school effects with estimates produced using a quasi-experimental approach utilising a regression discontinuity (RD) design. To this end, Section 4.4 reviews research estimating school effects using the RD design and compares the RD design to the VA design.

4.2 Evidence Concerning the Validity of School Value-Added Measures

4.2.1 Introduction to the Validity of Value-Added Measures of School Effectiveness

This section reviews the evidence on the validity of value-added measures and so looks at various reasons to doubt whether VA does in fact produce accurate and fair, ‘like-for-like’ measures of school performance as intended. Four issues are reviewed: The first two issues considered are bias and measurement error. Both of these are direct threats to validity as the presence of bias or error reduces the validity of value-added scores as a measure of the school effect. The second two issues considered are stability and consistency. These issues provide more indirect evidence on the validity of value-added scores; indeed, what stability and consistency evidence reveals about the validity of value-added is highly contested and rests heavily on the interpretation of the evidence, as is examined in Section 4.3. Stability and consistency evidence is indirect in the sense that one cannot observe the effect of many non-school factors (such as measurement error) directly, but the observed levels of inconsistency and instability suggest a greater extent of the value-added variance is constituted by non-school factor variance. As biases and error cannot be easily separated from school effects, they often cannot be observed directly.

These issues have been discussed at length by educational effectiveness researchers and others over several decades of research (see for examples Sammons, 1996, Goldstein, 1997, Coe and Fitz-Gibbon, 1998, Teddlie and Reynolds, 2000, Visscher, 2001, Gorard, 2010, Marsh et al., 2011). The core area of research which is reviewed is a strand of methodological research that has been conducted over several decades of educational research which is primarily concerned with ‘foundational’ methodological issues such as the reliability, validity and generalisability of school effectiveness measures (Teddlie and Reynolds, 2000, p.55). This area of research forms the body of literature in which this study could be said to be based and from which a large portion of its conceptual framework is drawn. Teddlie and Reynolds (2000, p.49) identify seven scientific properties of school effects that have been examined within the overall area of EER concerned with 'methodological issues', as follows:

1. Existence and nature of school effects
2. Magnitude of school effects
3. Context effects (between schools)
4. Consistency of school effects across outcomes at one point in time
5. Stability of school effects across time
6. Differential effects within schools
7. Continuity of school effects

(Teddlie and Reynolds, 2000, p.56)

As explained in Chapter 2, because experimental designs are generally unfeasible (Goldstein, 1997), it is the value-added conception of school effects (relative value-added) and the value-added method which is the basis for most research into school effects. Although there have been recent methodological studies exploring alternative approaches to estimating school effects (Sammons and Luyten, 2009) (see Section 4.4).

4.2.2 The Problem of Bias

VA modelling aims to produce estimates of school performance by statistically removing the influence of non-school factors. As Marsh et al. (2011, p.283) noted, 'perhaps the most basic assumption' behind value-added evidence 'is that models appropriately control for pre-existing differences so that VA estimates reflect the effects of the teacher or school being evaluated and not the effects of prior schools, prior teachers or other pre-existing differences' (also see Coe and Fitz-Gibbon, 1998). As a result, there has been ongoing work in educational effectiveness research to "discern different kinds of 'noise' or extraneous information in the analyses of effectiveness, and to get rid of it as far as possible" (Saunders, 1999, p.249). A failure to control for non-school factors will result in omitted variable bias. Burgess and Thomson (2013b, p.8) describe this problem in terms of fairness, explaining that an unbiased measure in relation to pupil attainment is one in which "every child in a 'neutral' school would have the same chance of being identified as causing concern whatever their prior attainment, and every 'neutral' school would have the same chance of being highlighted as under-performing whatever their attainment intake profile" (Burgess and

Thomson, 2013a, p.13). This description can be extended to relate to any other non-school factor, where bias is defined as a systematic relationship with any non-school factor.

The problem of controlling for non-school factors (and so avoiding bias) is complex. It is discussed below in relation to three overlapping issues: first, technical problems of model specification; second, issues of data quality and; third, theoretical problems of model specification. There is not always a clear division between these issues. The intention is to distinguish various related facets of the overall problem. As explained in the first sub-section, below, the technical issues which have been the key focus of much of the literature examine ways of improving validity but say little about how valid the measures are. The most pertinent papers for this thesis, therefore, are those which explore the nature of the problem itself and fundamental issues which are not readily amenable to technical solutions (notably Scheerens, 1993a, Coe and Fitz-Gibbon, 1998, Luyten et al., 2005, Gorard, 2010, Marsh et al., 2011).

Technical Problems of Model Specification

Recall the issue of model specification introduced in Chapter 2, Section 2.3.2. As was described, first, one must have adequate measures of all important non-school factors and, second, the relationships between school factors and non-school factors with pupil performance must then be suitably specified within the value-added model. Correct specification will involve the inclusion of the required non-school factor variables as controls and that the relationship between these and performance is modelled using the appropriate functional form (Ladd and Walsh, 2002). An example of an inappropriate specification of functional form is fitting a linear trend to a curvilinear relationship. This would result in biases which differentially favour particular intervals along the range of the non-school factor scale. There are also subtler biases caused by different specifications and different estimation (i.e. line fitting) approaches (cf. Burgess and Thomson, 2013b, who show the varying ability of various estimation approaches to make the performance measure wholly independent of prior attainment across the full range of prior attainment scores).

Consider the nature of the overall problem of bias: Imagine a continuum which at one end has a raw performance score and at the other has a perfect measure of the school effect (i.e. isolated from all other extraneous non-school factors) (Meyer, 1997). The technical problem of model specification can be seen as one of how to utilise the available data in order to move as far as possible (or required) to the latter end of the continuum, from the raw attainment score towards this perfectly controlled school effect (also see discussion of this general idea in Coe and Fitz-Gibbon, 1998). A great deal of research has been devoted to identifying and explicating all the technical considerations involved in appropriately specifying a value-added model in order to eliminate the influence of non-school factors and correctly identify the school effect (e.g. Aitkin and Longford, 1986, Bosker and Scheerens, 1994, Raudenbush and Willms, 1995, Hill and Rowe, 1996, Goldstein, 1997, Snijders and Bosker, 2011, Timmermans et al., 2011). This literature addresses the difficulties of fitting multi-level models to hierarchical educational data, correctly specifying a – potentially complex and multi-faceted – school effect and issues associated with estimating uncertainty by way of statistical significance tests. These points are discussed further in Section 4.3. For now, we remain with the issue of specification of non-school factor variables in order to remove biases as far as is possible.

Much of the research into issues related to model specification can be understood as a way of advancing along this continuum from raw scores towards the perfectly isolated school effect, sometimes in minute steps and often by way of extreme technical complexity. The general form of these studies is to identify a problem (e.g. pupil mobility between schools), discuss and estimate its seriousness and propose how to account for the problem within the model (in this example using multiple-membership models) (see Goldstein et al., 2007, for further details on this particular example). Rather than repeat this literature, the remainder of this section explains that such issues, while important for specific value-added measures, shed little light on the core research question concerning validity: How to maximise the validity of value-added measures is a different question to how valid value-added measures are. As is discussed and evidenced further below, even if technical best practice is followed, there is still a serious question about validity. The problem of controlling for bias is not considered reducible to an entirely technical problem (Sammons, 1996, Visscher, 2001, Creemers et al., 2010, Goldstein, 1997). The relevant question for this thesis, therefore, is how large the distance is between actual value-added measures and the end point of the continuum rather than the amount of movement along the continuum that is achieved by a technical fix to a known issue. Indeed, it is of greater interest to know the validity of value-added when technical best practice *is* followed: if the non-technical threats to validity are not serious, the question of validity reduces to ensuring that all researchers and other analysts producing value-added evidence have the necessary technical skills and that best practice is followed.

This is not to disregard technical issues associated with the specification of value-added measures. Poor specification can have serious implications for the validity of the value-added measure in question. In disregarding contextual variables, for example, English policy-makers risk serious biases in the English VA measure (see Chapter 3). Given that a simple solution (adding more contextual variables to the model) is readily available, however, the issue might be more meaningfully understood as a political problem associated with the potential effects of value-added on practice. In which case, observed biases relating to known and available contextual non-school factors reveals little about the validity of the value-added method *per se*. This returns to the above point that validity is best considered when technical best practice is followed so as not to confuse the limitations of the method with shortcomings of a particular analysis. It is the seriousness of the problems which are not amenable to technical solutions that reveal the most about the validity of value-added and are discussed presently.

Data Availability and Quality

A common concern with the value-added approach is that it makes high demands of the available data, both to measure the outcome and the numerous confounding factors which need to be controlled to isolate school effects (Coe and Fitz-Gibbon, 1998, Goldstein, 1997, Gorard, 2010). If non-school factors are not entirely controlled for, there is a risk of omitted variable bias within the estimates, where the scores reflect factors other than the school effect. The practical problem considered in this sub-section is twofold: first, there is the difficulty of obtaining the required variables. Second, there is the question of whether these variables adequately capture the constructs they are intended to measure and are therefore of sufficient quality to entirely remove the effects of non-school factors. Let us consider the first problem, that of data availability:

There are many non-school factors which have been identified as being associated with performance, yet these are not always available, even in relatively high-quality datasets. There are no measures of socio-economic status, for example, in the National Pupil Database (NPD) and Pupil-Level Annual School Census (PLASC) which are used to create the English CVA measures (Ray, 2006). Even in the former CVA measure, which contained over 50 variables to control for non-school factors (Evans, 2008), when researchers sourced and matched a measure of maternal education and added it to the CVA model, estimates changed considerably (Dearden et al., 2011b). ‘Pupils with mothers in the top qualification category score[d] on average 0.3 standard deviations higher than pupils with mothers in the bottom category, corresponding to a difference of about 20 CVA points’ (Dearden et al., 2011b, p.269). Dearden et al. (2011b) make the following recommendation based on their results:

> “The policy response to the problem identified in this paper is reasonably simple: to collect better background information in the PLASC data. There is a large literature on the factors that impact on educational outcomes. Some of this information, such as family income, would be impossible to collect in administrative data. But other important determinants, such as parental education and family size (for example, how many older and younger siblings each child has), could be collected as part of the PLASC return.”
>
> (Dearden et al., 2011b, p.277)

In a research context, there are some examples of educational effectiveness studies (e.g. Sammons et al., 2007) which have been able to collect a rich set of background variables (such as parental education and salary) and use these to correct for non-school-factor differences. This makes it more credible to claim that non-school differences have been largely accounted for. Yet, studies collecting data on this scale are rare and even in the highest quality studies will be missing important variables. It is estimated, for example, that 23% of 11-16 year olds in England and Wales are receiving private tuition (the figure is 37% in London) (Sutton Trust, 2014). Data are not commonly available to control for the effect of tuition or other non-school inputs, yet these are likely to have a substantial influence on school value-added scores. In sum, some level of bias is inevitable as it is practically impossible to measure and control for ‘all relevant variables even if a strong theory is available to help researchers select all of these’ (Creemers et al., 2010, p.45).

The second problem considered here is the extent to which available variables adequately capture the constructs they are intended to measure. The variables used to control for non-school factors in value-added models are often crude, with poor theoretical grounding, as Coe and Fitz-Gibbon (1998) explain:

> “An example of such ungrounded modelling is found in the use of variables such as 'sex' or 'ethnic origin' which 'explain' (in the statistical sense) part of the variation in outcomes, but which do not explain differential performance in any true sense - unless it is argued that effects result from purely biological differences, or from unfair discrimination. In the absence of supporting evidence, alleging biological effects or discrimination would seem to be ethically questionable. If these allegations are not intended then these variables are being used as a proxy for some unmeasured characteristic with which they are associated, and which would genuinely explain why some individuals perform better than others. Presumably if this characteristic were identified and adequately measured it would account for significantly more of the outcome variance than the crude proxy.”
>
> (Coe and Fitz-Gibbon, 1998, p.425)

Moreover, measures of gender, disadvantage (e.g. free school meals), ethnicity and many other pupil characteristics are generally not proxies for a *singular* characteristic but for a complex array of factors which are unevenly distributed across society and individuals. A binary operationalisation of gender, for example, assumes there is a single phenomenon of ‘(fe)maleness’ which is uniformly shared by all members of the gender. A more sophisticated approach would lead us to consider numerous underlying factors which produce average gender differences. This would involve measuring a vast range of attributes and capacities which differ across both genders and/or understanding how these interact with similarly complex cultural and psychological factors with each individual’s milieu.

It is mistaken to view this as a largely soluble issue of data collection and variable specification, where the variables collected must be sufficiently numerous, well-conceived, fine-grained and adequately structured such that they align with the constructs which they seek to capture (Tymms, 1996, Fitz-Gibbon, 1996, Willms, 2003). The issue is both practical and conceptual: even where an extensive data collection effort is possible, it is often difficult to know which variables would be required. In relation to gender, for example, differences in average male and female examination scores are poorly understood (Spinath et al., 2014). A similarly complex conceptual and practical problem is encountered for every variable (and this is to only consider the non-school factors we know about). Consider the difficulties of controlling for complex constructs such as socio-economic status, disadvantage, culture or personality. Even in excellent conditions, it is seriously questionable whether these phenomena can be represented numerically with any great precision. Even the most complex statistical models have the task of reducing the unfathomable complexity of social reality to a few dozen variables, tidily separating the school factors from the non-school factors through the specification of the statistical model (see below). The use of the word 'model' is apt: like any model (from toy train sets to architectural models) the aim is to produce a simplified representation of reality, to trace, imitate and approximate rather than capture reality in all its detail. It is important not to confuse these representations of reality for the real thing and assume that because a variable is available to act as a control for some non-school factor, it will have entirely ruled out this factor as a source of bias. Value-added models undertake the heroic task of accounting for all appreciable non-school factors. At best, the outcome of this can only ever be an approximation (Coe and Fitz-Gibbon, 1998) and it is difficult to know how good an approximation it is given that it is, by definition, impossible to know the extent to which estimates are influenced by any remaining unobserved differences. The key question is, therefore, to what extent does the 'messy stuff left over by the process of analysis' contaminate (or even constitute) the estimate of the school effect (Gorard, 2010, p.746)?

Theoretical Problems of Model Specification

The previous subsection examined the practical impossibility of fully measuring 'all relevant variables even if a strong theory is available to help researchers select all of these' (Creemers et al., 2010, p.45). This section explores the latter half of this statement: What are the problems with creating a 'strong theory' to correctly specify a value-added model? Up to this point, the terms school factor and non-school factor have been used unproblematically. A key difficulty, however, is that value-added evidence does not provide a way of distinguishing between effects (school factors) and bias (non-school factors), yet the underlying cause of any observed difference in performance is of vital importance for the correct specification of the model.

The difficulties in this area are best explored using the official English value-added measures as an example. This section examines the single issue of creating English value-added measures which are independent of prior attainment. Adequately controlling for prior attainment is the most fundamental task of value-added models but, as is discussed, even this has proved problematic. Another important reason for taking the independence from prior attainment in the English VA measures as an example is that this is one key focus of one of the empirical studies presented in Chapter 6. The empirical results directly address some of the issues raised here.

Analysing the first year of KS2-4 VA data in 2004, Gorard (2006c, p.239) found a 'surprising correlation' between the school VA scores and the raw measures of attainment. Gorard measured the correlation between the KS2-KS4 VA and the KS4 results as 0.96. Independence between value-added and the outcome measure is not strictly required as, other things being equal, one would expect schools which 'add more value' to get both higher VA and higher raw scores. Nevertheless, as Gorard, observed, the high correlations meant that the VA measure offered negligible additional information about school performance. This suggests that the model had not adequately controlled for non-school factors. In terms of the continuum considered above, the measure failed to advance more than a negligible way away from the raw scores. As the only factor controlled was prior attainment, this suggests either a problem relating to the predictive power or quality of the prior attainment measure or that further contextual variables were required to adequately distinguish 'raw' attainment from value-added.

Subsequent analysis of the 2005 CVA pilot data by Kelly and Downey (2010, p.184) investigated "how the addition of pupil- and school-level contextual demographic variables in the CVA model affected the relationship between the raw unadjusted threshold performance measure (the proportion of students attaining 5+ A*–C GCSE grades) and the value-added measures". Kelly and Downey (2010) found substantially lower correlations between raw scores and value-added in the new CVA measure compared to Gorard's analysis of the 2004 measure. At face value, then, the inclusion of contextual variables has moved along the validity continuum, further isolating the value-added from differences in raw scores. This suggests that the policy decision to include contextual variables (in 2005) improved the validity of value-added and, conversely, the decision to disregard contextual variables (in 2010) is likely to have reduced it. The actual effects of these policy moves is examined in Chapter 6. For now, the apparent solution (i.e. the addition of contextual variables) is considered in relation to the problem of correcting for bias.

Analyses conducted in this thesis (see Chapter 6, Section 6.1) found that the correlations in Gorard (2006c) stemmed from what is known as a compositional (or 'peer') effect on attainment. A compositional effect is when pupils' attainment is found to be associated with the attainment of their peers, over and above the association with their own attainment. In the CVA model, this had been corrected by the inclusion of two variables: mean cohort prior attainment and its standard deviation (Evans, 2008, Kelly and Downey, 2010, Wilson et al., 2008). By design, introducing contextual variables corrects for the bias. If a measure was required for purposes of parental choice rather than school evaluation, however, it would not be appropriate to do so (Raudenbush and Willms, 1995). The choice of whether to include the variable is ostensibly straightforward. The difficulty underlying this choice, however, is that there is considerable doubt about whether the compositional effect is genuine. Studies have shown that 'phantom' peer effects can arise from pupil-level measurement error (Harker and Tymms, 2004 , Televantou et al., 2015, Pokropek, 2014). Televantou et al. (2015), for instance, found that 'traditional approaches' to multilevel models produce positively biased compositional effects due to measurement error and He and Tymms (2014) found that OLS estimation of value-added leads to systematic bias relating to pupil average ability due to the asymmetric treatment of final and prior attainment measure errors in the estimation process. For now we remain with the theoretical problem; the empirical problem is returned to in the results and discussion chapters (see Chapter 7, Section 7.2.2, on measurement error).

The theoretical problem here is that by including school-level variables, the CVA measure will have 'mopped up' variance that was not explained at pupil-level, *irrespective* of whether this variance was due to measurement error, poor prior-attainment controls, poor control of other non-school factors, or a peer effect (Harker and Tymms, 2004, p.181). If one takes the meaning of the school-level variables in the value-added model at face value, all of these sources of variation would lead to an apparent compositional effect. Yet, on further inspection, 'the compositional effect is very difficult to pin down' (Harker and Tymms, 2004, p.195) and may well be partly or wholly spurious (Gorard, 2006a). Have the contextual variables added in 2005 solved the problems associated with the 2004 VA measure (Gorard, 2006c)? Or have they masked it with an arbitrary technical fix? What sources of variation have been captured by the new contextual variables? As Harker and Tymms (2004, p.195) point out, '...the really worrying thing is that the researcher can never be sure about what has been found.' This creates some serious problems in relation to separating school factors from non-school factors, effect from bias. Given this problem, it should be of no surprise that the existence of school compositional effects is still 'controversial' (Reynolds et al., 2014, p.209), despite major studies over many years and considerable statistical 'firepower' being brought to bear on the issue. Despite Liu et al. (2015) claiming that compositional effects have been 'consistently verified', there are many studies which fail to find effects (Lavy et al., 2012, Boonen et al., 2013a, Marks, 2015). Marks (2015, p.18), for example, "analyse[d] data with a large number of cases with reliable measures" yet found that "school-SES [socio-economic status] effects are trivial and do not warrant a policy response."

Let us bring this lack of consensus on the specific question of compositional effects to the finding that there is an apparent selective (or 'grammar') school effect in the English school system (Goldstein and Leckie, 2008), where the majority of selective schools form a cluster of schools with a positive value-added. The best available evidence suggests that there is a small grammar school effect, although it is difficult to rule out the problems associated with inadequate data and omitted variable biases (Coe et al., 2008). Even with the best data and high-quality analysis, it is difficult to know whether an effect is real: associations between non-school factors and performance say little about the underlying cause, variables only 'explain' the differences in the statistical sense of the word (Coe and Fitz-Gibbon, 1998). Consider the following explanations for this association:

(1) Selective schools are more effective with like pupils, maybe due to attracting and retaining better teachers.
(2) Selective schools are more effective because of student composition, where being surrounded by other high-ability pupils has a beneficial effect relative to a comparable pupil in a more mixed-ability environment (i.e. a peer effect).
(3) The difference is not a real effect and stems from omitted variable bias: i.e. there is some unobserved difference between high-ability pupils attending a selective school and high ability pupils attending other schools.

(4) School-level 'compositional' effects arise from measurement error at pupil-level, creating so-called 'phantom effects' which are 'mopping up variance at the second level' (Harker and Tymms, 2004, p.181, Televantou et al., 2015).

(5) The result arises not because the selective schools are generally more effective but rather are differentially effective; specifically, that they are especially effective with more able pupils as suggested in Foley and Goldstein (2013).

The crucial point is that some of these explanations are forms of bias, which need to be controlled for; others are effects, revealing that selective schools are indeed more effective. The selective school effect serves as a good example of the general problem that the difference between a bias and an effect is rarely clear-cut. Take one more example: the well-established finding that poverty is associated with lower school performance. On one hand, this could stem from external influences and not taking this into account would disadvantage schools in poorer areas. If, however, the difference reflected poorer standards of education in poorer areas, it would be inappropriate to control for this. Like the example above, we do not know what the underlying cause is and cannot even specify the level: it could be an individual effect (e.g. the effect of a deprived home life), a peer effect (e.g. social problems associated with concentrated poverty), a teacher effect (e.g. difficulties of attracting good teachers in challenging areas) or a school-effect (e.g. the lower quality of leadership in challenging areas). It is only by assumption that a difference which is highly consistent across a large number of schools reflects a non-school influence and a school-level cause.

The problem runs deeper if we consider the possibility that there is more than one reason behind any observed difference. What if the selective school effect was partly due to a better quality of teachers and partly due to unobserved attributes of selective school pupils? If this were the case, either one needs to obtain variables which can distinguish these or the analyst must decide which is the larger of two threats to validity: attenuation of the school effect (when controlling for the difference) or biasing the measure with a non-school factor (when the variable is not included in the model). This problem is pointed out in Coe and Fitz-Gibbon (1998) and Visscher (2001) who make the implications of this clear:

> "If one corrects for student intake differences between schools when constructing value added school performance measures, then unintentionally one also corrects for quality differences in educational practices. However, the intention was to estimate the effects of the latter. This riddle cannot be solved satisfactor[ily], and for that reason one has to bear in mind that the true, never to be known, school effect (i.e., that portion of the-between school differences in average achievement levels of students caused by the schools they attend) lies somewhere in between the gross outcomes and the value-added measures (cf. Grisay, 1997, Raudenbush and Willms, 1995)."
>
> (Visscher, 2001, p.207)

In summary, this section has highlighted a number of issues with controlling for bias in value-added models. Rather than this being a wholly technical matter, it has been argued that this is problematic, approximate and that the use of theory is indispensable in the construction of value-added models (Creemers et al., 2010). There are profound theoretical and practical difficulties with obtaining the necessary data on non-school factors and specifying a value-added model to estimate the school effect independently of these.

4.2.3 Measurement Error

This section overlaps with the discussion of data quality in Section 4.2.2. It is useful, however, to separate a number of conceptual differences and specific problems which have been discussed in the literature. The section explores other threats to validity and raises several questions which are addressed in the empirical results. Like non-school factor bias, measurement error is a direct threat to the validity of value-added scores. Measurement error could be described *as* a non-school factor. Nevertheless, it is valuable to use different terminology and discuss measurement error separately as its nature is likely to differ; vague uses of words like bias and error is not conducive to clear debates around the various complex threats to validity (Amrein-Beardsley, 2014, and see Section 4.3) The conventions followed here are as follows: all biases and errors from any source of any type are, following Amrein-Beardsley (2014, p.38), referred to as 'construct irrelevant variance' (CIV). CIV is simply anything other than the school effect. The term bias is used in preference to error to suggest that the CIV is potentially tractable and non-random. With a given dataset, bias can be observable (i.e. a factor which has been measured but not been included in the VA model or adequately specified) or unobservable (i.e. an unmeasured or unknown non-school factor); but in either case, the term is used to describe something somewhat systematic and potentially amenable to analysis. Conversely, the term error is used rather than bias to connote intractability. The primary sources of error will be those associated with measurement error and unpredictable 'chance' events that influence pupil performance which cannot reasonably be measured. The extent to which errors tend to be randomly distributed is contested (Muijs et al., 2011, and see below) and so the term error (or measurement error) is used to encompass both random and non-random error (although the distinction is often explicitly made). As CIV associated with the measurement and specification of non-school factors has already been discussed at length, this section focuses on CIV related to the measures of attainment.

The validity and reliability of the outcome measure itself is a fundamental issue for the validity of value-added (Meyer, 1997) and educational testing more generally (Koretz, 2008). Put simply, if the test does not adequately capture what students have learnt at a school, a value-added measure will not fully reflect the school effect. Without valid measures of the outcomes in question, value-added measures are a non-starter. Moreover, with any given testing system, there is a considerable degree of unreliability which stems from marking and other factors not relevant to the attainment that is measured (Newton, 2013). It is important to recognise that test validity and reliability are of fundamental importance to the validity of value-added measures: any imprecision or unreliability in the outcome measure will lead to CIV in value-added scores. The specific issue of the validity of raw performance measures *per se* is beyond the purview of this research. The issues that can be addressed are a) the extent to which measurement error in outcomes scores translates into school-level error in value-added and b) the relationship between the stability of raw performance scores and school value-added scores. Both of these issues are addressed to some degree in the empirical sections of this thesis. The remainder of this section reviews several issues raised in the literature relating to measurement error which are specific to, or have particular relevance for, value-added measures.

First is the issue of aligning performance measures over time, as required for value-added analysis. The English value-added models, for instance, estimate a composite score of 8 subjects at KS4 (which are different for different pupils) using measures of just 3 at KS2 (reading, writing and mathematics) (Kelly and Downey, 2011b). This is a serious and multi-faceted problem. The general problem is that academic performance must be wholly analogous at two different points in time, both conceptually (ontologically) in terms of what academic performance *is* but also in terms of what is captured by the measure used at each point. Any disparities will give rise to spurious differences in relative performance. Consider the finding that cognitive ability tests (CATs) can have a higher correlation with KS4 scores than the KS2 scores (Strand, 2006) and what this means in terms of measurement alignment. If CAT scores are a better measure of underlying academic performance, this means that some part of the KS2-KS4 VA scores stems from poor measurement alignment. Also consider the composite nature of measures of academic performance. There are known problems of comparability of the subjects within and across composite measures (Coe, 2010). Some differences in value-added will stem from changes in the composite performance measure which differentially affects pupils (i.e. where the emphasis between different pupils' relatively strong or weak areas of performance are (de)emphasised). This point can be made more generally: even within a single subject, mathematics for example, the nature of the subject changes over time in terms of the balance of topics and skills. Such shifts in demands will differentially affect different pupils, raising further implications for measurement alignment across two points in time. In sum, to the extent that measures of prior and final attainment fall short of being isomorphic measures of performance, bias will be introduced into the measure.

Second, Kelly and Downey (2010) point out the problem of ceiling and floor effects in the data and how approximately one third of the pupils in the CVA pilot got the top grade at KS2. This prevents the model distinguishing between 'genuine' pupils getting the score and very high-ability pupils who would have received a higher grade had the exam allowed this (and would be predicted a higher grade in future performance). This problem can apply to the final performance measure as well as the prior achievement measure used and similar problems can be identified for floor effects, especially at primary-level, as will be explored in the empirical results in this thesis.

The third and final issue considered is the impact of missing and erroneous data. Even in high quality dataset, missing data can be a problem, especially for contextual variables (Gorard, 2012a). Ideally, errors and missingness would be random, in the sense that they are independent of school membership, although this is not necessarily the case (Gorard, 2012b). One issue examined in the results chapter is the extent to which there are missing data in the NPD for key attainment and contextual variables. There has been a recent debate about the nature and seriousness of errors within value-added calculations. The impact of error on value-added measures is something which is very difficult to directly observe and how these should be understood is contested. As a result, this is something which is returned to in the section on interpretation (Section 4.3) and, more specifically, the sub-section on uncertainty, biases and error which examines the differences in how these threats to validity have been understood.

4.2.4 Stability

This section and the next concern stability and consistency of school value-added measures, respectively. Study of consistency and stability has formed an important and ongoing part of educational effectiveness research referred to as 'foundational studies' (Scheerens, 1993a, Teddlie and Reynolds, 2000). This strand of research has studied the properties of the school effect revealed through value-added analysis with a view to 'resolving basic conceptual questions regarding the construct of school effectiveness' (Scheerens, 1993a, p.17). For the construct to be meaningful, it must have certain properties, as (Scheerens, 1993a, p.21) explains:

> "The concept "school effectiveness" has connotations of duration and scope. That is, in order to call a school effective, high achievement levels should persist over time (stability) and effectiveness judgements should not be based on the functioning of just a partial segment of the total organization (scope)."
>
> (Scheerens, 1993a, p.21)

As is discussed at length in Section 4.3, trying to understand the properties of a construct without knowing the validity of the measure is highly problematic. If a school value-added score is unstable (or inconsistent), does this reveal that the measure is heavily influenced by the numerous sources of CIV discussed above, or does it just indicate that the effectiveness of schools is highly changeable (or multifaceted)? This simple question underpins many of the issues of interpretation discussed in Section 4.3. Presumably, the truth is that both of these are to some degree correct. If this is the case, this raises the question of how one judges when variation reveals information about the school effect and when it reveals something about validity. The next section considers the issue of interpretation further, drawing on the empirical evidence presented in this section on stability and the next (Section 4.2.5) on consistency.

Existing evidence has consistently shown a considerable degree of instability in value-added measures, although this has been interpreted in remarkably different lights by different researchers. Depending on your perspective and the particular dataset, value-added scores have a "fair degree of stability" over time (Teddlie and Reynolds, 2000, p.126); 'broad stability in some areas' but 'also a substantial degree of change over time in some schools' (Thomas et al., 1997, p.193); show 'considerable stability' in adjacent years but are 'much more variable for larger periods' (Thomas et al., 2007, p.277); are of little value for school choice as correlations over more than a few years are low and uncertain (Leckie and Goldstein, 2009), 'are not particularly reliable or stable over time' (Marsh et al., 2011, p.286), or are 'almost entirely useless for practical purposes because [value-added] is not a consistent characteristic of schools' (Gorard et al., 2012, p.8). Some of these differences presumably relate to the specific dataset, the application and the model specification. There also seems to be large differences in interpretation of correlation scores. Luyten and de Wolf (2011), for example, described correlations (between school mean raw scores) across consecutive years of 0.66 and 0.61 as demonstrating 'considerable stability across years'. In contrast, Goldstein and Leckie (2008, p.68) state that "the correlation between school-effects for cohorts of children taking such exams 6 years apart *is only* about 0.6" (emphasis added), adding that, "In other words, exam performance now is a poor guide to performance in 6 years' time." While these are not exactly like-for-like comparisons in terms of what is being compared to what, we can nevertheless see clear differences of interpretation. Also, note that one would expect the difference in interpretation to be in the opposite direction if anything: Luyten compared raw scores which are generally found to be relatively stable across only two years (Luyten et al., 2005), whereas Goldstein compared school effects, which are generally less stable, certainly over a period of 6 years where one might expect larger changes in school performance.

Another difference in interpretation which is suggested by reading how researchers summarise correlation scores is whether a Pearson r correlation is considered in terms of the percentage of variance common to both variables (i.e. R^2). Gorard et al. (2012), for example, whose interpretation is that stability is low explicitly give the latter (followed by the former in brackets), most others make no mention of this distinction. Obviously these are two different ways of presenting the same substantive result, but this may make some difference in the substantive interpretation given that a Pearson r of 0.6 gives an r^2 of 0.36. At first glance, these give markedly different impressions of the level of similarity between the scores. More profound differences in interpretation are discussed further in Section 4.3. The remainder of this section reviews the empirical evidence on the stability of VA measure, starting by looking at the English secondary CVA measure.

Two studies which have examined the level of stability in the English KS2-4 CVA measure (2005-2010) are Leckie and Goldstein (2009) and Gorard et al. (2012). Gorard et al. (2012) present the correlations of school CVA scores 1, 2, 3 and 4 years apart, finding correlations ranging from 0.58 to 0.79, 0.48 to 0.67, 0.56 and 0.46 respectively. These results show that, even 1 year apart, there is only a moderate correlation in school CVA scores. Gorard et al. (2012, p.7) reach the conclusion that the CVA scores appear to be ‘meaningless’. These correlations are in line with earlier results from Leckie and Goldstein (2009) who estimate correlations 1, 2, 3, 4 and 5 years apart as 0.80, 0.73, 0.57, 0.46 and 0.40, respectively. Note that these correlations (i.e. in Leckie and Goldstein, 2009) were from a model which included compositional variables (discussed further below).

There is also research pertaining to other measures, ages and school systems. This research has produced largely similar results to those regarding the English CVA measure. In general, school-level value-added scores at primary level (age 4-11) are found to be even more unstable than those at secondary level, as might be expected given smaller cohort sizes at lower ages. Dumay et al. (2013) looked at value-added performance of different primary grades across time for 1, 2, 3 and 4 years apart, finding correlations of 0.40-0.53, 0.40-0.43, 0.36-0.40 and 0.29 respectively. The only thing that was stable was that the vast majority of schools had ‘indeterminable’ effectiveness (Dumay et al., 2013, p.75). As they pointed out, this low level of stability ‘poses a significant challenge to the conventionally accepted view that we can make a generalized evaluation of how effective a school is, based on cross-sectional data from a single cohort’ (Dumay et al., 2013, pp.78-79). Similarly, research into systems other than England has found very low correlations in performance. Marks (2014, p.14) estimated year-on-year correlations in VA performance in grades 5 and 9 as ranging from ‘from a very low 0.10 to 0.30 for Year 5 and from 0.16 to 0.50 for Year 9.’

Another study examining primary school stability – this time in Portugal - is Ferrão and Couto (2013). Ferrão and Couto (2013) analysed the sign of the value-added scores across the 3 study years, finding that 26% of schools had positive VA for all 3 years, 15% had negative scores for all 3 years, and 85% had the same sign for at least 2 adjacent years within the 3 years. Note that some level of consistency would be expected by chance alone (Gorard et al., 2012): figures of 12.5%, 12.5% and 75%, respectively. Correlations are not presented but from the scatter plots presented appear to be quite small. Ferrão and Couto (2013, p.186) conclude that “the findings reveal a systematic pattern of educational units’ performance is more than just randomness.” At this (rather low) threshold for value, they conclude that Portugal should include a VA indicator into its system of evaluation.

All of this suggests that value-added scores exhibit some degree of stability but that this is less than might be desired. Primary-level stability correlations are generally being estimated at less than 0.50 and often far lower than this. These are very low correlations in this context and mean that scores even 1 year apart typically show marked differences. Scores separated by several years bear hardly any relation to one another. At secondary level, the correlations are moderate and so the issue of whether these are meaningfully stable (i.e. reflective of a valid measure of school effectiveness) is more contestable. There are certainly grounds for serious concern with these levels of stability and a strong suggestion that the measures are appreciably comprised of measurement error and unobserved bias.

A problem in terms of generalising about stability is that differences appear to relate to the specific dataset and the model specification. There are important links which need to be made between stability, validity, data quality and model specification (Dumay et al., 2013). Regarding model specification, for example, Leckie and Goldstein (2009) show that not including compositional variables tends to inflate stability because bias carries through the scores over a number of years and so mean school intake achievement is a ‘major driver’ of between-school differences in later years. Consider this in light of the continuum (from raw scores to perfectly valid VA scores) considered above. Raw scores tend to be highly stable over time (Luyten et al., 2005, Dumay et al., 2013, Gray et al., 2001). This is because the characteristics of schools in terms of intake characteristics tend to be relatively stable and possibly also because schools aim for similar standards with successive cohorts, smoothing performance - the ‘stable target hypothesis’ (Dumay et al., 2013, p.78). As one corrects for non-school factor bias, and moves along the continuum, two things happen: a) the measure is likely to become more valid (assuming appropriate specification) and b) the measure is likely to become less stable. Value-added is a residual after accounting for the effect of non-school factors. As further non-school factors are removed, measurement error and other sources of CIV become larger relative to the residual value-added. Moreover, as noted in the previous section, it is likely that more complex contextual models with too many and poorly theoretically grounded control variables may be over-correcting differences between schools, removing some genuine school effect (Willms, 2003, OECD, 2008, p.126). As Allen and Burgess (2011, p.253) note, “CVA is unstable because it results from fitting a complex model with many imprecisely measured parameters.” Similarly, Gorard et al. (2012) note variation in the model coefficients over time as well as stressing a number of problems with the quality of the underlying data, especially in relation to the measured contextual variables. This particular problem serves as an excellent example on the linkages between the issues discussed so far: the choices of model specification and the quality of the data have serious implications for both stability and validity.

4.2.5 Consistency

Another ongoing strand of research within school effectiveness relates to the consistency of school effects across outcomes. As with the issues of stability, there is a blurred boundary between whether the results reveal information about validity or about the nature of the school effect itself. On one hand, inconsistency may reflect CIV at the level of the subject. Alternatively, inconsistency may reflect differences in school performance across different areas. In either case, consistency across outcomes provides evidence towards generalisations about school effectiveness (Teddlie and Reynolds, 2000). Several types of consistency are identified by Teddlie and Reynolds (2000): consistency across subject areas, outcomes, grade-levels and across alternative test modes. The issue of consistency is often discussed in terms of ‘differential effectiveness’ which essentially addresses the questions, ‘Effective at what?’ and ‘Effective for whom?’ (Bogotch et al., 2007, Sammons, 1996). Note that this thesis has preferred the term ‘stability’ (see last section) to the synonymous ‘consistency over time’, which is sometimes used in the context of stability (e.g. Sammons, 1996, p.140). This review sorts issues of consistency into three groups: 1. Levels of consistency for the same pupils across different outcomes, 2. Consistency of school effects for different groups of pupils according to their characteristics (e.g. ability or socio-economic status) and 3. Level of consistency of school effects for different cohorts (year groups) of pupils at a single point in time. There is also a final section summarising the evidence on teacher-level value-added. This final sub-section considers both stability and consistency of teacher-level value-added scores. Stability is also considered alongside consistency in some of the other sub-sections in order to show the links between them.

Consistency of school effectiveness measures across outcomes

It has been common practice in school effectiveness studies to collect several outcome measures when estimating school effectiveness. As a result there are numerous examples of studies which have presented evidence concerning the consistency of school effectiveness estimates across subjects. This body of evidence suggests that there are moderate correlations between school effects across different academic subjects (Teddlie and Reynolds, 2000, Reynolds et al., 2014), although results varied across studies and are likely to depend on the subject and stage of education (Mortimore et al., 1988, Bosker and Scheerens, 1989, Luyten, 1994, Sammons et al., 1996, Reynolds et al., 2014). The key study in the feasibility of using VA as part of a national school performance monitoring system found correlations between subjects were higher at primary level than secondary level and recommended a profile of measures at secondary level (Fitz-Gibbon, 1997). There was a growing consensus in the research community that school effectiveness should be considered ‘multi-faceted’, with Thomas (2001, p.285) suggesting at least 4 dimensions are apparent: outcomes, pupil groups, cohorts and curriculum stages. Thomas also conducted research into wider outcomes such as pupil attitudes (Thomas et al., 2000), finding weak correlations between cognitive and affective outcomes. A subsequent review by Gray, drawing on Thomas’ work and earlier studies (e.g. Knuver and Brandsma, 1993, Smith et al., 1989) seemed to have confirmed this result: the links between school effectiveness in cognitive outcomes and wider outcomes such as attitudes, attendance or participation ‘appear weak to non-existent’ (Gray, 2004).

More recent research has continued to examine how value-added performance differs by outcome. Telhaj et al. (2009), for example, reviewed the case for performance indicators to be provided at departmental level, finding considerable differences between the performances of departments. Telhaj et al. (2009) also examined the stability of the measures for different departments and found that value-added departmental-level performance is ‘highly unstable over time’. Their conclusion was as follows:

> “Overall, these results suggest two possible interpretations. Either relative departmental performance is being largely driven by random fluctuations in exam performance that we have not captured in either our unadjusted or value-added measures or competitive pressures within schools tend to ensure that differences in relative performance are not sustainable over time.”
>
> (Telhaj et al., 2009, p.17)

It is valuable to consider the issues of stability and consistency alongside each other and note the conflict: Performance measures which aggregate the scores from a number of subjects tend to be considerably more stable over time (Teddlie and Reynolds, 2000). Given that Telhaj et al. (2009) found considerable differences between subjects, there is a strong case that departmental-level measures are more valid; yet, providing departmental-level measures exacerbates the problem of stability to the point that scores are hard to distinguish from ‘random fluctuations’ (Telhaj et al., 2009, p.17). Note that the level of stability of the English CVA considered in the last section should be viewed therefore in light of the fact that it is an aggregated measure. If the measure was reformed to reflect departmental-level differences, this is likely to increase its (construct) validity but this is also likely to make it less stable; yet lower stability suggests lower validity caused by a greater proportion of volatile CIV in the measure. There is an important question here as to the level at which measures are most meaningful and most valid as well as issues relating to the potential of aggregation to, on one hand, mask problems of validity and, on the other, smooth (cancel out) CIV, improving validity.

In sum, research has shown there to be moderate consistency in school effects between academic attainment outcomes. However, school effectiveness estimates for academic and other wider outcomes of schooling, while less extensively studied, have been found to be largely independent.

Consistency of school effectiveness measures across pupil groups according to their characteristics

School value-added measures give a mean value of pupil-level value-added estimates. Another ongoing concern of school effectiveness researchers is to break down these mean values by pupil groups such as gender, ethnicity, socio-economic status and academic ability (Teddlie and Reynolds, 2000). As well as giving further insights into school effectiveness (is an effective school effective for *all* its pupils?), this can be seen from the point of view of equity and ensuring the results of minority groups do not get hidden within the average results (Smith et al., 1989) as well as concerning the provision of information to parents about school performance which is targeted towards their children (Allen and Burgess, 2011). At the time of writing the *International Handbook of School Effectiveness and Improvement*, the evidence on differential effects by gender, ethnicity, ability and socio-economic status was 'inconclusive' despite there being a number of high-quality studies to draw on (Teddlie and Reynolds, 2000). Subsequent research tended to support this general finding, suggesting that while differences are evidenced in some studies, school effectiveness does not systematically differ (i.e. at system level) to a great extent according to pupil characteristics (Thomas, 2001, Strand, 2010, Reynolds et al., 2014). Consequently, this remains a school-level issue, where evidence is often sought regarding the relative effectiveness of particular schools in narrowing the gap between groups. Differential effectiveness within individual schools is now reflected in the English school performance tables where results are broken down for lower, middle and higher pupils and for pupils who are disadvantaged (DfE, 2015). Considering performance of sub-groups gives schools an incentive to focus on the performance levels of all pupils in the school. Justification for this can be found from studies such as Dearden et al. (2011a, p.225) who conclude that "even the most conservative estimate suggests that around one-quarter of schools in England are differentially effective for students of differing prior ability levels."

Consistency of school effectiveness measures across pupil cohorts

One final consistency issue is the differential performance of cohorts (otherwise known as year groups or grade levels) in a school at a single point in time. There are very few examples of studies which have been able to present evidence concerning the consistency of relative value-added performances across different cohorts at the same point in time in the same school (Teddlie and Reynolds, 2000). An early key study identified by Teddlie and Reynolds (2000) is Mandeville and Anderson (1987) who find 'discouragingly small' correlations and characterise consistency between grades 1 through 4 as 'very unstable' (p. 212 & 203, respectively). Their results pose(d) a serious question for school effectiveness research:

> "…the results should cause effective schools researchers to rethink the meaningfulness of the concept of an effective school. Apparently how well students in one grade in a school achieve, when achievement is gauged against prior achievement and SES, is only weakly related to the achievement of students in other grades in the school. This inconsistency contradicts a model that posits school 'main effects.'"
>
> (Mandeville and Anderson, 1987, p.213)

Subsequent work (Bosker and Scheerens, 1989) suggested moderate correlations between grade levels but urge caution, suggesting the figures were possibly inflated (see Teddlie and Reynolds, 2000). As Teddlie and Reynolds (2000, p.118) noted in 2000, the question had 'not been adequately researched'. A literature search conducted during this study was unable to find any more recent studies. One related study, however, is Taylor and Nguyen (2006, p.215) who found that the correlation between school performances at Key Stage 3 (KS3) had a very low correlation (r = 0.26) with performance across Key Stage 4. It is difficult to know to what extent this is related to problems with the KS3 tests but this suggests that consistency in the performance for different year groups within a school could be low. Cohort consistency is likely to have received little attention due to the shift in focus from schools to teachers in educational effectiveness research (Muijs et al., 2014) and the assumption (e.g. Bosker and Scheerens, 1989) that where performance is inconsistent across year groups, this reflects differences in the effectiveness of teachers rather than a problem of measurement. Increasing use of multi-level modelling to partition variance between pupils, teachers and schools found unexplained variance at teacher level tended to be higher than that found at the level of the school (Luyten, 2003), gradually leading to a greater focus on teacher effects (Muijs et al., 2014). Use of school-level measures by English policy makers continues nonetheless; the school tends to be the key unit of concern with respect to effectiveness from a policy and organisational perspective.

In sum, the very limited evidence available suggests that consistency between different cohorts' value-added performances within schools at a point in time is low. As a result, consistency is the focus of one of the key empirical studies presented in the following chapters. Given the lack of evidence in this area, one aspect of this consistency study that examines cohort consistency is one of the most original (and important) empirical contributions in this thesis. Consistency is an important concern for several reasons: first, the value-added measures in the English performance tables only concern a single cohort in each calendar year. Moreover, the measure concerns the year group which completed its examinations the previous year and is in many cases now no longer at the school. Whether the performance of a cohort can be generalised to the school more widely is an important issue for the interpretation of the measure and if this is not the case, the measures will have little formative value. Second, changes in the characteristics of different cohorts (year groups) is often suggested as a reason for observed levels of instability over time (Marks, 2014, Dumay et al., 2013). Examining the consistency of estimates between cohorts allows the source of instability across time to be more precisely located by ruling out real changes in school effectiveness over time as a source of the instability. If cohorts are consistent, it suggests that it is the schools' effectiveness that is changing. If inconsistent, this suggests that the differing cohorts passing through the key stage examination years may be a source of the variation; either that, or a more general problem of instability and inconsistency caused by CIV.

The Consistency and Stability of Teacher-Level Value-Added

As explained in the introduction, a comprehensive examination of teacher-level value-added is beyond the scope of this study. Nevertheless, the validity of school-level and teacher-level value-added are obviously very closely linked. On one hand, the properties of teacher or classroom-level effects can be seen as an issue of consistency as it concerns lower-level units within the overall construct of school effectiveness; on the other hand, teacher-level value added could run alongside all the sections above. As a result, although situated in a section on consistency, this sub-section looks at teacher-level value-added more generally and links key findings to the information above. Reviewing selected papers concerning teacher effects is especially valuable as there are a small number of studies which employ an experimental or quasi-experimental design. These designs can potentially account for unobserved differences between pupils. The evidence examined in this section comes from the United States of America where value-added is a current issue. Teacher-level value-added is being used as a substantial component of teacher evaluation in a large number of US states. New York City's Education department even went as far as publically releasing teacher value-added rankings, a move which proved highly controversial and has led to legal cases (Amrein-Beardsley, 2014). Also, there have been statements issued by the American Statistical Association (ASA, 2014) and the American Educational Research Association (AERA, 2015) outlining the limitations of value-added and cautioning against their use without sufficient evidence that a number of technical requirements have been met. Proponents of value-added have responded (to the former) that the concerns 'have been addressed in recent experimental and quasi-experimental studies' (Chetty et al., 2014a, p.112). These experimental studies are also cited in a response to the latter AERA statement:

> "We now know that teachers vary dramatically in their impact on student learning as measured by test scores (Chetty et al., 2011; Kane & Staiger, 2008; Nye, Hedges, & Konstantopoulos, 2004)"
>
> (Raudenbush, 2015, p.138)

Let us consider these three studies. The first study, Nye et al. (2004), reanalysed data from a project known as the Tennessee Class Size Experiment, or project STAR (Student-Teacher Achievement Ratio). This involved 79 elementary schools where within each school a cohort of pupils were randomised to classes (with various class sizes) and teachers were randomised to these classes. This cohort was followed for several years, maintaining the differences in class-size. In each successive year, new teachers were randomly assigned to the classes. This randomisation of pupils means that variation in outcomes can be assumed with a fair degree of confidence to reflect either a) the experimental differences in class size, b) the effectiveness of the teachers or c) threats to the internal validity of the study (such as measurement error or some kind of experimenter effect) (Shadish et al., 2002). In what was an apparently well-designed study with successful randomisation, Nye et al. (2004, p.253) found between-classroom (i.e. teacher) variance of about 13% in maths and about 7% in English. These estimates were broadly in line with previous correlational estimates reviewed which had a median value of 11%. School effects were found to be smaller than teacher effects with average estimates of 6% and 5% for Maths and English, respectively (although note that pupils were not randomised to schools) (see Luyten, 2003, Reynolds, 2008, for further information on the relative size of school and teacher effects). The estimated teacher effects were of an appreciable size:

> “The difference in achievement gains between having a 50th percentile teacher (an average teacher) and a 90th percentile teacher (a very effective teacher) is about one third of a standard deviation (0.33) in reading and somewhat smaller than half a standard deviation (0.46) in mathematics.”
>
> (Nye et al., 2004, p.253)

On the basis of these results Nye et al. (2004, p.253) concluded that teacher effects are therefore ‘real’ and the effect sizes are ‘certainly large enough effects to have policy significance.’ As they note, the experimental design provides far more robust estimates than previous studies.

The other two studies mentioned have not been able to match the strength of the randomised design in Nye et al. (2004). Kane and Staiger (2008) reported results not based on a true randomised experiment: instead of randomising pupils and teachers to classes, only the teachers were randomised to classes which had been declared acceptable by school principals (Amrein-Beardsley, 2014, pp.169-170). Another problem with this study is that Kane and Staiger (2008, p.4) found that “…the impact of the randomly assigned teacher on math and reading achievement faded out at a rate of roughly 50 percent per year in future academic years. In other words, only 50 percent of the teacher effect from year t was discernible in year t+1 and 25 percent was discernible in year t+2.”

The third paper cited by Raudenbush (2015), above, was Chetty et al. (2011). This was eventually published as two papers (Chetty et al., 2014b, Chetty et al., 2014c), the first of which employed a quasi-experimental design. Chetty et al. (2014b) examined whether cohort value-added scores changed as would be expected when teachers moved schools. As Chetty et al. (2014b, p.29) explained, “for example, in a school-grade cell with three classrooms, the loss of a math teacher with a VA estimate of 0.3 based on prior data should decrease average math test scores in the entire school-grade cell by 0.1.” Teacher VA was calculated using data outside the five year ‘event’ window (the year of the teacher’s move and two years before and after). They estimate that entry/exit of a high-VA teacher (based on the teacher being in the top 5% of the distribution) raises/lowers the mean test score in the grade by 0.04/0.05 (1SF) standard deviations, respectively. Entry/exit of a low VA teacher increases decreases/increases scores by 0.02/0.03 (1SF) of a standard deviation, respectively (Chetty et al., 2014b, p.2620). The authors concluded that "value-added models which control for a student's prior-year test scores provide unbiased forecasts of teachers' causal impacts on student achievement. Because the dispersion in teacher effects is substantial, this result implies that improvements in teacher quality can raise students' test scores significantly" (Chetty et al., 2014b, p.2630). This conclusion is surprising given the tiny size of these effects, especially without any estimates of stability or evidence ruling out the possibility that effects during periods of personnel change may not reflect typical levels.

As at school-level, stability is a known issue with teacher-level value-added. Summarising a large number of studies looking at the stability of teacher effects estimating using value-added designs, Amrein-Beardsley (2014, p.134) report that stability correlations of teacher-level value-added range between zero and 0.5, typically in the region of 0.2-0.4. The mid-point of this typical range (0.3) means that teacher value-added scores from a single year explain less than 10% (0.3^2) of the differences in scores one year later.

Collectively, these studies suggest that teacher effects are small but of a sufficient size to warrant some attention; there are, however, concerns about their stability which may limit the practical value of teacher-level value-added estimates and/or differences underlying teacher effectiveness. Even if the studies above were held to demonstrate that teacher value-added is unbiased, the size and instability of the scores suggests that teacher-level value-added is of very little practical value and likely to be highly damaging if used in a high-stakes context (Amrein-Beardsley, 2014).

4.3 Interpretation of Value-Added Evidence and Methodological Assumptions

4.3.1 Section introduction

One of the most striking things about the literature discussed above is the range of views which are based on what is a fairly large and broadly uncontested evidence-base (Gorard, 2010). On the stability of value-added, for example, positions range from value-added being 'rather stable' (Reynolds et al., 2014, p.207); to it being so volatile that it is 'meaningless with current datasets as well as useless or worse than useless for practical purposes' (Gorard et al., 2012, p.7) (see above). On validity in general, researchers such as Marsh et al. (2011, p.279) observe the 'fragility of causal inferences based on psychometric and econometric value added models', question whether VA models adequately control for pre-existing differences and summarise the properties of VA as follows (also see Coe and Fitz-Gibbon, 1998):

> "Estimates of CVA/VA are primarily summative measures based on one single indicator of effectiveness (student test scores). They are based on some problematic statistical assumptions, have large standard errors and are not particularly reliable or stable over time. They are not particularly useful for formative purposes of improving effectiveness in relation to the characteristics of teachers or what they do that contributes to effectiveness."
>
> (Marsh et al., 2011, p.286)

Consider this assessment alongside the finding in Chapter 3 that research in the *School Effectiveness and School Improvement (SESI)* journal was almost exclusively based on methods related to the value-added approach. It is difficult to square these negative evaluations of validity with the prevailing views and practices within the SESI research community - even when making distinctions relating to policy/research use and taking into consideration the more sophisticated and often teacher-level analysis currently being conducted by researchers.

It is held here that differences in data, models and application are indeed important (as discussed in Chapter 7). Nevertheless, differences in evaluations of validity are not held to be entirely reducible to specific contexts and applications. This thesis also argues that these differences relate to (sometimes profound) differences in interpretation and understanding. It is the aim of this section (Section 4.3) to review these differences and examine their influence on the debates around validity. To do this, I review a series of recent debates between Stephen Gorard, a critic of the use of value-added, and prominent educational effectiveness researchers who use value-added methods extensively in their research. These debates can be seen as a recent incarnation of more longstanding concern about methods and validity of value-added and educational effectiveness research (EER) (see Bosker and Scheerens, 1989, Scheerens, 1993a, Coe and Fitz-Gibbon, 1998, Luyten et al., 2005, Gorard, 2010, Reynolds et al., 2012). The spark for the Gorard-EER debates was a paper by Gorard (2010) which expressed 'serious doubts about school effectiveness' (as titled). This paper was followed by a British Educational Research Association (BERA) conference debate in 2011, a response piece from educational effectiveness researchers in Muijs et al. (2011), a reply in Gorard (2011b), further criticism in Gorard (2011c) as well as lengthier responses in Reynolds et al. (2012) and Gorard (2011a). This section (Section 4.3.2) and the next (Section 4.3.3) heavily draw on these debates which air conflicting perspectives on several key issues relating to matters of interpretation. To a large extent, these debates address core issues of the validity of value-added and represent the current state of thinking on the matter from proponents and opponents of value-added. The aim of these sections is to present the problem of justification and both sides of the argument as explicitly as possible and reach a considered, independent position.

4.3.2 The Problem of Justification: Bias, Error or Effect?

This section returns to the problem of justification first introduced in Chapter 1. Recall that value-added is, strictly speaking, an unexplained mean difference in pupil performance (i.e. the residual of a value-added model). Under what conditions, then, can it be said to be a school effect? This problem of justification does not have an easy or definitive answer. As Gorard (2010, p.758) observes, 'whatever the residuals [i.e. VA scores] are, we simply do not know if they are error or effect.' The only requirement for interpretations to be possible from value-added evidence is the presence of variation in the scores. Irrespective of the source of the variation, one can claim to have identified more and less effective schools. It is, as Gorard (2011c, p.39) observes, like backing all horses in a horse race but drawing attention only to the backing of the winner after the fact. How can one know that the school identified through the value-added method is in fact more or less effective?

Broadly, there are two possible approaches to justification (both of which are explored in empirical sections of this thesis): First, we might say it is a school effect if all the theoretically important non-school factors have been accounted for. As described in Section 4.2.2, however, controlling for non-school factors is problematic, incomplete and the result can only ever be approximate. Knowing how good this approximation is requires an estimate of the extent to which it is affected by unobservable bias and error; but how does one know if an estimate of a latent property is influenced by unobserved bias or error? The best approach to doing so would involve a robust research design which is capable of dealing with unobserved differences (e.g. experimentally) (Shadish et al., 2002, Gorard, 2013). As robust evidence is not available (other than the inconclusive evidence at teacher-level, see above), there is a question mark over the results (Coe and Fitz-Gibbon, 1998) and it remains difficult to answer Gorard's (2010) question: If the value-added estimates were largely comprised of measurement error or unobserved differences, how would we know?

The second approach to answering the question of validity considered here is to look estimates in context. Contextualising results is strongly encouraged by leading educational effectiveness researchers (Teddlie and Reynolds, 2000, Chapman et al., 2015) and forms part of the response to Gorard (2010) in Muijs et al. (2011) and Reynolds et al. (2012):

> "Researchers have repeatedly emphasised that school effectiveness is a relative and retrospective concept that is both outcome- and time-dependent, and that as a consequence there is a need to study consistency, stability and differential effectiveness covering variations in different outcomes, including departmental effects for secondary schools, and trends over time and for different groups of pupils (Luyten & Sammons, 2010; Creemers et al., 2010)."
>
> (Muijs et al., 2011, pp.3-4)

The reasoning here is that looking at the results in the context of others will give an indication of whether what is being captured is meaningfully stable and/or internally consistent and so avoid being misled by erratic isolated scores. As noted earlier, one would assume that VA scores for a given school ought to maintain some degree of consistency/stability (Bosker and Scheerens, 1989, Scheerens, 1993a). Faced with volatile and inconsistent school value-added scores, however, one must doubt either the validity of value-added or the common public understanding of the school effect as large, stable and consistent (as is the case in Marks, 2014). This choice cannot be sidestepped: if one is to maintain the assumption that what is observed is effect rather than error, it is also necessary to accept the properties of the school effect which are found. If these cannot be accepted, the face validity justification does not hold. This reasoning is ostensibly an approach to identifying invalidity, even if it does not establish validity (as it cannot account for stable or consistent biases).

The principal difficulty of this approach is that it requires prior knowledge (or assumptions) regarding the level of stability and consistency in a valid measure in order to identify an invalid one. Herein lies the contradiction of examining the properties of value-added school effects to, on one hand, understand more about school effectiveness and, on the other, to avoid erroneous inferences. Allowing understanding of the construct to change according to stability and consistency evidence exacerbates the problem of justification: The evidence reviewed above suggests that value-added has considerable levels of instability and inconsistency; yet, what evidence could possibly persuade us that the value-added scores were largely comprised of error or bias rather than effect if all instability or inconsistency is interpreted as changeable school performance, 'differential school effects' or, generally, 'complexity'? Our understanding of the construct will simply change to fit the evidence. At the extreme, when all differences are, by *a priori* assumption, effect rather than construct irrelevant variance (Gorard, 2011a), value-added evidence will become completely unfalsifiable (Popper, 2005). Two other problems related to the conceptual framework are that, first, when school effectiveness factors are defined vaguely (Coe and Fitz-Gibbon, 1998), this reduces the strength of a justification of value-added based on the consistency of EER findings (Reynolds et al., 2012). As Coe and Fitz-Gibbon (1998, p.430) note, "…the apparent repeated confirmation of these factors depends in part on the vagueness of their formulation." A second additional problem for justification is the practice of identifying school effectiveness factors in a post-hoc analysis of the data, referred to as 'fishing' for correlations (Scheerens, 1993b, pp.67, cited in, Coe and Fitz-Gibbon, 1998, p.429, Luyten et al., 2005, p.254) or data 'dredging' (Gorard, 2015, p.86). All of these problems raise the concern that the conceptual framework within which value-added is interpreted may be so flexible that just about any eventuality will allow a plausible explanation of the results in terms of school effectiveness.

The only way to guard against erroneous inferences (but see Section 4.3.3) is to have some limit on what one considers plausible regarding the properties of value-added scores (i.e. one must rely on the face validity of the evidence) and the central question becomes one of where to draw the line and so whether the extant evidence suggests this line has been crossed. The issue tends to come down to answering the following question, posed by Gorard (2010):

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

Critics such as Gorard answer in the negative, concluding that the effect is too small, inconsistent and unstable and the error is relatively large, intractable and non-random (i.e. biased) for the value-added method to be of value. Conversely, defenders of value-added view the effect as something that is sufficiently large, stable and consistent to be of value and the error as a much smaller component of the variance that tends to be random (Reynolds et al., 2012) and amenable to technical solutions (Muijs et al., 2011). This positive view is significantly aided (see above) by the prevailing understanding in EER that a fair degree of instability and inconsistency is to be expected due to school effectiveness being a changing and complex construct (Sammons, 1996, Reynolds et al., 2014); while proponents accept 'uncertainty' in the estimates, they hold that it is nonetheless possible to distinguish effect from bias and error to a practically valuable degree and dismiss weaknesses of conceptualisation and measurement as 'primarily a technical problem with a technical solution' (Muijs et al., 2011, p.4). Understanding error and uncertainty is a crucial issue of interpretation to which we now turn.

4.3.3 Understanding Uncertainty, Bias and Error in Value-Added Estimates

Section Introduction

The last section described difficulties and differences of interpretation relating to inference in general. This section specifically examines how different conceptions of error and approaches to dealing with it have led to further and more profound differences of interpretation. Understanding error and uncertainty is a theme throughout the Gorard-EER debates. Approaches to dealing with error that make use of statistical testing and/or present results along with confidence intervals to convey their uncertainty are ubiquitous in all areas of use reviewed in Chapter 3. Yet, these are strongly opposed by Gorard who claims that the use of inferential statistics to deal with error in EER is 'fatally flawed' (Gorard, 2010, p.755).

Moreover, the principal issue on which Gorard (2010, p.748) bases his view that value-added is ‘rather meaningless’ concerns the seriousness of error - in particular the potential for errors to ‘propagate’ (see below).

In Section 4.3.3, several issues are considered in detail given the serious implications for the interpretation of value-added and the difficulties of understanding and communicating uncertainty in value-added estimates. These issues are organised in three main sections: first, a section on the nature and seriousness of error; second, an examination of the theoretical justification of construing uncertainty as stemming from sampling error and; third, a consideration of the merit of applying the theoretical justification in a practical context.

The Nature and Seriousness of Error

A central part of Gorard’s (2010, p.748) case against value-added and school effectiveness research more generally is his view that measurement error within value-added models will ‘propagate’. Gorard observes that value-added scores are created by finding a difference between the actual and the predicted performance; both of these are measured with error, the former directly (in an examination) and the latter because of the errors in the measured variables (e.g. prior examination attainment) which form the prediction model. Both of these errors can be either positive or negative and so can ‘propagate’ as well as cancel out; moreover, the difference between predicated and expected performance (with expected value, zero) will be considerably smaller than the original exam scale. Potential propagation combined with the smaller magnitude of the new value-added scale will mean that the error on value-added estimates will be much larger relative to that on the original examination scores, which Gorard argues will already be sizable.

Educational effectiveness researchers (Muijs et al., 2011, Reynolds et al., 2012) have responded with the following criticisms of Gorard’s argument: first, that the predicted score is not an attempted measurement. As per the description above, this is technically true but irrelevant since the prediction model uses variables (which are not error-free) as its basis. The Key Stage 2 scores, like the Key Stage 4 scores are subject to considerable measurement error (He et al., 2013). Second, Reynolds et al. (2012) point out that the example of relative error in Gorard (2010) gives the range of the error relative to a hypothetical estimate rather than the commonly accepted meaning of ‘relative error’. Again, this seems to be true but a point of semantics beside the substantive point about the potential size of errors in relation to value-added estimates (Gorard, 2011a). The more fundamental disagreement relates to whether, as Reynolds et al. (2012, p.8) claim, errors ‘tend to be randomly distributed’. If this is the case, pupil-level errors will generally cancel out and error is ‘unlikely to be systematically different in different schools’ (Reynolds et al., 2012, p.8). Moreover, because errors are random, they can ostensibly be estimated within the framework of statistical significance (see below). Reynolds et al. (2012) cite several EER studies which examine the influence of random measurement errors within multi-level models which suggest that measurement error is more of a problem for fixed effects coefficients of value-added models (see Chapter 2) than school value-added (Goldstein et al., 2008, Ferrão and Goldstein, 2009). Gorard’s response is that errors are not necessarily (or even presumably) random and ‘it is unfair and unethical to assume glibly that they will appear random once aggregated for each school’ (Gorard, 2011a, p.18).

Without further empirical evidence or a clearer understanding of the extent to which either school-level errors occur or pupil-level errors translate into school-level estimates, it is difficult to know which of these positions is closer to the truth and the seriousness of the problem. Gorard (2011a) discusses the example of the free school meals (FSM) variable (an indicator of poverty in the English system) and the finding (later published in Gorard, 2012b) that rates of missingness of FSM data are high and that these – rather than being a random subset – appear to be a super-deprived group. This is an example of where missing data and measurement error will almost certainly have sizable effects on school-level value-added for schools which take a large number of pupils with missing FSM. As in previous areas, some of the disagreement may relate to the outputs of the value-added method (see Chapter 2). The reply to Gorard in Reynolds et al. (2012) largely concerns the effect of (random) error on model coefficients and the proportion of variance situated at school-level. Gorard (2010), on the other hand, largely discusses the effects of (non-random) errors in relation to school value-added scores. While this difference of focus certainly does not dissolve the disagreement, it may have some bearing on what are marked contrasting views on what is ostensibly the same issue.

One final consideration concerning the nature of error relates to its magnitude. While error is certainly present, Gorard (2010) holds that value-added is disproportionately comprised of error. Proponents of value-added replying that they are well aware of the problem of error but take a ‘more prosaic, less hysterical interpretation’ of the likely rates of error in the estimates and claim that such errors can be taken into account using confidence intervals and related tests of statistical significance (Muijs et al., 2011, p.3). This application of statistical tools related to significance testing is an issue to which we now turn.

Uncertainty as Sampling Error

A key theoretical issue for the interpretation of value-added scores is how to understand and communicate ‘confidence’ in the value-added scores. Given the importance of this issue for many practical uses, it is of value to outline the conventional understanding in relation to uncertainty and discuss whether this framework is suitable for capturing and communicating the threats to validity which are considered in this thesis.

In a paper titled ‘Understanding Uncertainty in School League Tables’, Leckie and Goldstein (2011) consider how the CVA measure is presented by the Department for Education (DfE) and subsequently communicated to the public at large via ‘league tables’ in national newspapers. Leckie and Goldstein (2011) discuss confidence intervals and criticise the omission of confidence intervals from ‘league tables’ produced by the media before going on to analyse issues relating to school comparison. The following is what is written about confidence intervals and their omission in the media-produced league tables:

> "In the CVA tables, each 95 per cent confidence interval quantifies how precisely each CVA score is estimated… By not presenting the confidence intervals, the media present CVA scores as if they were free from sampling error and therefore as if they were completely reliable estimates of school effectiveness. This is far from true. Indeed, the sampling error of CVA scores is so great that, nationally, only around a half of schools are statistically distinguishable from the national average."
>
> (Leckie and Goldstein, 2011, pp.209-210)

This passage exhibits several commonly-held ideas: first, that unreliability in the CVA estimates is due to sampling error. Second, that confidence intervals estimate this sampling error and so provide an indication of the error rates in the CVA estimates (although the difficulties of this are discussed in the paper). Third, that statistical significance is the/an appropriate benchmark against which to judge differences in the CVA of schools. Leckie and Goldstein's paper is typical of the vast majority of commentators in this area construing uncertainty in terms of the conceptual and analytical framework of inferential statistics. Wilson et al. (2008), for example, state that confidence intervals 'take account of the uncertainty involved in using a set of test results, achieved by one set of pupils on one day, as a measure of the underlying effectiveness of the school." Similarly, leading educational effectiveness researchers hold that the use of confidence intervals 'mark[s] a recognition that the effectiveness measures are estimates subject to error' and have 'consistently advocated' their use (Reynolds et al., 2012, p.14 & 8). This understanding of confidence intervals as being the appropriate tool for estimating and communicating uncertainty is also shared by the DfE, whose guidance describes confidence intervals as 'the range of scores within which each school's underlying performance can be confidently said to lie' (DfE, 2014a, p.9).

This understanding is based on probability theory and the methods of inferential statistics and so rests on a longstanding but nonetheless current debate about the value of inferential statistics within social science data. There are numerous criticisms of inferential statistics including problems with the underlying logic, the relevance to data typical of social science and potential for misunderstanding from both researchers and users (Gorard, 2015). Consideration of the underlying logic of inferential statistics is beyond the purview of this project (but see Trafimow and Rice (2009) and Gorard (2015) for criticism and Neale (2015) for a response). There are two particular issues, however, which require attention given their importance to the interpretation of value-added measures. The following two questions can be raised in response to the dominant view of uncertainty as described above: first, are inferential statistics (statistical significance, confidence intervals, p-values etc.) suitable for use with data such as the NPD given that they are population data as opposed to random samples (Gorard, 2015)? The second question concerns the extent to which uncertainty can be reduced to what is a narrow, technical (and therefore tractable) issue. Of course, if the answer to the first question is negative, the second question is irrelevant.

First let us consider whether calculation of statistical significance is even appropriate with population data such as the NPD. The objection to this practice is that, as these data are for all pupils in England, there is no need to draw on sampling theory to estimate whether the results will generalise to the wider population from which the sample is drawn. What is one generalising to? Moreover, even if the data were construed as being a sample from a larger ‘super-population’ and we were to accept this as meaningful, the data cannot be a random sample of this super-population. No sampling has taken place, let alone random sampling. What does it mean when Leckie and Goldstein (2011), above, explicitly cite sampling error as a problem (Gorard, 2015)?

In a published open dialogue on the ‘widespread abuse of statistics by researchers’, Gorard (2014) strongly criticises the use of techniques designed for random samples with non-random samples or populations. Many respondents agreed: Glass (2014, p.12) stated that “the fiction that probability statements are meaningful in the absence of random acts underlying them is preposterous”, Howe (2014, p.14) commented that “it is certainly inappropriate to use techniques drawn from sampling theory when working with populations” and White (2014, p.25) described the practice as ‘meaningless’. Putwain (2014, p.18) was more equivocal, arguing that the use of non-random samples is a ‘pragmatic decision’, where a non-random sample is ‘treated *as if* it were random’ but notes that “if one is a purist, these practices are not acceptable; if a pragmatist, then they are permissible.” The only respondent opposing Gorard’s view at any length is Styles (2014) who rehearses the case given by proponents of the practice (see below) using VA as an example:

> “The concept of a virtual population is often used without acknowledgement, for example, when assigning a confidence interval to a school’s value-added results even when all students are measured (Goldstein, 2008)… we imagine the trial being run many times on students in the same schools at the same time in a virtual population from which we did sample randomly. This allows us to quantify chance and gives meaning to the p-values and confidence intervals used. Whilst the concept is abstract, ignoring uncertainty is far worse and may result in concluding that things work when they do not and vice versa; even if this is just for the sample in question.”
>
> Styles (2014, p.21)

It is valuable to unpack this position as we are no closer to understanding what it means to ‘quantify chance’ in the context of data which are not randomly sampled:

A common justification for applying the derivatives of probability theory in this area is to make a distinction between design-based and model-based inference (Reynolds et al., 2012, Plewis and Fielding, 2003, Goldstein and Noden, 2004, Snijders and Bosker, 2011). Model-based inference is an 'attempt to formulate and evaluate the structure of relationships between response variance of interest and relevant explanatory variables' (Plewis and Fielding, 2003, p.411) rather than generalise from a (conventional) sample to a population. A relatively clear exposition of the difference between model-based and design-based inference is given in Snijders and Bosker (2011) who explicitly recognise the lack of awareness and confusion surrounding this distinction:

> "It may be noted that in many introductory textbooks there is confusion because statistical modelling is often argued by underlining the importance of random sampling combined with making assumptions about independent and normally distributed residuals, thereby confounding the distinction between design-based and model-based inference."
>
> (Snijders and Bosker, 2011, p.218)

Let us consider the model-based perspective: Goldstein and Noden (2004) observe that differences will arise between schools even if pupils are allocated to them at random and even when the entire population is considered. Such randomness within the sample/population proves a problem for model-based inference because, on every comparison which could be made, there will be some difference, however small. Every variable considered would be included in the model if all non-zero differences are considered important. For instance, even with a non-random sample, there are a potentially enormous number of ways of splitting the sample into two even groups. How can we know differences pertaining to our chosen variable (gender, for example) are distinguishable from the myriad other dichotomous variables we could conceive? One approach to take is to calculate the expected differences which could be expected from a random split and use this as a benchmark to judge the statistical significance of an observed difference. Similarly, one can use a variable consisting of random values (with the same mean and standard deviation as variable of interest) as a point of comparison. Use of inferential statistics in this way separates the systematic 'signal' from the random background 'white noise' when making model-based inferences. This approach is analogous to the permutation test of random samples in data from a randomised control trial discussed in Gorard (2015) in the sense that, in both cases, the meaning of the statistical significance can be verified within the sample itself (although the meaning differs in other ways, see below). Statistical estimates have some meaning in this light as a point of comparison to identify differences larger that could be expected from such random sorting and comparison. In the same vein, Gorard et al. (2012) compared the number of schools identified as adding high/low value-added over a number of years to the number which would be expected if schools were sorted to either high or low VA at random. Note that this comparison is still meaningful even if, as was the case, no such randomisation had taken place; the aim was to reach model-based inferences about 'the probabilistic mechanism that could have generated the values of the dependent variable' (Snijders and Bosker, 2011, p.217) rather than generalise to a wider population.

This conception is of a specific meaning of statistical tests and confidence intervals. This meaning is limited and whether this is of practical value is considered in a moment. Presently, note that there is a difference between comparing results to a hypothetical random mechanism and claiming that the results *arise* from a random mechanism. It is important to distinguish these: the former is verifiable, as described above; the latter, as White (2014) notes, is a philosophical, rather than statistical appeal. We can, for instance, certainly compare a coin toss with a hypothetical fair coin using probability theory (i.e. a Bernoulli distribution with p=0.5) and use this as a benchmark for conclusions about coin bias. The claim that results of the actual coin are (or are like) instances of sampling from a super-population with a Bernoulli distribution but with an unknown, but discoverable, probability is considerably more difficult to justify. Consider this in relation to schools: We may compare the differences in school value-added in light of variance expected by a random sorting of pupils to schools (see Goldstein and Noden, 2004, cited in Reynolds (2010) as a response to Gorard) but it is another matter to say that pupil scores *are* instances of random sampling.

There may indeed be grounds for positing some random component to the underlying mechanism generating value-added scores. For instance, one finding in support of this is the empirical evidence suggesting that the volatility of value-added scores is inversely proportional to the size of the school, with small cohorts having greater spread of results (Gorard et al., 2012). It is difficult to explain this relationship without ascribing some portion of the difference to random sources of variation (which cancels out as samples increase). Note, however, that this is different to saying that *all* variation around the school VA mean is an example of so-called 'sampling error'. Drawing inferences about what is causing observed differences brings us back to the difficulties relating to inference given a potentially changing and complex school effect (see Section 4.3.2). Unlike with the example of a coin, it is questionable whether a school value-added score can be considered a reflection of a single, stable mean. How can we know what is generating the differences in pupil value-added performance which are observed other than by *a priori* assumption about what is error or bias and what is effect? *Ad absurdum*, we might well say that there is no random 'sampling' error at all, only a school effect that is so heterogonous that there is a different *effect* on all pupils in the school. On what basis can we say that some differences are effect, some are error (see Section 4.3.2)?

It is important not to conflate the limited use of statistics for model-based (as described above) with claims about the possibility of wider generalisations. Plewis and Fielding (2003), who are cited in Reynolds et al. (2012), describe the rationale for fitting statistical models which reflect real-world structures and then thinking in terms of super-populations as follows:

> "We are interested in rather more general conclusions than just about this set of pupils but also about pupils (and teachers) in similar contexts (possibly in previous and subsequent years who can reasonably assumed to be rather similar to the year in question). A well-formulated model enables us to adapt to these more general questions."
>
> (Plewis and Fielding, 2003, pp.410-411)

While Plewis and Fielding (2003, p.411) note that the connection between different types of statistical inference and scientific inference is 'profound', they do not address the issue. Yet it should not be treated as a philosophical aside. The use of inferential statistics to make wider generalisations commits the analyst to empirical assumptions about how the data were generated which generally have little or no basis yet are strictly necessary for statistical significance to have any meaning or value (Berk and Freedman, 2003). The distinction between limited model-based inferences and wider generalisations is most clearly explained by analogy: imagine a number of coin tosses designed to test whether a coin is fair. One may imagine a hypothetical super-population (or underlying generation mechanism) which has generated the results and suppose that there is a mean value which exists or is meaningful in some sense (in the case of a fair coin this should be 0.5 where heads is scored as 1, tails as 0). With these conceptual apparatus in place, it is possible to draw on probability theory to reach inferences about the underlying generation mechanism of the (specific) coin toss and whether it is fair or biased. Suppose that one wishes, however, to generalise to other coins of the same denomination, to coins more generally, or maybe to small disc-shaped objects. This cannot be done through inferential statistics and requires an inductive scientific argument. Yet, Plewis and Fielding (2003) seem to be arguing that it can. They make two points: first, that a specific group of pupils should be considered a sample of the actual (or even hypothetical) pupils more generally, which seems reasonable; and, second, that this is a *random* sample of these, which does not. Without this latter assumption, we return to the point that statistical inference cannot be used to make this generalisation: scientific (inductive) inference is required.

Similarly, this problem is apparent in Snijders and Bosker (2011) who explain that inferences to an 'infinite population' would require two steps of inference: a design-based approach to generalising from a random sample to finite population and a model-based application to generalise from the finite to the infinite population. The limited conceptions of model-based inference above described the use of statistical tools in terms of generalising about the finite population itself (the 'underlying' mechanism); Snijders and Bosker (2011), however, are claiming that *wider* generalisations can be made using inferential statistics. Again, this is to assume that the finite population is a random sample of the infinite population. As much as statisticians may wish this to be true, the assumption has little firm basis (Berk and Freedman, 2003). As Gorard (2015) notes, statisticians have gone to great lengths to justify the use of statistical tools in the specific context of value-added (Gorard cites Camilli, 1996, which contains a revealing and fascinating discussion of this issue).

In conclusion, this discussion has argued that the results of statistical tests are not meaningless outside the context of a non-random sample (or population) but, rather, can only have a very specific and limited meaning. Broader conceptions of model-based inference do not appear to be justified. Statistical inference cannot, therefore, be used to generalise to any larger (real) population in the absence of a design-based application based on a random sample. It might be that such generalisations are possible in certain circumstances (given further philosophical assumptions, see above) but they are not – it is argued here – defensible in the context of value-added. The only meaning of a statistically significant result in this context is, as per the definition, a difference which is larger than that which would be expected from random chance. If randomisation has not taken place in the design, random chance is hypothetical and profoundly unrealistic in most cases. Within the framework of model-based inference we might choose to interpret this difference as systematic; although, when a difference is not statistically significant, we cannot say it is not 'real' or that it is *because* of chance. It is either effect, error, bias, or most likely some combination of these. The difference might entirely be comprised of effect and if this is small relative to other sources of variance, a type 2 error will be made. In other words, statistical tests in the context of model-based inference therefore provide a benchmark rather than an explanation. Compare this to the context of a randomised control trial. Here statistical tests are used to estimate whether a difference is greater than would be expected merely from the randomisation to groups. If a non-significant difference is discovered this is an *explanation* for the difference relating to the random sampling error known to have been introduced by the design rather than a benchmark used to draw a line between 'random' and systematic difference.

The Practical Value of Significance Testing for Value-Added

This final sub-section on the issue of uncertainty and statistical testing brings the position developed above to the more practical context in which the results of statistical tests are interpreted and used. It is argued that, while statistical tests may be of some limited value during the specification of econometric models, there are numerous serious problems with their practical application to any other context.

A key practical limitation of the use of significance testing is that it 'emphasises random errors at the expense of explanations' (Coe, 1998, p.2). As noted, within model-based inference, lack of statistical significance is better viewed as a benchmark than an explanation. Even if a model-based estimator is found to be statistically significant, this is no guarantee that it reflects a larger finite sample (if applicable) or even that it is an unbiased estimator in its own right (Snijders and Bosker, 2011). As Gorard (2015, p.92) notes, 'even if they worked as intended, [confidence intervals] and p-values could not address measurement error, missing data or bias.' When using population data, p-values and confidence intervals are simply not the right tools to examine the typical threats to validity which should concern us (Gibbs et al., 2015). It is impossible for confidence intervals to account for many if not all of the numerous and serious threats to validity considered in Section 4.2, above. Yet, these are vital considerations for how 'confident' one should be in the results.

In the context of large datasets such as the National Pupil Database (NPD), statistical significance is an almost entirely irrelevant consideration for analysts. With a cohort size of over half a million pupils, even the smallest differences are statistically significant. In relation to smaller samples, it is not clear that statistical significance is any more valuable. Statistical significance does not equate to substantive (or 'scientific') significance nor bring the practical, human costs of type 1 or type 2 errors into the analysis (Wainer and Robinson, 2003, p.27). All of this suggests that there is no substitute for judgement (Gorard, 2006b). Yet, compare this apparently uncontroversial position with that of Muijs et al. (2011) who state that 'there are weaknesses in the measurement and conceptualisation of many studies, but this is primarily a technical problem with a technical solution.' In their view, the best response to 'uncertainty' is the 'appropriate modelling of error terms' using techniques such as item response theory and other latent variable models. This position contrasts sharply with that of Gorard and that reached here (see Chapter 7 and 8).

The extent to which any of the difficulties discussed in this section are understood by the research community might be questioned. By claiming that such difficulties have been 'extensively refuted' and quoting an example of the (less problematic) conception of model-based inference, Reynolds et al. (2012, p.10) knowingly or unknowingly divert attention from some serious fundamental problems with the use of significance testing. It is so conventional to discuss statistical significance in the context of social science, it is very difficult to know what the speaker has in mind with regards to the assumptions being made. While it might be the case that some producers of statistical tests do in fact have a sophisticated understanding of the various underlying assumptions in mind when interpreting them, the consumers outside of statisticians almost certainly do not. Herein lies the problem with the view (e.g. Leckie and Goldstein, 2011, Reynolds et al., 2012) that confidence intervals and other derivatives of inferential statistics are suitable for conveying uncertainty. As will be explored further in the discussion chapter (Chapter 7), maybe one of the biggest problems with confidence intervals is that, at best, ‘confidence’ is a misnomer which invites misinterpretation on the part of users, especially those with no statistical training. Confidence intervals may even be counter-productive, encouraging misplaced confidence in measures by distracting from more substantive threats to validity.

4.4 Alternative Measures of School Effectiveness

4.4.1 Introduction to Alternative Measures of School Effectiveness

One of the studies included in the empirical sections of this thesis tests inter-method reliability of value-added by comparing school value-added estimates with scores produced using a regression discontinuity design. Unlike the variations of value-added models considered earlier in this chapter, regression discontinuity (RD) design is a fundamentally different approach to measuring school effectiveness. This section gives details of this design and research examining and utilising it. It also gives details of another type of design known as a ‘seasonal’ design. The seasonal design and the application of the regression discontinuity design to estimate school effects are relatively recent advances in educational effectiveness research (Sammons and Luyten, 2009).

4.4.2 Alternative Measures of School Effectiveness

The ideal design to estimate differences between school effectiveness would be a randomised control trial where pupils were randomly allocated to schools. Of course, this type of experiment would be very dubious in ethical terms and such an experiment has not been tried. Nonetheless, there are some examples of randomised allocation to classes (Nye et al., 2004, see Section 4.2.5). In contrast, the value-added method uses model-based comparisons of statistically-similar pupils to separate value-added from non-school factor differences and, until recently, was the only feasible approach to estimating school effectiveness. Recent advances in educational effectiveness research, however, have identified two alternatives which can make design-based comparisons of pupils. These are the seasonal design and the regression discontinuity design (Sammons and Luyten, 2009).

Regression discontinuity designs match whole cohorts to consecutive cohorts in the same school. As the regression discontinuity design is used within one of the empirical studies in this thesis, there are several sections dedicated to it below. Seasonal measures look at rates of progress for the same pupil across time, using the difference in rates of learning for a holiday period compared with rates of learning during term time as a measure of the school effect (Verachtert et al., 2009, von Hippel, 2009). The idea is that all the effects of the non-school factors are operating both during the summer break and during the school term whereas the school factors are only operating during the latter. The difference in the rate of progress can, therefore, be attributed to the effect of the school.

The key reason that the RD design was preferred to the seasonal design in this study for an alternative to which value-added scores can be compared was data availability. Seasonal measures involve obtaining estimates of performance from at least three time periods: before the summer break, after the summer break and at a later stage in the school term.

4.4.3 Measuring School Effectiveness using a Regression Discontinuity Design

This section gives technical details and reviews all research on the use of the RD design to estimate school effects. The RD design is not formally specified here but the specific model used is included in the Appendix D, as referred to in the results chapter.

Introduction to the Regression Discontinuity Design

The use of the RD design in social research dates back to the mid twentieth century (Shadish et al., 2002) and it has become increasingly used in educational research (e.g. Allen, 2012, Vardardottir, 2013). The use of the RD specifically for the estimation of school effects is far less common. While there are early examples (Cahan and Davis, 1987, Cahan and Cohen, 1989) of RD-based school effects estimation, the practice has only recently come to more general awareness amongst educational effectiveness researchers following research by Luyten (2006) which illustrated and assessed RD-based school effectiveness estimation and subsequent work which extended, tested and applied the method (Luyten et al., 2008, Kyriakides and Luyten, 2009, Luyten et al., 2009). Building on these promising results, researchers are beginning to make use of RD in educational effectiveness studies (Heck and Mahoe, 2010) and it is being recognised as a 'fruitful' methodological development in school effectiveness measurement (Reynolds et al., 2014, p.204).

RD-based measures estimate treatment effects by considering the outcomes either side of a known cut-off point for the treatment in question. A sudden break in an otherwise continuous regression line yields strong evidence regarding a programme's effectiveness and the magnitude of the discontinuity can be used as an estimate of the programme's effect (Trochim, 1984, Shadish et al., 2002, Bloom, 2009). This design can be applied to the estimation of school effectiveness: Many school systems admit young children to the first year of schooling on the basis of their age relative to a given cut-off date. In England, those born on the 31st August will have received a whole year extra of schooling than pupils of almost the same age born a day later on the 1st September. Within cohorts, age has a clear, positive association with academic performance and there is a strong tendency for older pupils within a particular school year to out-perform younger members of the same year group (Crawford et al., 2010). These organisational features raise an opportunity for school effectiveness measurement as, using RD, this school entry cut-off can be used to separate the effects of age and schooling and, thereby, estimate the (absolute) effectiveness of schools in improving the measured outcome (Luyten, 2006, Luyten et al., 2009).

Threats to Validity when using a RDD

Basic RD designs only need a measure of age and test scores for two consecutive year groups. Assuming a valid measure of the outcome is obtained, the key threat to validity of a RD design is non-adherence to the cut-off (Shadish et al., 2002) (i.e. pupils in a different year to that predicted by their chronological age). It is common practice in some school systems to 'hold-back' lower-attaining pupils by a year or to 'promote' higher attaining pupils to a higher year. The extent of these practices differs substantially by country (Luyten and Veldkamp, 2011). Less than 5% non-adherence is often considered a level which will give reliable estimates (Trochim, 1984). A study by Cliffordson (2010, p.50) found the effect of a non-adherence rate of 3.5% on the estimates was 'generally relatively small'. English rates of non-adherence are generally found to be relatively low at around 1% to 2% (Luyten et al., 2008, Luyten and Veldkamp, 2011, Luyten et al., 2009).

The RD design uses the lower of two consecutive cohorts in a school as the baseline against which the absolute effect of schooling can be estimated. This raises a second problem: Cohort characteristics in a school fluctuate from year-to-year and this may lead to unreliability in estimates of the effect of an additional year of schooling, a problem also faced by VA models (see Teddlie and Reynolds, 2000, p.72).

One final difficulty is a relative age effect within a school year. It may be the case that there are relative age effects, where being the oldest or youngest in a year group has an influence over and above this general function describing the link between performance and age. Previous research in this area, however, has concluded that the absolute age effect is approximately linear and that the pupil's age when taking the test rather than a relative age effect is the overriding factor explaining the link between age and examination performance within a given cohort (Crawford et al., 2010, Crawford et al., 2013, Cliffordson, 2010).

Practical Use of Regression Discontinuity Designs

Use of the RD design to estimate school effectiveness is currently quite rare. Nevertheless, the existing evidence gives a positive picture of the design and its potential as well as identifying issues which must be considered. Early pieces of research showed clear absolute effects of an extra year of schooling using a RD design (Cahan and Davis, 1987, Cahan and Cohen, 1989). More recently, Luyten (2006) applied RD to the Third International Mathematics and Science Study (TIMSS), for eight countries whose adherence to the age-grade cut-off date was high and calculated estimates of grade effects (of the 4th grade). The results were in line with previous studies, showing clear grade effects, relatively high grade-age effect ratios and sizable differences between the performances of different schools. The possibility of including interaction effects in the model and thereby analysing whether other factors are associated with the 'added year effect' was also demonstrated.

Luyten followed this study with several other studies over the coming years exploring the possibilities of the regression discontinuity design. As well as yielding research findings in their own right, these studies demonstrated, tested and developed the possibilities of the design. One of these studies made use of PISA 2000 data and found only a small effect on reading performance in year 10-11 English secondary education and no grade effect for reading engagement or reading activities (Luyten et al., 2008). This grade effect on reading performance was not found to vary significantly between schools. In a recent follow-up study to Luyten et al. (2008), Benton (2014) points out various difficulties in the use of the OECD PISA data and found that the small effect on reading performance disappears when these difficulties are taken into account. Benton (2014, p.10) puts this non-effect down to the lack of alignment between the PISA tests and the English secondary school curriculum.

Another example is Luyten and Veldkamp (2011) who apply RD to estimate the effect of 1 year of schooling on attitudes and achievement in mathematics and science using TIMSS-95 data, a large cross-national survey. The method is extended to include a 'correction factor that expresses the effect of the *unmeasured* variables determining assignment to grades' (Luyten and Veldkamp, 2011, p.267). They found the added-year effect of schooling for mathematics and sciences to be positive with some variation across countries. With regard to mathematics attitudes, the added-year effect was found to be negative in all cases but quite small, with RD (with the correction factor) typically explaining less than 1% of the variance. The grade effect on attitudes to science was found to be negligible, with contradictory signs and little variance explained.

Other recent studies have demonstrated the possibility of extending RD to encompass multiple-cut off points (i.e. a series of added-year effects across a number of consecutive school years) (Kyriakides and Luyten, 2009). This was achieved by Kyriakides and Luyten (2009) whose results, using both curricular and non-curricular outcomes for 577 students in 6 schools across 6 grades of secondary education, provide further evidence for the value of the RD design for the estimation of school effects. With a sample containing only 6 schools, significant differences between schools were not found in the schools' relative effects (Kyriakides and Luyten, 2009). Also, 52 pupils of the sample of 629 (8%) were dropped from the analysis due to being allocated to year groups without strict adherence to the cut-off (i.e. were retained or promoted to another year group). Despite these difficulties, Kyriakides and Luyten (2009) provide another clear example of the successful use of the RD design and, moreover, demonstrate that it can be extended to model performance across numerous consecutive year groups rather than just two.

Of particular relevance to this study is Luyten et al. (2009), the only study known to this author which has, as is the intention here, compared cross-sectional and longitudinal estimates of school effectiveness. Luyten et al. (2009) drew on data from the baseline assessment used within 'Performance Indicators in Primary Schools' (PIPS) project (see http://www.cem.org/primary, (Tymms and Albone, 2002, Tymms, 1999)), estimating the added-year and school effects for 4- and 5-year-old pupils. The PIPS data used contained 'less than 1.5% of the pupils were in the "wrong" grade given their date of birth' (Luyten et al., 2009, p.146) and is therefore excellent for the calculation of RD-based estimates. Luyten et al. (2009) applied the RD design to both cross-sectional data and longitudinal data to test whether the estimates are consistent (nb. in Chapter 6, Section 6.2, these are compared alongside the VA design and are referred to as RD and longitudinal RD (LRD) measures, respectively).

Luyten et al.'s findings indicate that the overall effect (for all schools) of an additional year of schooling (for all schools) is 'very similar' in both the cross-sectional and longitudinal dataset for all three outcome areas and this is the case across models accounting for linear and quadratic effects of age for which there were 'hardly any difference' between the cross-sectional and longitudinal data (Luyten et al., 2009, pp.152, 156). In terms differences in the grade effect for individual schools, variance in the cross-sectional (RD) data was consistently higher than estimates based on longitudinal data. Correlations of the school effectiveness estimates between cross-sectional RD and longitudinal RD estimates for individual school effects were .78, .71 and .52 for reading, mathematics and phonics respectively. The latter appearing to exhibit ceiling effects in the assessment across the two years. These results suggest that school-level estimates produced by each method are fairly consistent yet have some level of disagreement. As Luyten et al. (2009) compared the estimates for the effect of only one year group for 18 schools, there is great value in replicating these results in a larger dataset including more schools and a greater range of ages.

4.4.4 Comparing Regression Discontinuity and Value-Added Designs

This section describes the major differences between value-added and regression discontinuity designs. This provides the basis for meaningful comparison in the later empirical section.

Absolute vs. Relative Effect

A major difference between VA and RD is that the former estimates the relative effectiveness of schools, whereas the latter estimates the absolute effect. Other things being equal, where a school is highly effective and its pupils made higher rates of progress, this should be reflected in both measures. A problem with using a RD design for the comparison of schools is that it may prove to be inequitable given that the absolute effect of schooling or the age effect may systematically vary with prior achievement or other factors. Where there are strong cohort-level differences which influence rates of progress, the absolute measure obtained through RD will differ from the relative measure produced by the VA design which takes such differences into account.

'Like-with-like' Comparison

The biggest design difference between the VA and RD is the choice of comparators. RD assumes that two consecutive cohorts are from a single population, with the only systematic differences accounted for being the pupil age and extra year of schooling received by the upper cohort. Particularly when effectiveness estimates for individual schools are sought, RD is at risk of differences between cohorts distorting the measured absolute effects. A lower cohort with relatively poor performance, for example, would exaggerate the absolute school effect.

In contrast, the value-added method makes the assumption that a statistically constructed average pupil for given measured characteristics is the appropriate comparator. In other words, pupil performances are compared to the performance of other pupils with the same statistical information as them. This raises two key threats to the validity of value-added measures: first, the number of pupils studied may not be sufficient for the effect of unmeasured or imperfectly measured variables to 'even out' and, second, that variables which have appreciable school-level effects may be unmeasured or even unmeasurable. These omitted variables could result in unobserved biases in the estimates of (relative) school effectiveness even where group sizes are large enough to eliminate other extraneous 'noise'.

At present, it is unclear which of these assumptions is more problematic. It is important to note that each design is strong where the other is weak: if the cohort assumption from the RD design holds, the major advantage is that the cohorts will be equivalent on both measured and unmeasured variables. If unobserved variables are a major problem for validity, the quasi-experimental logic of RD may yield more robust estimates. On the other hand, while VA is vulnerable to unobserved school-level differences, the VA measure is likely to be less influenced by differences in the characteristics of cohorts given that it draws on the strength of the whole sample (across all schools) to generate estimates of expected performance.

Underlying Measures and Common Problems

Finally, there are many problems common to both measurement designs which could render both measures invalid even if they prove to be in agreement. These problems include the validity of the underlying measure of performance used and problems of generalisation such as the likelihood of differential school effectiveness across ability levels, groups and various outcomes (Sammons, 1996, Thomas, 2001). These problems can raise slightly different problems for each design, such as the added requirement for the outcome measure used for both year groups in the RD design is equally applicable for both year groups in a way which appropriately measures progress across the two years (see Cahan and Elbaz, 2000). This study focuses on differences between the robustness and design of VA and RD measures rather than these common problems, although they are certainly important considerations.

5. Methods

5.1 Chapter Introduction

5.1.1 Introduction to the Four Empirical Studies

Chapter 4 explored the evidence pertaining to numerous issues relating to the validity of value-added. In the absence of clear experimental evidence that provides a definitive single test of validity, researchers have pieced together evidence relating to bias, stability and consistency in order to form a view on the validity of value-added. The empirical portion of this thesis reflects this by presenting results from a total of four main studies which examine the following issues: first, *bias and error*, which examines sources and seriousness of bias and error, primarily using the official English value-added measure; second, *inter-method reliability*, where estimates produced using value-added are compared with estimates produced using the quasi-experimental regression discontinuity design; third, *stability over time*, where the stability of English value-added measures and the stability of the performance of cohorts over a number of years is examined; and, fourth, *cohort consistency*, where the consistency of estimates for different cohorts within a school at a single point in time is examined.

These four studies are presented across the methods, results and discussion chapters rather than as four separate studies (i.e. each with its own methods, results and discussion sections). Within this, the methods chapter is designed to provide introductory and explanatory information for each study and detail the specific sample used. Detailed description of the analysis is left until the results chapter. Model specifications for the statistical models used and descriptive statistics are placed in an appendix unless these are essential to the analysis. Organising the four studies across chapters in this way avoids duplication of information, about the data sources for example; allows the results to be considered collectively, preventing the discussion becoming fragmented or repetitive; and enables the results chapter to concentrate on posing, explaining and then answering each research question in turn. The main intention is to create a results chapter which is as clear and self-supporting as possible. Readers wanting to scrutinise the samples and statistical details of the analysis are able to consult this methods chapter and the appendices for more detailed information.

5.1.2 Chapter Organisation

The methods chapter begins by introducing the research questions and the overall design and approach of the four studies (Section 5.2). The core research question (Section 5.2.1) is broken down into several primary research questions which are sorted into the four studies (Section 5.2.2). After this, there is a sub-section which explains selected aspects of the overall analytical approach taken in the four studies (Sections 5.2.3).

The following section of the methods chapter (Section 5.3) gives details of the three key data sources which are used across the four studies. Details are given about general analytical decisions and aspects of how the data in the data sources are cleaned and organised.

The methods chapter closes with four sections (Sections 5.4 to 5.7) which discuss the general approach, sample and methods of each of the four studies in turn. This information is designed as an overview only. More detailed information on the analyses is given in the relevant sections in the results chapter alongside presentation of the results. Model specifications and further sample details are placed in the appendices unless these are central to the findings.

5.2 Research Questions and Design

5.2.1 Core Research Question

The core research question is as follows:

Are school value-added measures valid measures of school effectiveness?

Where 'school effectiveness' is operationally defined by the value-added method as the relative effect of schools on measured outcomes. This is examined here in terms of the estimated size of the residual school-level differences. Note that the term *school effect* is frequently used elsewhere (e.g. Luyten, 2003) to describe the overall variance attributable to schools relative to all other non-school factors (also see Willms, 2003). In this alternative use of the term, subordinate levels such as the classroom may or may not be encompassed within the term, depending on the purpose. Also note that school value-added measures are viewed here in terms of being an estimate of causal effects of schools rather than merely as school-level unexplained differences in performance. It is the interpretation of school value-added scores as a causal estimate of school effects which raises the question of validity. In sum, a school value-added measure is valid to the degree that it captures the relative causal effect of the school and therefore can be used to draw conclusions about a school's performance relative to other schools.

5.2.2 Primary Research Questions

The core research question is broken down into the following primary research questions:

Table 5.2.2a - Primary Research Questions

	Study 1 – Biases and Error
RQ 1.1	*Are there observable biases in the current English value-added measure?*
RQ 1.2	*What is the level of missing data in the National Pupil Database?*
RQ 1.3	*What is the influence of measurement error on value-added scores?*
	Study 2 – Inter-Method Reliability
RQ 2.1	*How similar are estimates of effectiveness produced by value-added (VA), cross-sectional regression discontinuity (RD) and longitudinal regression discontinuity (LRD) designs?*
	Study 3 – Stability over Time
RQ 3.1	*How stable is the current English value-added measure across several years?*
RQ 3.2	*Is the rate of stability in value-added scores associated with school performance?*
RQ 3.3	*How stable is the contextual value-added performance of a given cohort over time?*
	Study 4 – Consistency across and within Cohorts
RQ 4.1	*How consistent are value-added estimates of performance across cohorts from within a single school in a single year?*
RQ 4.2	*How consistent is performance within cohorts?*
RQ 4.3	*Does within-cohort consistency vary by mean school performance?*

These questions form the primary questions for each study. Within each study there are also supplementary questions which provide key supporting information as well as more fine-grained questions which address specific issues or aspects of the data which are used.

5.2.3 Analytical Approach

At the start of this chapter (Section 5.1.1), it was asserted that the best practicable way to examine the validity of value-added is to draw on several types of indirect evidence. This section provides further justification and explanation for this assertion and so links the core research question to the primary research questions for which empirical evidence is provided. Two issues are discussed: first, the combination of various sources to reach conclusions about validity. Second, the interpretation of inconclusive, indirect or ambiguous evidence. These issues underpin the approach taken in the four empirical studies as well as how the findings are dealt with in the discussion section to reach overall conclusions.

Combining Validity Evidence

Each study and analysis follows what is intended to be a self-contained, self-explanatory argument pertaining to a property of value-added evidence in the given context. To a large extent, therefore, the four studies presented here and the analyses and sub-analyses within them could be considered a number of discrete issues which happen to be grouped by a common topic (properties of value-added). Despite this, the results are understood to form a larger picture of value-added measures and their methodology, rather than merely being a series of independent findings so it is useful to briefly comment on how these analyses and findings can be brought together to reach more general conclusions.

The basic approach to address all research questions has been to aim for a concrete and objective answer within a single analysis using the highest quality data and most robust research design available. Often, however, this is not sufficient to provide a definitive answer to the specific research question or the core question of validity. As this is the case, multiple sources of evidence are brought together with a view to building coherence between a body of empirical findings and a theoretical understanding (Kvanvig, 2008). By looking across studies rather than within them it is possible, for example, to examine whether instability over time is likely to have been caused by inconsistency in the performance of cohorts, changes in school performance, measurement error or changes in the value-added model. While the evidence of any single analysis may fall short of being conclusive, often results from other studies will be able to narrow down the possible explanations for the properties of value-added observed (and whether these are likely to be error or effect). In order to do this, the research questions have been designed to complement each other, addressing the questions raised by other analyses. Most analyses within the studies were conducted around the same time, only being separated into the four main studies afterwards for purposes of clear presentation. Note that many of the research questions relate to properties of school value-added rather than to directly inform validity (see Chapter 4, Section 4.2.1 for discussion of direct and indirect validity evidence).

Bringing together all findings across all studies and discussing the evidence in terms of validity (rather than properties) is a task undertaken in the discussion and conclusion chapters, where the results are considered collectively in order to reach conclusions about the validity of value-added which can be reconciled to the greatest possible extent with the findings of all four studies. The conclusions could be described as being underpinned by a coherentist approach to justification (Kvanvig, 2008, Little, 2013), where explanations for variance are considered alongside the collective results to develop an evaluation of validity which is most consistent with all of the available evidence.

Dealing with Inconclusive, Indirect or Ambiguous Evidence

While the approach to justification described above will generate some general conclusions about the validity of value-added, it is unlikely that highly specific conclusions are possible with the evidence available. As well as the interpretative difficulties discussed in Chapter 4, there are difficulties generalising from the specific data and value-added models used to value-added evidence more generally (see Chapter 2). Rather than stop at general conclusions based on findings which are inescapably ambiguous to some degree, the discussion chapter shifts the consideration of the results to the various areas of use across research policy and practice which have been reviewed (Chapter 3).

As an example, consider the issue of stability. A specific conclusion on reliability cannot be reached as there are (at least) two unknowns: the reliability of value-added and the changeability of school performance. While it might be possible, drawing on various other results, to get an indication of the extent to which these factors are driving the observed level of stability, a precise answer is not possible. But rather than leave this problem open, it is pursued to the practical context of decisions where, even if conclusions about the validity of value-added evidence cannot be reached, conclusions about the validity of arguments for the use and interpretation of this evidence can (Kane, 2013). This approach bears some similarity with that taken in research of Leckie and Goldstein (2009) and Allen and Burgess (2013).

Both of these studies examine performance data in relation to a specific use (parental choice of school). One can construct the following pragmatic argument: for a measure to be of value for purposes of school choice, it must be able to provide a meaningful estimate of future school performance. In this practical context, if the measure is not sufficiently stable to provide a meaningful measure for purposes of school choice, it is arguably of lesser importance whether this is due to the measure validity or phenomenon stability. It can simply be concluded that the uses and interpretations which are demanded by the particular context are or are not valid given the properties of the measure. In the case of the study by Leckie and Goldstein (2009), for example, the measure is found to be insufficiently stable over time and have too much statistical uncertainty for it to be a meaningful measure for purposes of parental choice. Allen and Burgess (2013) estimate the extent to which several performance measures improve choice of school compared to a random choice and concludes that the measures (including the value-added measure) improve on the random choice. Both of these cases are able to reach a concrete conclusion in relation to a practical application of the measure without needing to address more fundamental questions about the underlying source of the observed variance.

The value of this pragmatic position is that it allows more fine-grained validity conclusions, despite ambiguous and inconclusive results. This approach can be taken with all applications of value-added where the general idea is to look at the practical application of value-added evidence and consider the requirements for beneficial use of the measure under these circumstances.

5.3 Key Data Sources

5.3.1 Performance Tables Data

In England, details of school performance and characteristics for all state-funded schools are published annually on the 'performance tables' on the Department for Education (DfE) website (DfE, 2015). Approximately 7% of pupils nationally attend private schools and so data are generally unavailable for these pupils. Where data are available for privately educated pupils, measures are not always comparable due to differences in qualifications which are taken. School performance data go back to 1994 and full school-level datasets are freely and readily available from 2005 onwards (DfE, 2015). This research uses school-level data from 2011-2014, where 2011 was the first year in which the current VA measure was used and 2014 was most recent year for which performance data were available during analysis. It is possible to match schools across years of data using unique school reference numbers recorded with each record in the data. Even when schools change school name or type, it is generally possible to match schools using local authority codes and establishment codes.

The data provided in the performance tables is used in two of the studies in this thesis. First, the study of reliability of value-added scores across years and, second, the study into observable biases and error in the official value-added scores. Further details of the exact measures used will be given later in this chapter and the relevant results sections.

5.3.2 National Pupil Database Data

As well as the school-level data collected in the performance tables (above), the DfE collects pupil-level data which combines data on pupil and school characteristics with examination results from key stage examination years. It is possible to apply for pupil-level data through application to the DfE who encourage use of the data for appropriate research or school improvement purposes. The National Pupil Database (NPD) is a very large dataset and has performance data going back to 1996; these performance data have been matched with the School Census data (formerly the Pupil-level Annual School Census, PLASC) since 2002 (DfE, 2015). A large number of variables are collected relating to achievement and to pupil and school characteristics. These fields change over the years with new policies and improvements in the data, but the general trend has been towards data of a greater quantity and quality. More information can be found on the NPD Wiki, which is maintained by NPD users in the research community and is updated regularly as the NPD changes (Allen, 2015a).

The main study (Study 1) in which the pupil-level NPD data are used concerns bias and error within the official value-added data used in policy and practice. Study 1 looks at some of the difficulties in constructing the measure and so uses the more fine-grained pupil-level data as well as the school-level data described in the last section. NPD data are also used in the study presented here regarding the consistency of value-added estimates across cohorts within schools in a given year (Study 4). As will be described in the next section, this research also obtained a dataset containing teacher-assessed performance data and numerous contextual variables. This dataset was collected as part of previous research funded by the DfE and the data were matched with NPD data. The dataset described in the next section, therefore, is a combination of data collected in the DfE study and NPD data.

5.3.3 'Making Good Progress' Data

The NPD (see above) contains performance data for pupils at National Curriculum Key Stage years (where Key Stages 1 to 5 correspond to ages 7, 11, 14, 16 and 18). Several studies in this thesis, however, required performance data for year groups who were between these years. Study 2 compares value-added estimates of school effectiveness with those produced using a regression discontinuity design, requiring performance data for consecutive year groups. Study 3 examines the stability of performance for a given cohort followed over time. Study 4 examines the consistency of performance between all year groups in a key stage at a single point in time. All of these require performance data for year groups outside of key stage years.

During initial searches for existing data to meet this need, a DfE study known as 'Making Good Progress' (MGP) was identified. This was obtained through an application process which allowed access to the MGP dataset for use in this thesis. The MGP study looked at how pupils progressed during Key Stages 2 and 3 (DfE, 2011), collecting teacher-assessed performance data for all cohorts within this age range for three study years. It is a very large dataset which is well-suited to the intended analyses. Summary details are given presently, before more study-specific details are given in later sections. Further details of the MGP dataset, including more details on the local educational authorities included, variables collected, the validity of the teacher-assessed data and methodology of the data collection, can be found in the DfE statistical report based upon it (DfE, 2011).

The MGP dataset is large with data for 148,135 pupils spanning 342 schools, 10 local authorities, 6 consecutive school year groups (UK years 3-9) across 3 years. There were 100,000 pupils in 2007/2008 with pupils being fairly evenly spread across years 3 to 9 (age 8 to 14). This overall number dropped to just over 70,000 by the third year, again spread fairly evenly across the age range. The MGP report compares the achieved sample across a range of pupil background variables with national data for these year groups, finding it to be 'broadly representative' of pupils in years 3 to 9 nationally (DfE, 2011, p.6).

The analyses of the MGP data make use of the teacher-assessed data based on National Curriculum (NC) levels framework and guidance. NC levels are designed to be a single scale tracking attainment from age 5 to age 14. It is questionable, however, whether the NC levels can be considered an interval scale (where the difference between level 3 and 4 can be assumed to be the same size as between 4 and 5, for example) and whether levelling is consistent across teachers across the full age range. There is also evidence to suggest that teacher-assessed levels can be unreliable in some circumstances, although it may be possible to improve this by way of moderation procedures and well-designed assessment criteria (Harlen, 2005). The evidence base on both reliability of teacher assessments and the effectiveness of moderation as a way of improving it is considerably lacking at present, however (Johnson, 2013). The MGP report discusses the quality of teacher assessments and includes an annex which compares the teacher assessed levels to those obtained in the key stage 2 and 3 examinations (DfE, 2011). This gives some indication of how consistent teacher-assessed and examination assessed grades were in this instance. Agreement between the teacher-assessed levels varied from 56% to 77% in KS2 writing, 36% to 95% for KS2 reading and 64% to 89% in KS2 mathematics. Some of the discrepancies will stem from differences in timing between the two measures, with teachers’ scores being lower than the examined results due to being recorded some time earlier (DfE, 2011).

Because of these differences, analyses of the MGP data use only the mathematics performance data which tended to have higher levels of agreement with the examination-assessed data (see DfE, 2011, pp.41, for a chart showing the correspondence between teacher-assessed and examination assessed KS2 mathematics). The report comments on the moderation activities which took place in schools during the study, noting that during the pilot study concerns were expressed about the initial quality of teacher assessment but that the quality improved as the ‘processes bedded’(DfE, 2011, p.7). The correspondence between the teacher-assessed levels and the examination levels tended to increase over the time period from 2008-2010. Although another factor is that the sample in the third year was reduced which is likely to reduce robustness, especially at secondary level (where the school numbers were lower). One other factor relating to the consistency of teacher- and examination-assessed attainment is that agreement was generally on the higher end of the above range for pupils of average or above average ability and lower for lower attaining pupils.

These problems of validity are inherent in educational and psychometric measurement in general and so concerns about the quality of the data used, while certainly noteworthy, are not held to be especially problematic for this study in particular. The quality of the MGP dataset is comparatively high, with the main difficulty being the fact that performance is teacher-assessed. Of course, it is not the case that examination-based measures of academic performance are entirely valid, especially when based on a single examination in a high-stakes context (Stobart, 2008). Moreover, teacher-assessments are used as part of the predominantly examination-based key stage results in the NPD (above) and are widely used in practice in schools (see Chapter 3 for details and Chapter 7 for further discussion of how this study limitation influences the implications of the results for policy and practice, respectively).

5.3.4 Approaches and Actions Common to all Studies

The majority of the information for each study is contained within the sections below and in the corresponding results sections. This section briefly outlines some steps taken with the data that are common to all analyses and so saves repetition. Several explanatory points are made to ensure key distinctions made within the various analyses are clear.

All analyses contained within the results chapter were conducted using either Stata (v13) or SPSS (v22), the final analysis has been completed almost exclusively using Stata. Syntax has been stored and can be produced on request. Next, all the analyses concern state-funded mainstream schools only. Special schools and independent schools and pupils can be identified in all of the available datasets and so are removed from all analysis. The reason for omitting the former is that special schools take pupils with highly individual and specialist needs. This makes use of value-added measures (which seek to produce like-for-like comparisons) highly questionable. Special schools, at best, could be considered a more challenging application of the methodology, whereas the intention is to consider its use in more favourable circumstances. Similarly, using private schools brings in issues surrounding the measure of performance as many private schools take alternative qualifications which are not counted within the official data and many pupils lack the prior attainment measures used in the calculations. For present purposes excluding private and special schools from the analysis allows the core issues to be addressed.

Another point relevant to several of the analyses is that there has been a great deal of reform of English schools; part of this is to change the legal and financial arrangements of schools so that the school is funded directly from central government (rather than local government) and the designation of such schools as ‘academy’ schools. One result of this is that many of the data sources used record the pre-academy and post-academy schools using different school reference IDs. Despite the nominal change to the school, it was thought appropriate to match schools operating on the same site using local authority and establishment codes rather than the school IDs as the latter remain unchanged through academisation.

In terms of presentation of the results, the approach has not been to put model output in the main text unless it is the object of discussion. Output from results which is not discussed directly but underpins key results is included in the appendices.

5.4 Study 1 - Bias and Error

5.4.1 Overview and Research Questions

This first study concerns bias and error within value-added measures. This is an issue which is somewhat specific to particular datasets and value-added models. This study primarily concerns the current English VA measure and former CVA and VA measures. The data come from the National Pupil Database. Analysis of the English (C)VA measures and the NPD has clear policy relevance. The method used for the forthcoming ‘Progress 8’ measure is likely to be changed on the basis of the recommendations in Burgess and Thomson (2013a). Despite this, results from the current model are thought to be highly generalisable to the Progress 8 measure given Burgess and Thomson’s results showing the high levels of similarity between the estimates from the models studied: the more complex models were found to ‘add little or nothing’ to the more simple Progress 8 model recommended (Burgess and Thomson, 2013a, p.8). Also, although the analysis concerns the specific context of the NPD and focuses on specific model formulations (i.e. the official DfE value-added measures), the analyses are designed to address general issues which are common to value-added analysis relating to bias and error. There is wider relevance to the degree that the measures and data bear similarities with other systems or EER studies. For example, all attainment data are likely to have some degree of measurement error but there are likely to be large differences in the quality of the examinations used in different countries or studies. The NPD is a relatively high-quality dataset which is practicable to maintain on a national level with data for the entire population of English school pupils (see Section 5.3.2 for further details). As a result, this is taken as a relatively good, but not ideal, basis for value-added estimates and so is a fair and realistic dataset on which to base analysis intended to examine school value-added measures more generally.

Research Questions

The primary research questions given in the general section above (*RQ* 1.1, *RQ* 1.2 and *RQ* 1.3) are supplemented here by sub-questions and supplementary questions.

Table 5.4.1a - Primary and Secondary Research Questions for Study 1

RQ 1.1 ***Are there observable biases in the current English value-added measure?***

RQ 1.1.1 *Are the current and former KS2-KS4 (Secondary) English value-added measures unbiased in relation to attainment?*

RQ 1.1.2 *Are the current and former KS1-KS2 (Upper Primary) English value-added measures unbiased in relation to attainment?*

RQ 1.1.3 *Is the current KS2-KS4 (Secondary) English value-added measure unbiased in relation to available theoretically important variables?*

RQ 1.1.4 *Is the current KS1-KS2 (Upper-Primary) English value-added measure unbiased in relation to available theoretically important variables?*

RQ 1.2 ***What is the level of missing data in the National Pupil Database?***

RQ 1.2.1 *What is the level of missing data in the variables used in the former KS2-KS4 English contextualised value-added measure and current value-added measure?*

RQ 1.2.2 *What is the level of missing data in the variables in the variables which would be required for a KS1-KS2 contextualised value-added measure?*

RQ 1.2.3 *Are ceiling effects, floor effects or scale discontinuities present in the main Key Stage (1-4) performance scores?*

RQ 1.3 ***What is the influence of measurement error on value-added scores?***

RQ 1.3.1 *To what extent does measurement error which is random at pupil-level influence school-level KS2-KS4 value-added estimates?*

RQ 1.3.2 To what extent does measurement error which is random at pupil-level influence school-level KS1-KS2 value-added estimates?

5.4.2 Sample

RQ 1.1 requires both school-level and pupil-level data for English primary and secondary schools. The first two sub-questions require data from former value-added models up to and including the most recent value-added model data available. Data for 2011-2014 are readily available to the public from the DfE's performance tables website (DfE, 2015). Data were obtained for primary and secondary level for these years. For 2007-2010, it was not possible to download the complete school-level dataset for all schools online. For these years KS4 school-level data were obtained by way of a NPD application. This NPD application also obtained pupil-level data going back to 2004. The lack of independence between VA and raw scores found in Gorard (2006c) was based on 2004 secondary-level data (see Chapter 4, Section 4.2.2, in the theoretical problems of model specification section) and so KS4 data were obtained back to this point. For 2004-2006, school-level data for these years were obtained by aggregating the 2004-2006 pupil-level data to school-level. In sum, for RQ1.1, school-level data for all maintained, mainstream schools in England were obtained for 2004 to 2014 at KS4 and for 2011-2014 at KS2.

The second two sub-questions within RQ1.1, RQ 1.2 and RQ 1.3 required pupil-level data. The pupil-level data obtained from a number of NPD data requests were used for this. At the point of final analysis the most up-to-date pupil-level data sets were the 2013 extract for KS4 and the 2012 extract for KS2 and KS1. This reflected the uses for which the data were to be put and the value, constraints and costs of further NPD applications for more up-to-date data. As noted in the results section, the quality of the data seems to be gradually improving. This is unlikely to affect any of the substantive results however so is not considered overly problematic.

5.5 Study 2 - Inter-Method Reliability

5.5.1 Overview and Research Questions

The second study involves comparison of a value-added measure with three variations on a regression discontinuity measure. The regression discontinuity design is discussed at length in Chapter 4, Section 4.4. These comparisons are a test of inter-method reliability or ‘alternate forms reliability’(Allen and Yen, 2001, p.78). Ostensibly, all designs are measuring the same phenomena (i.e. school effectiveness), although, as was discussed in Chapter 4 (Section 4.4.4), the different designs will produce different results and have different threats to validity. Ideally, both measures will be highly correlated, in which case this is strong evidence for the validity of each. In the more likely case that the measures differ to some degree, inferences will be drawn by examining the threats to validity for each measure. This is why several variations on the regression discontinuity design are valuable: this allows the sources of observed differences to be examined. All four measurement designs are described in the results chapter (Section 6.2.1) immediately prior to the main analyses.

The main objective of this study is to compare a VA design with different RD designs in order to test the validity of the VA design. As explained in Chapter 4, Section 4.4, however, the application of regression discontinuity to estimate school effects is a relatively new innovation in educational effectiveness research and, while early results are promising, development of RD designs to estimate school effects is ongoing. Before comparing all four measurement designs, therefore, there are some initial research questions which examine RD designs in their own right (RQ 2.1.1, RQ 2.1.2 and RQ 2.1.3). These initial research questions, below, are designed to test the validity of the RD design rather than the value-added measure. This serves two main purposes: first, to gather information on the validity of RD designs to enable it to be used as a test of the validity of VA. Second, RD is a possible alternative approach to measure school effectiveness for certain purposes. Understanding the strengths and limitations of RD to measure school effectiveness allows this thesis to consider the validity of VA alongside the possible alternatives (See Chapter 7 for further discussion).

Research Questions

Several research questions for this study follow on directly from those of Luyten et al. (2009, p.148) who compares cross-sectional and longitudinal RD designs and whose results the study seeks to replicate. In addition to these questions, further questions are added to extend the study to a greater range of concerns allowed by the more extensive data and because of the intention to critically compare RD and VA designs as well as the two RD designs. The more extensive data, for example, allows interaction variables to be included and estimates from across the age range studied (ages 7 – 14) to be compared (RQ 2.1.2). As noted above, as well as providing valuable information on the RD design, this feeds into the main intention of comparing VA and RD designs and being able to isolate which aspects of each design are driving any major differences.

Table 5.5.1a - Primary and Secondary Research Questions for Study 2

RQ 2.1 ***How similar are estimates of effectiveness produced by value-added (VA), cross-sectional regression discontinuity (RD) and longitudinal regression discontinuity (LRD) designs?***

RQ 2.1.1 *What is the effect of 1 extra year of schooling on achievement and what proportion of this is accounted for by schooling?*

RQ 2.1.2 *Do RD school effects differ according to ability or other contextual factors?*

RQ 2.1.3 *To what extent does the effect of 1 extra year of schooling vary between schools?*

RQ 2.1.4 *How similar are estimates of effectiveness produced by value-added (VA), cross-sectional regression discontinuity (RD) and longitudinal regression discontinuity (LRD) designs?*

5.5.2 Sample

For this study the MGP dataset is used as this allows analysis of performance for consecutive school years. In addition to the weaknesses which have been discussed earlier in this chapter relating to the quality of the performance measure, there is also a weakness of the MGP data relating to the ability to produce RD estimates. Namely, that the pupil date-of-birth (DOB) is given by month and year and so the specific day is not identified. This is not considered a major difficulty given that age effects, in keeping with previous studies (Luyten et al., 2009), are estimated as fixed effects for the entire sample. Moreover, given that the age effect is found to be linear and calculated using the large MGP sample, this lack of fine-grained information will have little effect on a linear trend. It will mean, however, that there will be some level of error in controlling for the effect of maturity, especially for smaller groups or individuals. As each pupil's DOB is recorded as the 15th of the month, pupils may be as much as half a month away from the recorded age value against which their results are adjusted. Given the small likely size of this discrepancy and the likelihood that some of the bias will be smoothed when considering school-level effects, the problem posed by this weakness for the present purposes is small.

The MGP dataset is designed to be a nationally representative sample of English pupils (DfE, 2011). The following table, Table 5.5.2a, gives the year group and number within each cohort within the sample for each of the three time periods. This table is referred back to when discussing the four measures used in the study. In particular, note the year groups which are emboldened; these are time-period/cohort combinations for which it was possible to estimate overall and individual school effects for all of the measurement designs. The results section gives subscripts with each measure denoting the cohort/time combination. Subscripts i, ii … to x refer to estimates from T2-Yr4, T2-Yr5 … to T3-Yr9, respectively. Note that the RD requires the earlier year group in the same school which rules out Yr3 and Yr7 (the first year group in each key stage). The longitudinal regression discontinuity design requires estimates from the same cohort in the previous time period – this rules out all of T1.

Table 5.5.2a – Year group and number within each cohort by time period

	Time Period					
Cohort	**T1** (2007/08)		**T2** (2008/09)		**T3** (2009/10)	
	Year Group	*N*	*Year Group*	*N*	*Year Group*	*N*
A					Yr3	9,831
B			Yr3	13,132	***Yr4***	***10,232***
C	Yr3	13,356	***Yr4***	***13,401***	***Yr5***	***10,469***
D	Yr4	13,895	***Yr5***	***14,031***	***Yr6***	***10,584***
E	Yr5	13,964	***Yr6***	***13,848***	*Yr7*	*10,081*
F	Yr6	14,210	*Yr7*	*14,555*	***Yr8***	***10,441***
G	Yr7	14,673	***Yr8***	***14,305***	***Yr9***	***9,738***
H	Yr8	14,869	***Yr9***	***14,129***		
I	Yr9	14,934				
Total*		99,901		97,401		71,376

*There were 141,057 unique pupils with a recorded score in at least one year.

An initial inspection of the data indicated that the data were suitable for VA and RD analysis. A small minority (3.6%) of cases has one of the following problems a) no recorded year group corresponding to any study time period or b) conflicting information between two or more time periods or c) expected year group based on DOB did match one or more recorded year group. For 3.1% of these, there were no recorded year groups for any of the time periods for which data was collected. At least 0.5% of pupils, therefore, were in the 'wrong' year group according to the cut-off, but this figure may be as high as 3.6% depending the levels of discrepancy in the missing data. The actual figure is likely to be somewhere in between these figures. These 5372 (3.6%) cases were omitted from the analysis.

Initial estimates showed generally consistent age and grade effects (see results), although some volatility was apparent in the monthly scores (age effect). In this dataset there are approximately 1000 pupils for each month and this volatility did not obscure the clear overall linear trend. This might, however, be a concern for future use of the RD as this linearity is likely to break down to exhibit a 'saw tooth' pattern when looking at smaller groups. These data suggest that the effects of maturity will be difficult to discern in smaller datasets.

5.6 Study 3 – Stability over Time

5.6.1 Overview and Research Questions

The third study addresses the temporal stability of the English value-added measure and the stability of estimated CVA performance for given MGP cohorts over time. This study is made up of two main parts. The first is a replication of Gorard et al. (2012) and of Leckie and Goldstein (2011) using more up-to-date data. These previous studies examined the stability over time of the former CVA measure. The first part of this study brings this up-to-date by examining the stability of the current VA model over time. School-level VA scores from 2011-2014 from publically available data are matched and correlations over successive years found. After finding these correlations, these are compared with the results of Gorard et al. (2012) to see whether the VA scores are more or less stable than the CVA scores. Following on from this, the second question addressed is whether the rates of stability of school value-added scores vary by initial estimated school performance. The final analysis within this section looks at the stability of cohort performance over time. This requires use of the MGP data described in the previous study. A CVA measure of performance is created to estimate performance from the last key stage for primary and secondary age pupils. Then, the relative performance of a school's cohort (compared to the overall cohort) can be examined over time.

Research Questions

Study 3 addresses the following research questions:

Table 5.6.1a - Primary and Secondary Research Questions for Study 3

RQ 3.1 ***How stable is the current English value-added measure across several years?***

RQ 3.2 ***Is the rate of stability in value-added scores associated with school performance?***

RQ 3.3 ***How stable is the contextual value-added performance of a given cohort over time?***

There are no sub-questions as it is possible to address these directly. Note that examination of English value-added measures (RQ 3.1 and 3.2) involves examination of both primary and secondary level results.

5.6.2 Sample

For the first analysis, the publically-available data discussed in the first study in Section 5.4.2 is used. This consists of school-level results for all maintained, mainstream English schools. The current VA model has been used since 2011 and the most recent performance data available are for 2014. As a result 2011 to 2014 data are combined using school and local authority establishment codes.

The second analysis draws on the MGP sample, as described in Section 5.3.3 and Section 5.5.2. The CVA performance of cohorts B to H (see Table 5.5.2a) could be estimated for two or more consecutive time periods. CVA could be estimated for three consecutive time periods for cohorts C, D and G. Note that cohorts E and F had results which crossed from primary to secondary education and so only two consecutive years' results from the same phase are considered. Further details of the number of schools used in the analysis are given in the results section alongside the estimates.

5.7 Study 4 – Cohort Consistency

5.7.1 Overview and Research Questions

The fourth and final study examines the issue of consistency within value-added measures. This study uses the MGP dataset and the analysis was conducted at the same time as the second part of the last study (concerning cohort CVA stability over time). As a result, the same data are used as well as the same CVA scores described in the previous section. This study is merely an analysis of the consistency between the CVA scores of different cohorts within a school rather than the stability of a given cohort over time (in the last section). This has been organised into a distinct study because of the issue addressed by each of these questions.

This study also poses two more research questions. These consider the extent to which an overall mean school effect reflects the underlying pupil value-added (RQ 4.2) and whether this varies by mean school performance (RQ 4.3).

Research Questions

Study 4 addresses the following research questions:

Table 5.7.1a - Primary and Secondary Research Questions for Study 4

RQ 4.1	***How consistent are value-added estimates of performance across cohorts from within a single school in a single year?***
RQ 4.2	***How consistent is performance within cohorts?***
RQ 4.3	***Does within-cohort consistency vary by mean school performance?***

There are no sub-questions as it is possible to address these directly.

5.7.2 Sample

RQ 4.1 uses the MGP sample, as described in Section 5.5.2. The CVA performance of all cohorts (A to I) is estimated for all three study years and cohorts within the same school at a given time are compared (i.e. NC years 3-6 in primary and NC years 7-9 in secondary).

RQ 4.2 and RQ 4.3 draw on 2013 pupil-level data at secondary level (KS4) and 2012 pupil-level data at primary (KS2) level. These were the most recent data available at the time of analysis. Further details are given on these in Section 5.3.2 and previous studies using these data (see Section 5.4.2). One further step taken was to drop very small schools from the analyses. At secondary level, cohorts with less than 20 pupils were dropped. In total, 77 (0.01%) pupils in 12 schools were dropped. At primary level, cohorts of less than 5 pupils were dropped. In total, 887 (0.17%) pupils in 301 institutions were dropped.

6. Results

6.i Chapter Introduction and Organisation

(N.B. This sub-section is marked as Section 6.i rather than 6.1 so that all subsequent section numbers correspond to the research question addressed, aiding navigation.)

This chapter is structured as a series of research questions followed by an empirical answer, all organised within four sections, one for each study. Immediately after each question, the section explains value of the question for addressing the core research question (see Chapter 5, Section 5.2.1) and how it relates to previous research reviewed in Chapter 4. Note that introductory and explanatory information for the overall study and details of the specific sample used is given in the methods chapter; model specifications for the statistical models used and descriptive statistics are placed in an appendix, unless these are essential to the analysis; and the majority of the discussion of the results is left until the dedicated discussion chapter (Chapter 7). The overall intention is to allow this results chapter to concentrate on posing, explaining and then answering each research question in turn.

6.1 Study 1 - Bias and Error

6.1.1 RQ 1.1 - Are there observable biases in the current English value-added measure?

The first study on bias and error is relatively straight-forward in terms of the problem of judging validity. Where biases (i.e. associations between a value-added score and factors which are ostensibly non-school factors) are observed, this is evidence that the measure in question is not entirely valid. In other words, value-added scores should not be systematically related to non-school factors. Observable bias and error are highly specific to the data and value-added measure in question. As the analyses in this section concern the English value-added measures and National Pupil Database (NPD), the results have direct implications in this context. Also, as explained in the methods chapter (Section 5.4.1), these results have wider relevance to the degree that the measures and data bear similarities with other systems or EER studies.

RQ 1.1.1 Are the current and former KS2-KS4 (Secondary) English value-added measures unbiased in relation to attainment?

Chapter 4, Section 4.2.2, discussed the problems associated with specifying the English value-added model such that it is independent of prior attainment. A problem identified by Gorard (2006c) was the high correlation between value-added scores and raw attainment scores. This problem was apparently fixed by the addition of contextual factors to create a CVA measure (Kelly and Downey, 2010) (but see Section 4.2.2 for further discussion of this point). The current English value-added measure (from 2011) no longer includes contextual variables. This raises the question of whether problems of non-independence between value-added and attainment scores will have re-emerged. This is the concern of this research question.

Table 6.1.1a, below, presents school-level correlations between English value-added scores and a) the measure of prior attainment used to create the VA measure and b) the measure of final attainment used as the outcome. The correlation between VA and final attainment estimates the extent to which VA gives unique information about school performance (Gorard, 2006c). Some level of correlation is expected with final attainment because schools with higher value-added will also tend to get higher final attainment as a result. The correlation between VA and prior attainment, however, raises concerns about non-school factor bias in the measure. The correlation with prior attainment scores should be minimal as value-added is intended to measure effectiveness independently of (prior) intake characteristics. The difficulties of this are discussed at length in Chapter 4, Section 4.2.2, which discussed how school-level correlations between VA and prior attainment can stem from inadequate (pupil-level) prior attainment measures leading to ‘phantom’ school-level relationships and/or the possibility that school-level ‘compositional’ effects are also present.

This analysis involved computing the pairwise correlations between these variables for all years from 2004 to 2014 for all maintained, mainstream schools in England. The specific NPD variables used are included in Appendix C (Table C.1a).

Table 6.1.1a - Correlation between KS2-4 value-added measures, prior attainment and final attainment 2004-2014 (School-level Data)

Year	KS2-KS4 Value-added Measure	KS2 Average Point score	Capped GCSE and equivalents point score.
2004	**KS2- Equivalents VA score**	.50	.85
2005	**KS2 to KS4 contextual value added score**	.01	.52
2006		.03	.43
2007	**KS2 to KS4 contextual value added score with shrinkage factor.**	.00	.42
2008		.01	.29
2009		.01	.27
2010		.00	.27
2011	**Best 8 VA measure**	.17	.75
2012		.13	.76
2013		.15	.77
2014		.29	.75

Table 6.1.1a shows how changes in the value-added measure can influence the correlations with measures of prior and final attainment. These results show that the very strong link found for the 2004 data in Gorard (2006c) is no longer present. However, a small correlation between prior attainment and value-added remains in the current data, with a slight increase in 2014. This means that the current value-added is not entirely independent of intake prior attainment.

An important methodological point arising from this analysis is that controlling for prior attainment at pupil-level does not necessarily translate into independence at school level. The value-added models above were calculated using pupil-level data, controlling for the relationship between prior attainment and final attainment without consideration of school membership. The re-emergence of a link between prior attainment and value-added when the scores are aggregated to school-level suggests that pupils whose peers had high prior attainment tend to outperform similar pupils whose peers had lower prior attainment. The strength of this school-level link has varied with changes to the value-added measure. The contextual value-added (CVA) models completely removed the influence of prior attainment at school level. The CVA models included a school-level mean KS2 score (and its standard deviation) (Kelly and Downey, 2010) and so, by design, would correct any linear association between attainment and value-added. By only using pupil-level data, more recent models have not been able to account for this difference. Schools therefore benefit from taking intakes with high average prior attainment. There is also a change in the correlation between VA and final attainment between 2007 and 2008. The most likely cause of this is a more predictive VA model, either due to inclusion of a greater number of contextual variables or an improvement in the quality of contextual or attainment (prior or final) data.

Later in study 1 (RQ 1.1.3), there is found to be an apparent ‘grammar school effect’, where selective schools consistently perform above expectations. It is possible that the small positive associations between VA and prior attainment found above in the 2011-2014 data were due to this grammar school effect. This possibility was examined by removing selective schools from the analysis in the 2014 data and re-estimating the relationship. The resulting correlation between VA and KS2 scores fell from .29 to .20, suggesting that selective schools were a particularly marked case of a more general pattern rather than the reason for the small school-level association between prior attainment and value-added (also see RQ 1.3.1).

RQ 1.1.2 Are the current and former KS1-KS2 (Upper Primary) English value-added measures unbiased in relation to attainment?

Similar to the secondary level results in RQ 1.1.1, the association between value-added scores and attainment scores at primary level was calculated for all maintained, mainstream schools in England. The results are in table 6.1.1b, below. Note that RQ 1.1.1 (above) went back to 2004 to correspond with results in Gorard (2006c) and show key model changes. This time, correlations are presented for recent data only (2011-2014) as these reasons do not apply. Again, the specific NPD variables used are included in Appendix C (Table C.1b).

Table 6.1.1b - Correlation between KS1-2 value-added measures, prior attainment and final attainment 2011-2014 (School-level Data)

Year		Average Point Score at KS1	Average Point Score at KS2
2011	**KS1-2 English and Maths VA Measure**	-.15	.59
2012		-.20	.54
2013		-.20	.56
2014		-.18	.58

As with the secondary results, there are small but appreciable correlations between the measure of prior attainment (in this case KS1 average point score) and the school value-added score. This time, the correlation is negative, indicating that cohorts with high average attainment at KS1 tend to do worse in terms of value-added at KS2. It is worth noting that English schools tend to take pupils from the start of KS1 (often with a reception year before this) to the end of KS2. KS1 scores are generally, therefore, a measure from the mid-point of the age range for most English schools. Compare this with the KS2-4 scores (above), where KS2 is generally recorded at a separate primary school prior to entering the secondary school. What this means is that primary schools who perform well at KS1 tend to be disadvantaged for KS1-2 value-added; in contrast, secondary schools who take cohorts with high prior performance at KS2, tend to be advantaged. The most likely explanation for this is that schools who push pupils to the limits of their capabilities at KS1 or artificially inflate pupil scores through (teacher-assessed) test preparation make value-added more difficult in KS2 without any other benefit. Secondary schools taking high-attaining KS2 cohorts may face a similar disadvantage in relation to the baseline but this is offset by more favourable intake characteristics associated with higher *average* attainment (also see RQ 1.3.1).

RQ 1.1.3 Is the current KS2-KS4 (Secondary) English value-added measure unbiased in relation to available theoretically important variables?

Chapter 3, Section 3.4.4, introduced the new English ‘Progress 8’ measure. It was pointed out that, like the current VA measure, this would not include contextual variables and there were concerns over the likely level of bias in the measure as a result (Burgess and Thomson, 2013a, p.7). Burgess and Thomson (2013a) do not provide school-level estimates of bias nor estimates at primary level (see RQ 1.1.4, below). The magnitude of biases in the school results is not clear from pupil-level results alone as it will depend on how pupils with different levels of performance are distributed across schools. The extent to which school-level VA scores are biased according to intake characteristics due to the policy decision to ignore contextual variables is addressed by this research question. The starting point for this is a school-level multiple regression analysis of 2014 KS2-4 school value-added scores against a number of theoretically important (school-level) pupil background variables. See

Appendix C2 for the specification of the regression models and a note about the use of school-level data.

Table 6.1.1c – Results of a School-Level Multiple Regression Analysis of the 2014 KS2-4 Best 8 Value-Added Measure on a Number of Intake Characteristics

	Coefficient Value	**Standard Error**	**t**
Proportion* of Pupils[†] with Special Educational Needs (SEN)	-3.2	7.6	0.42
Proportion of pupils with English as an Additional Language (EAL)	74.0	2.6	30
Proportion of Pupils on Free School Meals (FSM) or with Looked After Status	-76.8	3.7	21
Total Number of Pupils at the End of KS4	0.014	0.006	2.1
Cohort Average KS2 Attainment (APS)	1.5	0.4	4.2
Coverage (Proportion of Pupils Included in the Measure)	-8.8	8.5	1.0
Proportion of female pupils	18.3	2.2	8.5

Model $R^2 = 0.35$, n = 2990 schools, all figures to 1DP apart from total pupils, reported to 3DP

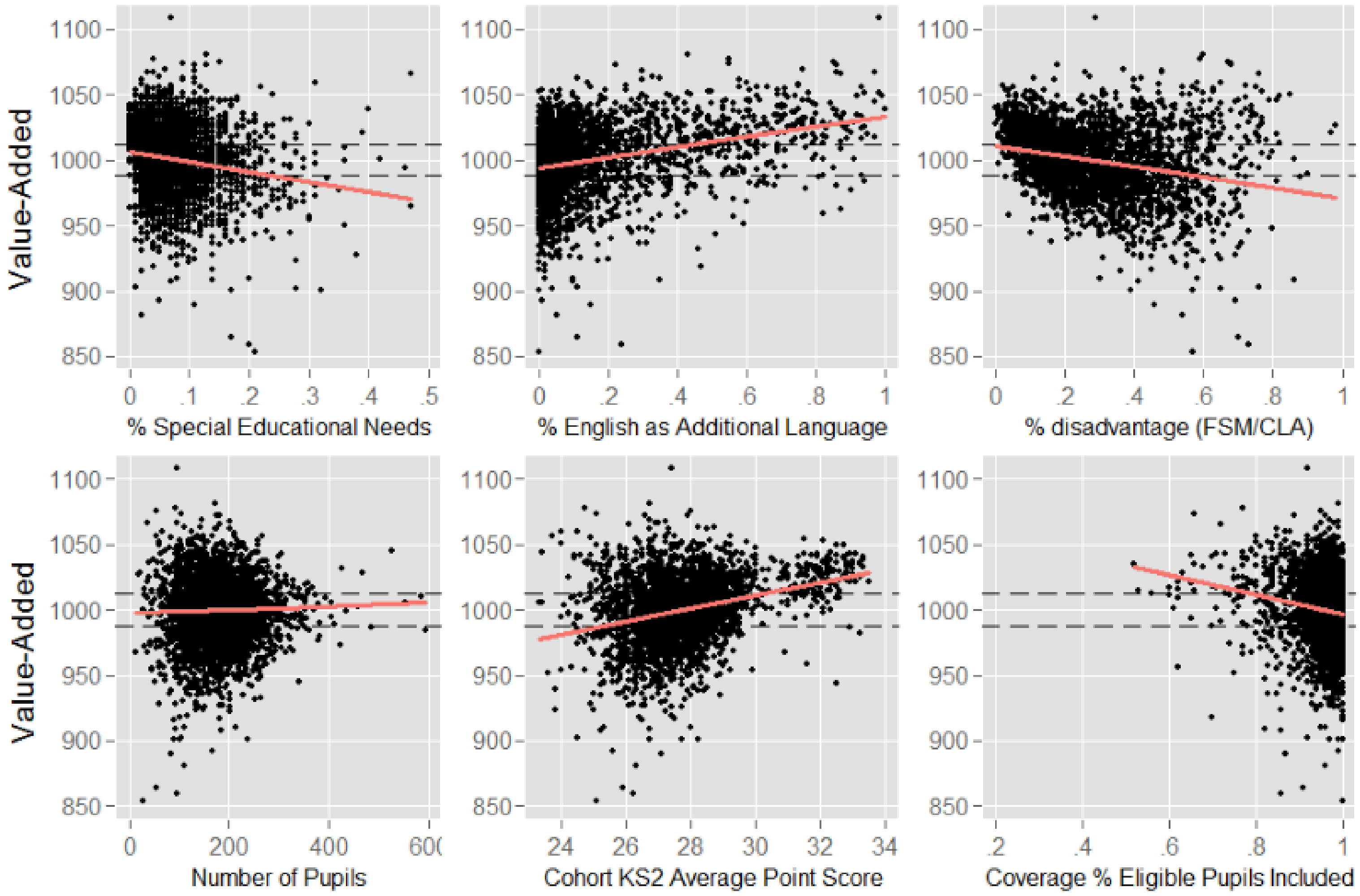
[†] Figures relate to the cohort rather than the overall school

These results show a number of positive and negative biases on the value-added scores. The analysis was conducted with unweighted school-level data and therefore take schools as the unit of analysis, without accounting for their size. This is appropriate for school-level analysis and allows us to examine bias on school value-added scores. Note that a weighted or pupil-level analysis would be required to draw conclusions at pupil-level. A school with 50% of its pupils with English as an additional language, for example, would on average receive a score 37 points (0.5*74), higher than a school with none. This example equates to 6 GCSE grades per pupil. Over a third (35%) of the variance in VA scores can be explained using these variables. The most important of these are, first, the percentage of pupils with English as an additional language (EAL), which is positively associated with school value-added; second, the rates of disadvantage, as measured by pupils in receipt of free school meals (FSM) and/or looked after by the local authority, which has a strong and consistent negative association with school value-added performance. These two variables, when included as the only independent variables in the regression account for 33% of the variance alone. There are appreciable differences related to other variables but these are smaller and/or more inconsistent in comparison to these two variables. One exception is the gender balance which is looked at further below.

To illustrate these differences, the school variables above (other than the percentage of girls, considered separately) are plotted against school value-added scores. This is shown in Figure 6.1.1a, below. Figure 6.1.1a gives a number of univariate comparisons. The limitation of presenting univariate comparisons is the likelihood that there will be association between the various characteristics of school intakes. If there is a high level of association between rates of disadvantage and rates of pupils with English as an additional language, for example, the plots will give a misleading indication of the bias expected for schools with these characteristics. Despite these drawbacks, the graphs give a clear picture of the distribution of outcomes and how they relate to various intake factors. Linear trend lines fitted using ordinary least squares are added as well as horizontal (dashed) references lines at 12 points above and below the statistical expectations on the 'Best 8' attainment measure. A difference of 12 equates to 2 GCSE (or equivalent) grades on average per pupil above or below those expected across pupils' best 8 examination grades. A pupil expected to achieve 8 GCSE C grades at a school on the upper reference line, for example, would instead receive 2 B grades and 6 C grades.

Figure 6.1.1a clearly shows a number of relationships between value-added scores and a number of intake background variables. As with the regression results, the number of pupils with English as an additional language stands and the proportion of disadvantaged pupils stands out as a large source of bias. Another feature which stands out particularly from the plots is the presence of a cluster of selective 'grammar' schools in the cohort KS2 average point score graph. This is in line with previous research (Leckie and Goldstein, 2009, Coe et al., 2008) and recent analysis undertaken at the independent research centre, Education Datalab (Allen, 2015b). In Section 6.1.1, looking at 2014 data, it was found that a similar but smaller relationship remains without this group of schools so it is likely that this cluster of selective schools is a particularly marked example of a particular phenomenon.

Figure 6.1.1a – 2014 School KS2-4 Value-Added Scores against selected Contextual Variables at School-Level*†

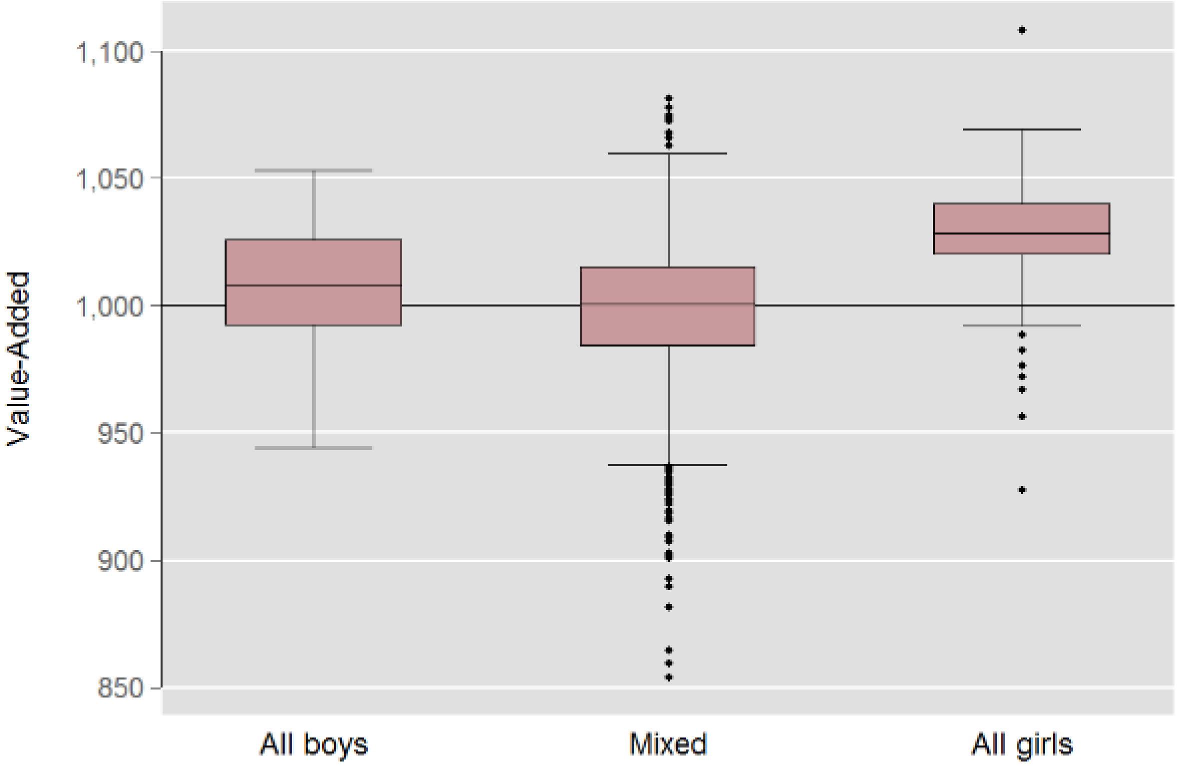


* The solid red line is a fitted trend line, estimated using ordinary least squares

† The dashed horizontal line gives 12 points above and below the neutral value-added score of 1000 on the 'Best 8' attainment measure (with English and Maths bonus), equivalent to approximately 2 GCSE grade changes across the Best 8 GCSE scores for all pupils at the school.

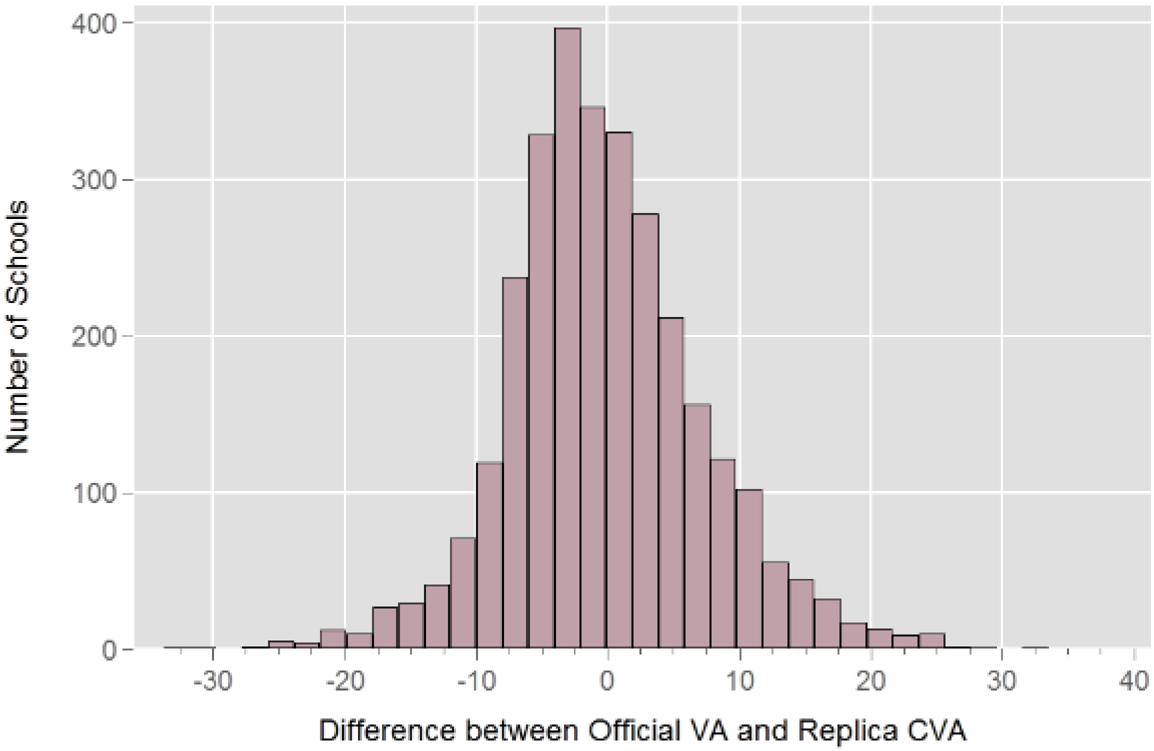
Similar to the grammar school ‘effect’, a positive effect on value-added was found for schools with single-sex intakes (in the regression model but not shown in Figure 6.1.1a above). Adjustments for gender in former CVA models tend to adjust expectations downwards for boys (Evans, 2008). Analysis showed that all-girls schools tended to have higher VA scores, as expected, but all-boys schools, rather than having lower than average scores, showed a positive bias, albeit much smaller than that for all-girls schools. The bivariate relationship between value-added and single-sex status is given in Figure 6.1.1b, below. There is considerable overlap between the grammar school effect and the single-sex school effect. The single sex effect does, however, appear to be related to value-added in its own right. When a single-sex dummy variable was included in the regression analysis in Table 6.1.1c, above, the single-sex effect was estimated at 3.9 points. This suggests that there is a single-sex school ‘effect’ over and above a) the relationship between gender and value-added and b) the relationship between average cohort prior attainment and value-added. As has been discussed in Chapter 4, it is difficult to know whether this and the grammar school effects are genuine peer (or compositional) effects or produced by measurement error or unobserved differences.

Figure 6.1.1b – Boxplots of School-Level Value-Added by Intake Single-Sex Status



The final analysis presented within this section is a simulation of what the school scores would be if a CVA model was used. This involved replicating the actual 2013 DfE VA model (DfE, 2013a) and then adding contextual variables to create a replica CVA measure. This replica CVA measure was then compared to the official VA measure to estimate the impact of the policy-decision to ignore contextual factors on the school-level value-added scores. As well as the original VA variables, the replica CVA model took the following variables into account: Pupil gender, pupil SEN status, pupil EAL status, pupil FSM status, pupil IDACI score (a measure of neighbourhood deprivation) and the IDACI score squared. The model specifications for the VA and CVA replicas and further technical details are given in Appendix C3. School-level scores were produced from the CVA model and the difference between the official VA score and this replica CVA score is calculated. This difference variable shows the change to school actual VA scores which would take place if the theoretically important contextual variables that have been discussed were also taken into account in the measure. The frequency distribution for this variable is shown in Figure 6.1.1c, below:

Figure 6.1.1c – The Effect of Taking Pupil-Level Contextual Factors into Account on 2013 School KS2-4 Value-Added Scores



Recall that a score of 6 equates to 1 GCSE grade *per pupil* across the Best 8 GCSE measure. This graph shows that CVA scores can be up to 30 points above and below the official VA scores (scores range from -33.7 to 33.5). This simulated CVA measure took 5 pupil-level contextual factors into account. Differences are likely to be larger if school-level variables or other variables, such as those used in the official CVA measure used between 2005 and 2010, were also taken into account.

The histogram above is presented below as a table giving the number of schools for each magnitude of difference between the official VA and the replica CVA score. The totals given are cumulative. These results show that of the 3017 mainstream, maintained schools included in the analysis, 1116 – over a third - would change their performance score by 1 or more grade at GCSE per pupil if a contextualised VA measure was used. Just over 10% of schools (310) would see results change by 2 GCSE grades or more per pupil.

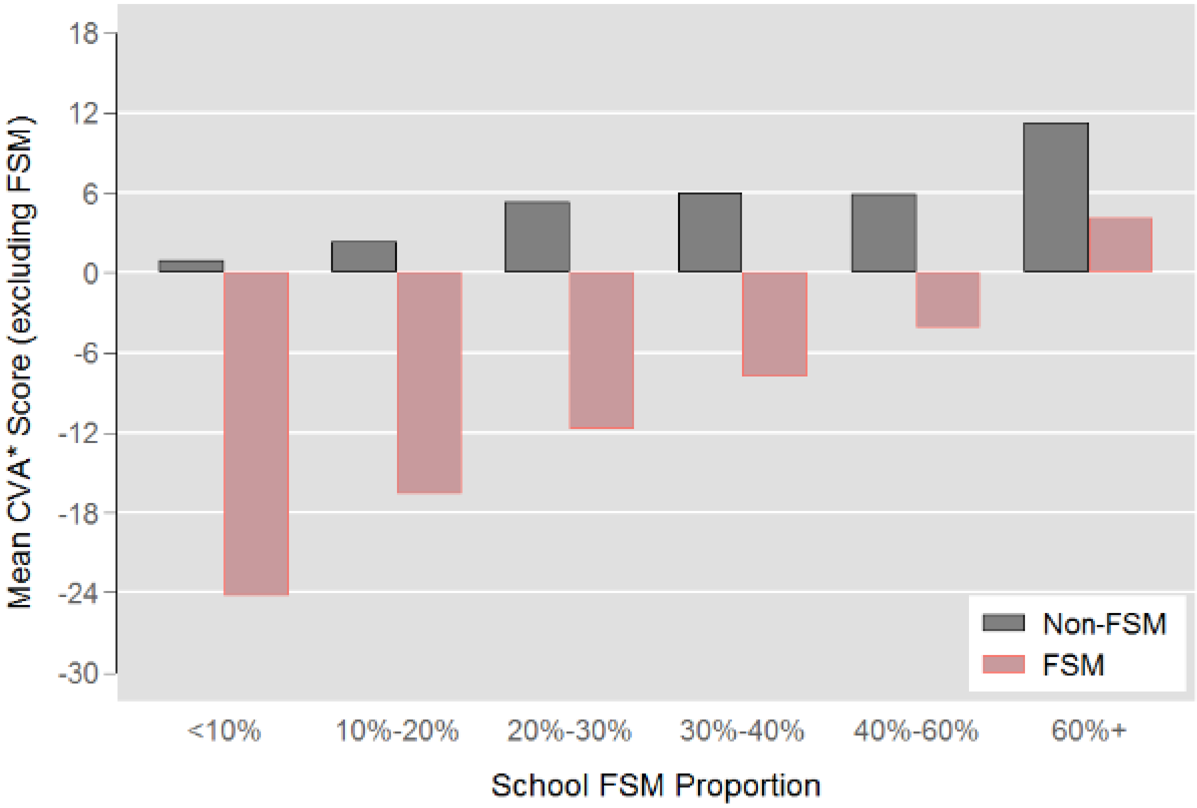
Table 6.1.1d – Number of Schools for each Level of Difference between the Official 2013 KS2-4 Value-Added Measure and a Simulated CVA Measure

Size of Change	**Increased Score** (no. schools)	**Decreased Score** (no. schools)	**Total**
<3 Points Change	481	550	1031
3 Points or more	890	1096	1986
6 Points or more	552	564	1116
12 Points or more	181	129	310
18 Points or more	49	34	83
24 Points or more	13	9	22
30 Points or more	1	2	3

Total number of schools = 3017

One final noteworthy finding arising from the analysis above relates to the free school meals variable. In the simulated model for the CVA measure above, free school meals eligibility predicted a lower Best 8 score by 18 points. During the creation of this model, it was investigated whether this negative free school meals effect varied according to the proportion of disadvantaged pupils in the given school. The school-level FSM proportion was estimated and this was multiplied by pupil-level FSM eligibility variable to form an interaction variable between these two. When this interaction term was included in the model in addition to all other variables, the original FSM variable now predicted 28 points lower attainment. The new interaction variable had a large positive coefficient of 44. This coefficient can be multiplied by the proportion of pupils at the school on FSM to estimate the magnitude of the effect. For example, if the proportion of pupils in a school was 50%, this interaction term would offset the negative association between attainment and FSM by 22 points (44 x 0.5). This relationship is most clearly seen in the following graph which plots the relative performance of FSM and non-FSM pupils according to the proportion of pupils in the school on FSM status. This was created by excluding the two measures of disadvantage (FSM and IDACI, see method) from the CVA model used above so the effect of disadvantage could be seen independently of all other factors.

Figure 6.1.1e – Mean Difference in CVA performance for Non-FSM and FSM status pupils by School FSM Proportion in a CVA model excluding measures of disadvantage



*CVA model excluding the FSM and the IDACI and IDACI squared variables

This graph shows the remarkable finding that the greater proportion of pupils in a school on FSM, the lower the penalty associated with the status. There is also a small non-linear relationship between school FSM proportion and the performance of non-FSM pupils. As with findings earlier in this chapter, it is hard to know whether these differences reflect differences in school performance, measurement error or unobserved differences. It could be, for example, that the thresholds used to identify pupils as disadvantaged reflect different levels of disadvantage depending on the proportion of disadvantaged pupils in the local area. Whatever the case, the systematic nature of the relationship suggests this should be regarded as bias.

RQ 1.1.4 Is the current KS1-KS2 (Upper-Primary) English value-added measure unbiased in relation to available theoretically important variables?

RQ 1.1.4 follows the same analyses in RQ1.1.3 but using KS1-2 data rather than KS2-4 data. As with the secondary results, the first analysis was a multiple regression analysis to estimate the strength of the correlation between (KS1-2) school value-added and a number of theoretically important contextual variables. As before, this analysis was at school-level and was conducted on unweighted school-level scores, not taking school size into account. As such the conclusions drawn all pertain to the school-level scores. The variables used were the same as in the secondary-level analysis, with the exception of prior attainment now corresponding to KS1 rather than KS2 attainment (see Appendix C2 for the model specification for the secondary results for reference). A summary of these regression results are shown in Table 6.1.1e, below.

Figure 6.1.1e shows that the contextual variables account for a smaller proportion of the VA variance at primary level than at secondary level, with an R^2 value of 10%. This may stem from the greater variability in the smaller cohorts at primary level, or possibly the greater ability of education for earlier ages groups to countermand the influence of social background. The background factors which have the greatest substantive significance are the percentage of pupils with SEN, the percentage of pupils with EAL, the percentage of pupils on FSM or looked after and the KS1 attainment of the cohort. It should be noted that none of these associations necessarily reflect a causal relationship between the factor and value-added performance.

Table 6.1.1e – Results of a School-Level Multiple Regression Analysis of the 2014 KS1-2 Value-Added Measure on a Number of Intake Characteristics

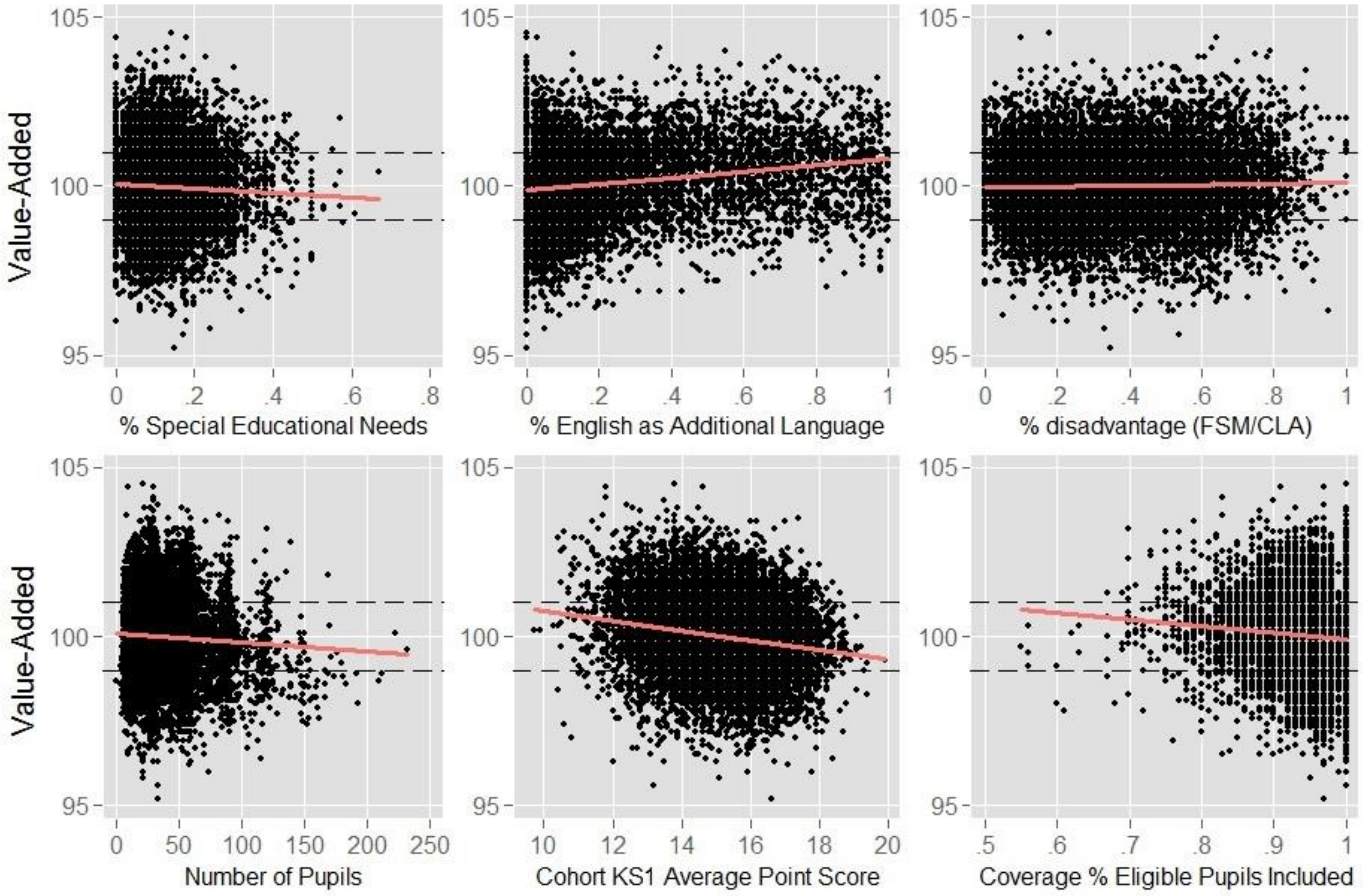
	Coefficient Value	Standard Error	t
Proportion of Pupils[†] with Special Educational Needs (SEN)	-1.5	0.1	13.1
Proportion of pupils with English as an Additional Language (EAL)	1.1	0.0	24.2
Proportion of Pupils on Free School Meals (FSM) or with Looked After Status	-0.9	0.1	17.2
Total Number of Pupils at the End of KS2	-0.0	0.0	13.6
Cohort Average KS1 Attainment (APS)	-0.2	0.0	25.0
Coverage (Proportion of Pupils Included in the Measure)	0.0	0.2	0.1
Proportion of female pupils	-0.0	0.1	0.2

Model R^2 = 0.10, n = 14,292 schools, all figures to 1DP

[†] Figures relate to the cohort rather than the overall school

As with the secondary level results, the relationships are now shown on a series of graphs (see Figure 6.1.1f, below). Each graph puts school value-added on the y-axis against a school context factor in a bivariate comparison. Each data point is a primary school in England. There is a linear trend line on each scatter plot showing the systematic relationship between value-added performance and the factor in question. Also, each plot has two horizontal reference lines set at 1 National Curriculum (NC) average point score *per pupil* above and below expected performance, this equates to about 4 months' extra/lower progress since the previous key stage, 4 years earlier.

Figure 6.1.1f – 2014 School KS1-2 Value-Added Scores against selected Contextual Variables at School-Level[*†]



* The solid red line is a fitted trend line, estimated using ordinary least squares

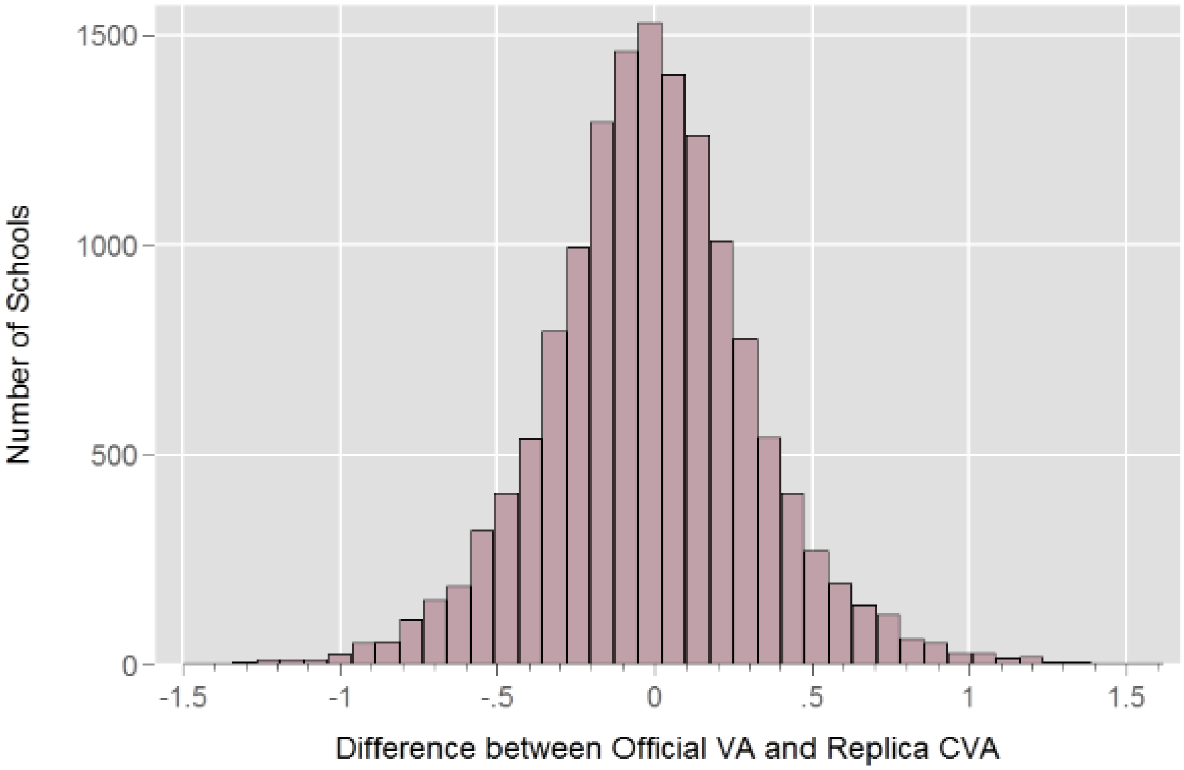
† The dashed horizontal line gives 1 National Curriculum (NC) average point score per pupil above and below expected performance, this equates to about 4 months' extra/lower progress since the previous key stage, 4 years earlier.

These scatter plots show a number of small to moderate biases. In line with the regression analysis, however, Figure 6.1.1f suggests that pupil background factors have less systematic effects at KS2 level. There are some noteworthy differences with the multivariate analysis such as the lack of association between disadvantage and value-added. It is likely that countervailing effects (such as a tendency for schools with high EAL rates to also have high FSM rates) are at play and so the multivariate results are likely to be more accurate as an estimate of effect.

As with the previous section, a replica pupil-level CVA model is produced to estimate the extent to which official VA scores would change should a CVA score be produced taking the variables from the regression analysis (above) into account. 2012 KS2 data are used as these were the most recent available. The model specification is identical to that used at secondary level, using the corresponding variables at KS1-2 (see Appendix C3 for the specification and analysis at secondary level for reference and see Appendix C4 for the model output and other technical details at primary level). The replica CVA model took the following variables into account: pupil gender, pupil SEN status, pupil EAL status, pupil FSM status, pupil IDACI score (a measure of neighbourhood deprivation) and the IDACI score squared. The difference between the official VA measure and the new replica CVA measure was saved as a new variable and plotted as a histogram, shown in Figure 6.1.1g, below.

Figure 6.1.1g shows that taking contextual factors into account in the primary measure in terms of the average NC levels of value-added. To put this into context, pupils are expected to progress by 3 points per year. 1 point, therefore, represents about 4 months' progress. This graph shows that CVA scores can be up to about 1.5 NC points, or 6 months' progress away per pupil from the official VA scores. Much of the distribution for the official school value-added scores lies between about -3 and 3 (over 98% of all schools). Therefore, in terms of range, the differences between the CVA replica and the official VA is about a third of this.

Figure 6.1.1g – The Effect of Taking Pupil-Level Contextual Factors into Account on 2012 School KS1-2 Value-Added Scores



Finally, this histogram is presented as a cumulative frequency table 6.1.1d, below, giving the number of schools above each difference threshold. It shows, for example, that 1841 of 14,321 schools would see their result altered by about 0.5 points (2 months' worth of progress per pupil) if this CVA measure were to be used.

Table 6.1.1f – Number of Schools for each Level of Difference between the Official 2012 KS1-2 Value-Added Measure and a Simulated CVA Measure.

Size of Change	**Increasing Score** (no. schools)	**Decreasing Score** (no. schools)	**Total**
<0.25 Points Change	4127	4437	8564
0.25 Points or more	2700	3057	5757
0.5 Points or more	859	982	1841
0.75 Points or more	266	267	533
1 Points or more	79	59	138
1.25 Points or more	13	12	25

Total number of schools = 14,321

6.1.2 RQ 1.2 - What is the level of missing data in the National Pupil Database?

RQ 1.2.1 What is the level of missing data in the variables used in the former KS2-KS4 English contextualised value-added measure and current value-added measure?

Previous authors have raised concerns about the extent of missing data in former versions of the National Pupil Dataset (Gorard, 2010). Where data are missing, these cannot be accounted for within a VA (or CVA) analysis. To estimate the scale of this problem the most recent pupil-level data available to this researcher were examined to record the level of missing data for the attainment and contextual variables studied in the last section (Section 6.1.1). This research question therefore examines whether more recent NPD extract contain high rates of missing data in the main variables which would be used for a VA or CVA measure. Below is a table of the level of missing data for the contextual variables used earlier in this study. 2013 pupil-level data are used. These are results for all maintained, mainstream schools.

Table 6.1.2a – Levels of missing data for attainment and contextual variables in 2013 National Pupil Database KS4 data.

Variable	Missing Values	Non-Missing Values
Best 8 Score Plus Bonus	0	567,273
KS2 Average Point Score	0	567,273
KS2 English Score (Fine graded including teacher assessment)	0	567,273
KS2 Maths Score (Fine graded including teacher assessment)	0	567,273
SEN Provision Status	0	567,273
Language Group Status	0	567,273
FSM eligibility	5,964	561,309
Gender	0	567,273
IDACI Score	7,375	559,898

These results show low rates of data marked as missing. All variables bar two had no missing observations (although see RQ 1.2.3, below). The highest rate of missing data was for the Income Deprivation Affecting Children Index (IDACI), with 1.1% of observations missing a value. As is discussed in the next chapter, the practice of using default values such as the mean or the modal status (Evans, 2008) may disguise the actual level of missing data.

RQ 1.2.2 What is the level of missing data in the variables which would be required for a KS1-KS2 contextualised value-added measure?

The analysis in RQ 1.2.1, above, is now repeated in relation to a KS1-2 (C)VA measure using KS2, pupil-level data from 2012 (the most recent pupil-level data available to this researcher). Below is a table of the level of missing data for the attainment and contextual variables likely to be used for a CVA measure. These are results for all maintained, mainstream schools.

Table 6.1.2b – Levels of missing data for attainment and contextual variables in 2012 National Pupil Database KS2 data.

Variable	Missing Values	Non-Missing Values
KS2 APS (Fine graded)	548	531,621
KS1 APS (Fine Graded)	23,989	508,180
KS1 Maths Deviation (from APS)	24,057	508,112
KS1 English Deviation (from APS)	24,022	508,147
SEN Provision Status	0	532,169
Language Group Status	0	532,169
FSM eligibility	1,228	530,941
Gender	0	532,169
IDACI Score	2,799	529,370

As at secondary level, these results generally show low rates of missing data. The exception to this is the KS1 results, for which just under 5% of the total are missing. If these pupils are concentrated by school, there is the potential for appreciable bias in the scores for affected schools. Contextual variables show much lower rates: the IDACI score, for example, is missing in 0.5% of cases. As with the secondary results, the use of default scores and categories is likely to disguise substantial rates of missing data.

RQ 1.2.3 Are ceiling effects, floor effects or scale discontinuities present in the main Key Stage (1-4) performance scores?

The quality of the main attainment measures are of paramount importance to the quality of a value-added measure. Although the validity of the actual scores is difficult to verify, one could expect the overall distribution of scores to be approximately normal with no clear ceiling effects, floor effects or obvious discontinuities. Positive or negative skews may stem from the actual distribution of 'real' attainment levels; ceiling effects, floor effects and discontinuities are highly suggestive of artefacts of- or limitations with- the measurement itself. This analysis involved visual inspection of the Key Stage (KS) 1 to 4 attainment distributions for the most recent data available to the researcher. Distributions for KS results KS1, KS2 and KS4 are given below. The first of these is the KS4 Best 8 GCSE measure (plus English and Maths score bonus) in 2013, see Figure 6.1.2a, below:

Figure 6.1.2a – 2013 KS4 Pupil Attainment Distribution on the Best 8 Score (Plus Bonus) Measure

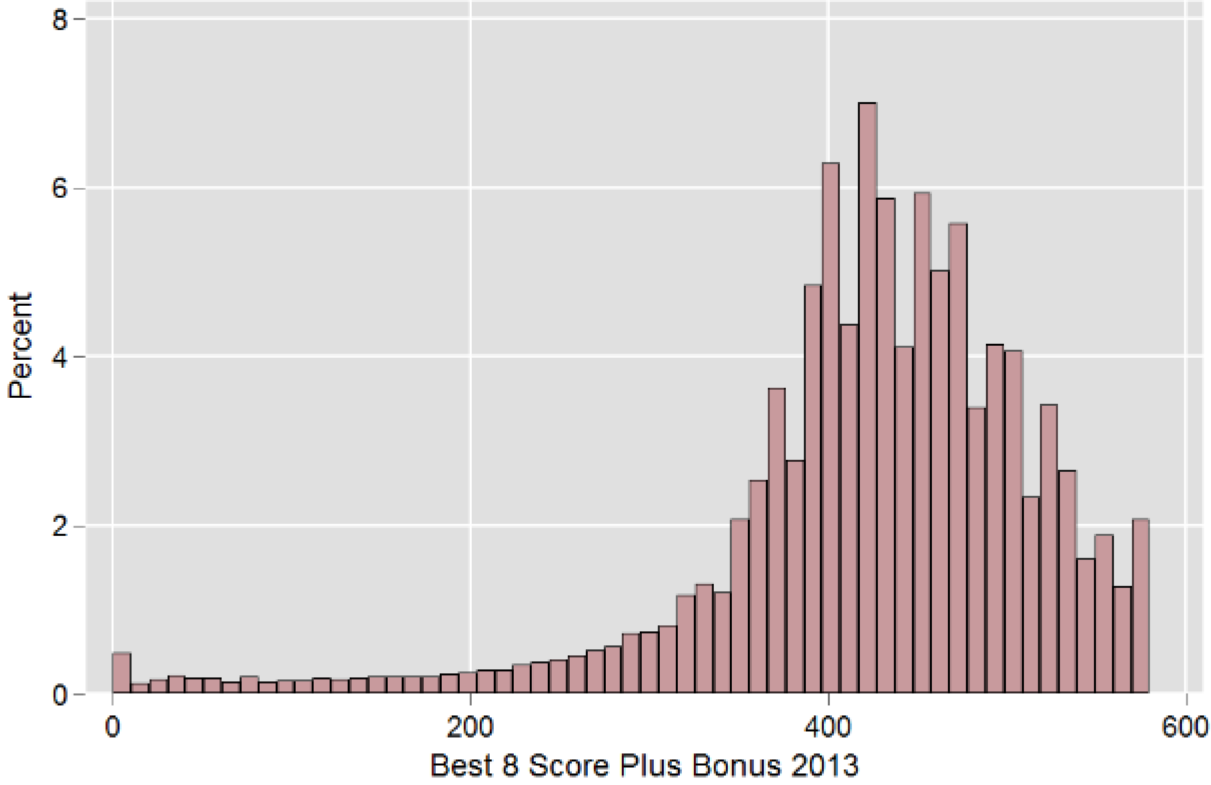
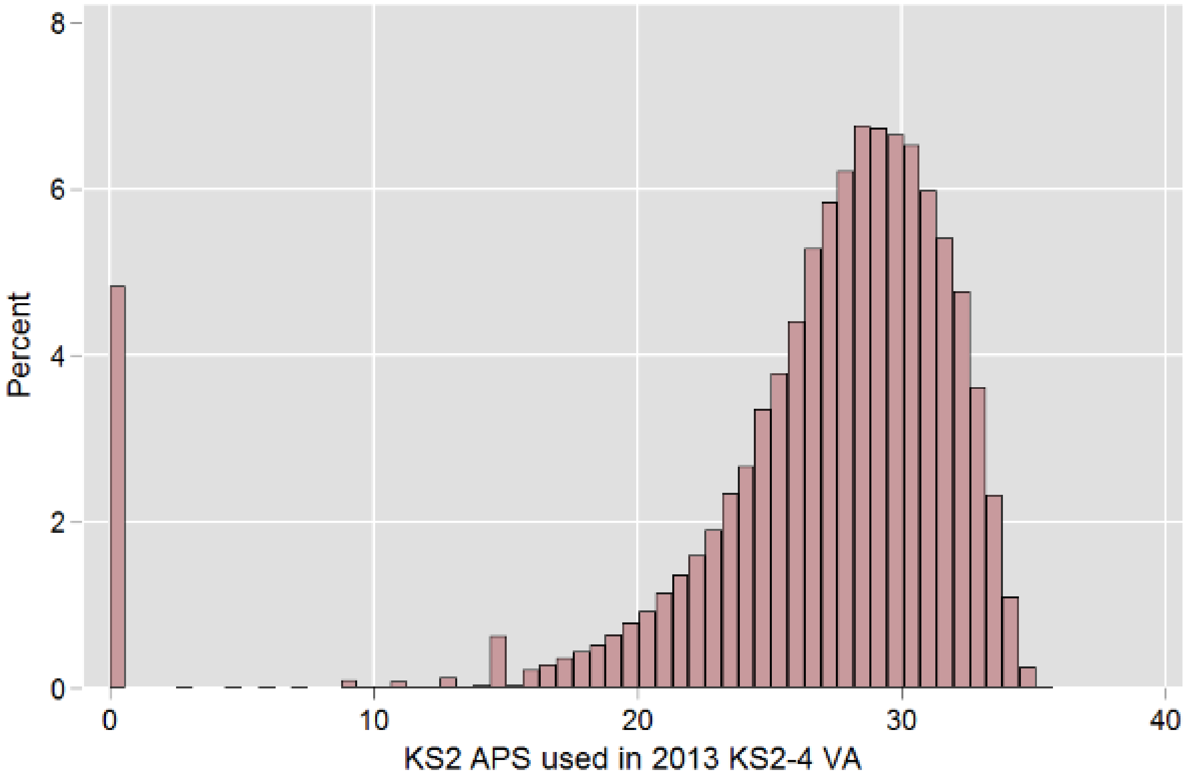


Figure 6.1.2a (above) raises two suggestions of non-normality in the data: First, the presence of a long bottom tail of low attaining pupils. This does not necessarily reflect a measurement problem but may have implications for models assuming data normality. More problematic is the apparent ceiling effect on the right of the distribution. The maximum score of 580 has apparently artificially capped the scores of approximately 1% of the national cohort (estimated through inspecting the excess frequency on the final bar over the normal distribution trend).

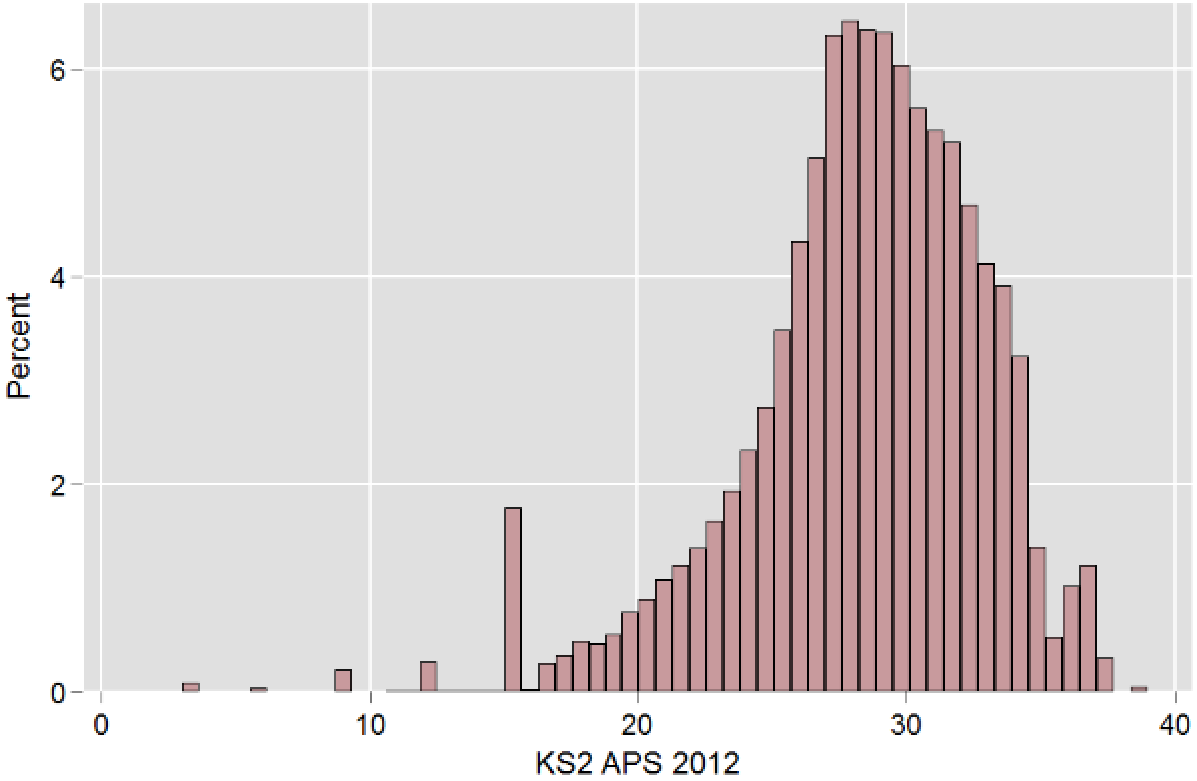
Figure 6.1.2b, below, shows the KS2 scores used for the 2013 KS4 Best 8 VA measure. The KS2 scores for the 2013 KS4 cohort were recorded in 2008. This distribution shows a large percentage (around 5%) of pupils who had failed to score on the test and were given a score of zero. At the other end of the distribution, frequencies fall sharply past a score of about 33.

Figure 6.1.2b – 2008 Pupil KS2 Attainment Distribution



It is useful to contrast this with a more recent KS2 cohort. The most recent available at the time of analysis was for the 2012 KS2 results. These are shown in Figure 6.1.2c, below. Figure 6.1.2c shows that both of the problems identified the 2008 data (above) have been reduced: the proportion of pupils with the floor score, now 15, is reduced and there are examples of pupils achieving scores above 35, due to a greater number of submissions to the level 6 KS2 tests.

Figure 6.1.2c – 2012 Pupil KS2 Attainment Distribution

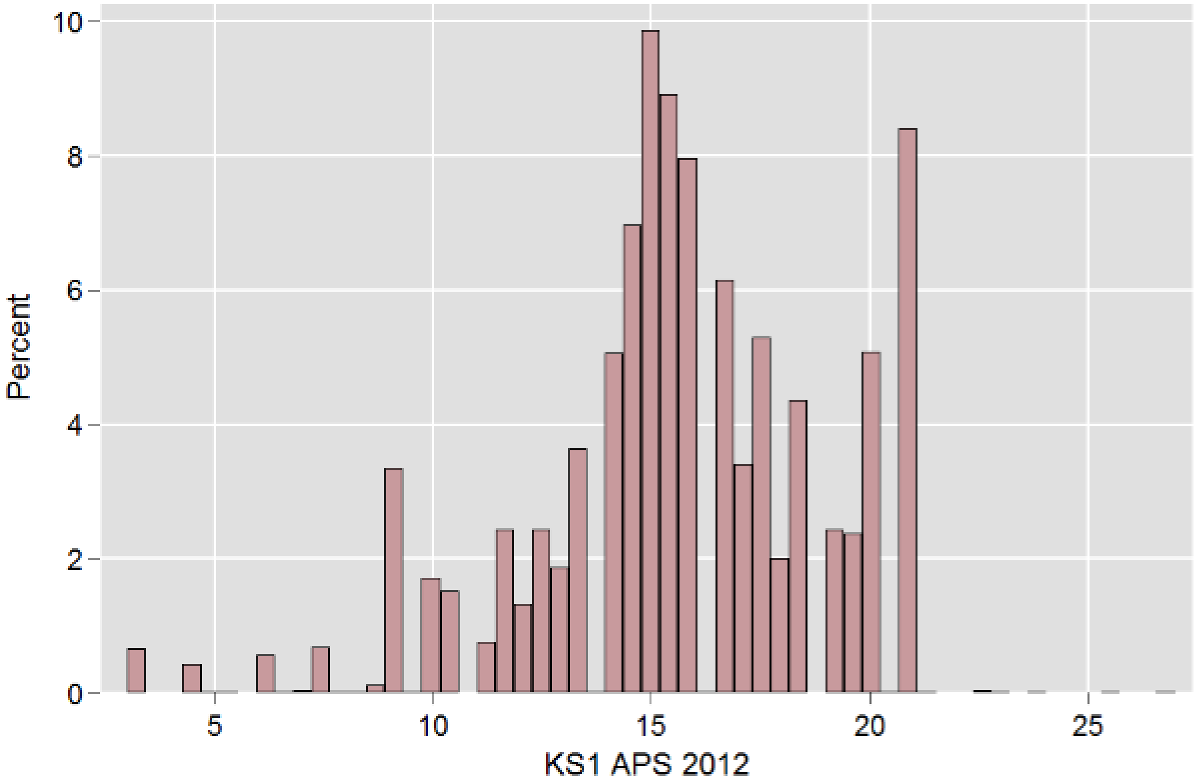


Although Figure 6.1.2c still is some way from a 'neat' normal distribution, the data appear to be improving over time.

The final distribution examined is the KS1 attainment distribution for 2012, the most recent available to this study. This is given in Figure 6.1.2d, below. Figure 6.1.2d shows that KS1 scores are some way away from being a continuous, normally distributed variable as one would expect from a robust attainment measure distribution. Scores appear to be less fine grained and there are suggestions of floor and ceiling effects at 8 and 22, respectively. These problems have implications for the use of the KS1 attainment measures as a baseline for the KS1-2 value-added measure.

Before moving to the final research question in this study, let us consider the results within this section (concerning RQ 1.2.3) collectively. Each attainment distribution examined has raised appreciable concerns about the underlying measure of attainment used in the value-added measures. KS2 and KS4 distributions are more problematic for pupils at either end of the distribution. The KS1 distribution does not suggest that it is produced from a highly robust measure of achievement, as might be understandable for this age group. The comparison across years suggests that the quality of the data is gradually improving.

Figure 6.1.2d – 2012 Pupil KS1 Attainment Distribution



6.1.3 RQ 1.3 - What is the influence of measurement error on value-added scores?

RQ 1.3.1 To what extent does measurement error which is random at pupil-level influence school-level KS2-KS4 value-added estimates?

As discussed at length in Chapter 4, Section 4.3.3, the seriousness of error within value-added scores has been an area in which researchers have fundamental disagreements, especially in relation to whether measurement error can be considered random. Reynolds et al. (2012, p.8) claims that measurement errors tends to be random, will therefore generally cancel out and so error is ‘unlikely to be systematically different in different schools’. On the other hand, Gorard (2010, p.748) holds that errors are likely to be non-random and their effects would ‘propagate’. As was noted, it is difficult to resolve the issue of the extent to which there are biases (non-randomness) in pupil-level measurement errors according to schools. Without further empirical evidence to identify biases in measurement error, more detailed examination of measurement is beyond the purview of this study. What can be examined is how random measurement error ‘plays out’ within value-added calculations. It is possible that different areas of the prior and final attainment distributions are more sensitive to the effects of measurement error. The ceiling effects reported in Section 6.1.2, for example, mean that a small range of pupil scores at KS2 may be required to discriminate between a large range of scores at KS4; it is possible that even random error could have systematic effects on school-level value-added scores if this was the case. This section, therefore, poses the question, "If random errors were introduced into the KS2 and KS4 attainment scores, to what extent would this translate into school-level VA score errors?"

The effect of error would be most clearly seen if there are no other sources of variation (such as school effects). Imagine a scenario in which all school value-added scores are zero and there is a completely deterministic relationship between KS2 and KS4 attainment. In such a scenario, how much (spurious) 'value-added' would the addition of pupil-level random measurement error lead to in the school-level scores? Would errors cancel out, as Reynolds et al. (2012, p.8) claim? This hypothetical question is addressed by completing the following analytical steps: a) a realistic (see below) deterministic relationship between KS2 and KS4 performance is created, this has a school effect of zero by definition; b) several levels of random pupil-level measurement error are added to both the prior and final attainment scores; c) the resulting data are treated as if they were actual pupil prior and final attainment scores and the school-level 'value-added' (which comprised entirely of error) is calculated.

The first step was to create a realistic deterministic KS2-KS4 relationship. To do this, the actual KS2 data and school memberships are used from the 2013 pupil-level KS4 NPD data extract. This means that the prior attainment score used is identical to the actual distribution (this corresponds to the KS2 2008 scores, see Figure 6.1.2c). Then, a deterministic KS4 score was produced by creating a predicted KS4 in a slightly simplified version of the official DfE VA measure. This model was almost identical to the actual model (r = .995) but had the advantage of producing a 1-to-1 correspondence between KS2 and predicted KS4 average point scores (see Appendix C5 for the model specification and further details). It is imagined that the predicted score from this value-added model *is* the actual KS4 of the pupils in question. Treating the predicted scores as if they were actual scores means there is a) a deterministic KS2-KS4 relationship with zero VA by definition and b) the new KS4 distribution (as it is based on model predictions) is highly similar to the actual KS4 distribution. The means of the new and actual distributions were identical to 4 decimal places. As would be expected, the standard deviation of the predicted KS4 scores was smaller at 58 compared to 89.

From this deterministic relationship, random measurement error is added to both KS2 and KS4 scores in order to simulate the effect on final value-added scores at pupil and school-level. Where the school-level scores differ from zero, it can be said that the given level of error at pupil level translates into an error at school level given by the (spurious) value-added scores. To introduce error at KS2 and KS4, 6 normally distributed random variables were created. These error distributions all had a mean of 0. The following error rates are added (given as standard deviations on the error distribution):

Table 6.1.3a – Random error introduced into KS2 and KS4 scores during simulation

	Error to be added at KS2	**Error to be added at KS4**
Small error	Standard deviation = 1 NC point (1/6 of a NC level).	Standard deviation = 6 B8 points (1 GCSE grade across the 8* subjects on the Best 8 Measure)
Medium error	Standard deviation = 2 NC point (1/3 of a NC level).	Standard deviation = 9 B8 points (1.5 GCSE grades across the 8* subjects on the Best 8 Measure)
Large error	Standard deviation = 3 NC point (1/2 of a NC level).	Standard deviation = 12 B8 points (2 GCSE grades across the 8* subjects on the Best 8 Measure)

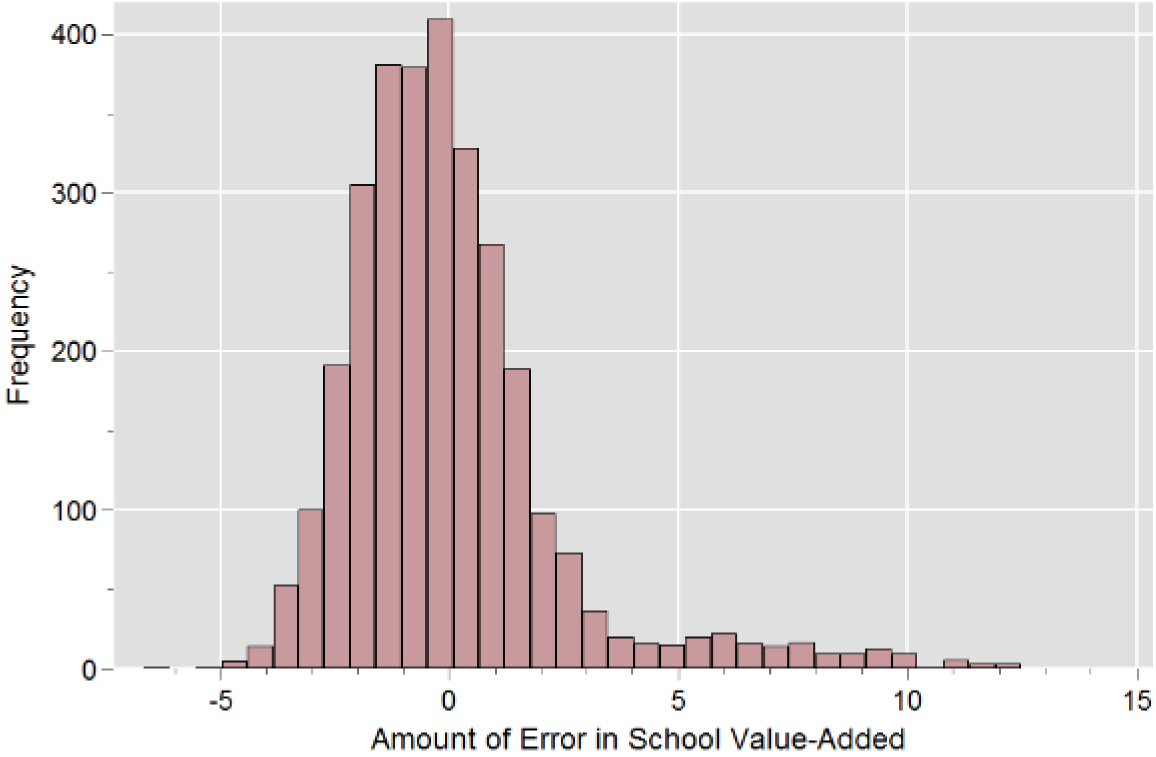
*NB: The Best 8 measure includes a bonus which doubles GCSE English and Maths, the score is therefore the equivalent of 10 GCSEs, with the score difference being applied across all of these.

After introducing these errors, the same VA model was used to estimate the spurious 'value-added scores' for schools. The model outputs for each of the three levels of error are given in Appendix C5. The amount of error translating into school-level value-added from each error level is shown in a series of graphs, below, starting with the small error rate.

The first level of error introduced was a 1 NC point (1/6 of a NC level) standard deviation error at KS2 and a 6 point standard deviation error on the Best 8 point score measure at KS4, corresponding to 1 GCSE grade across the 8 subjects (plus English and Maths bonus).

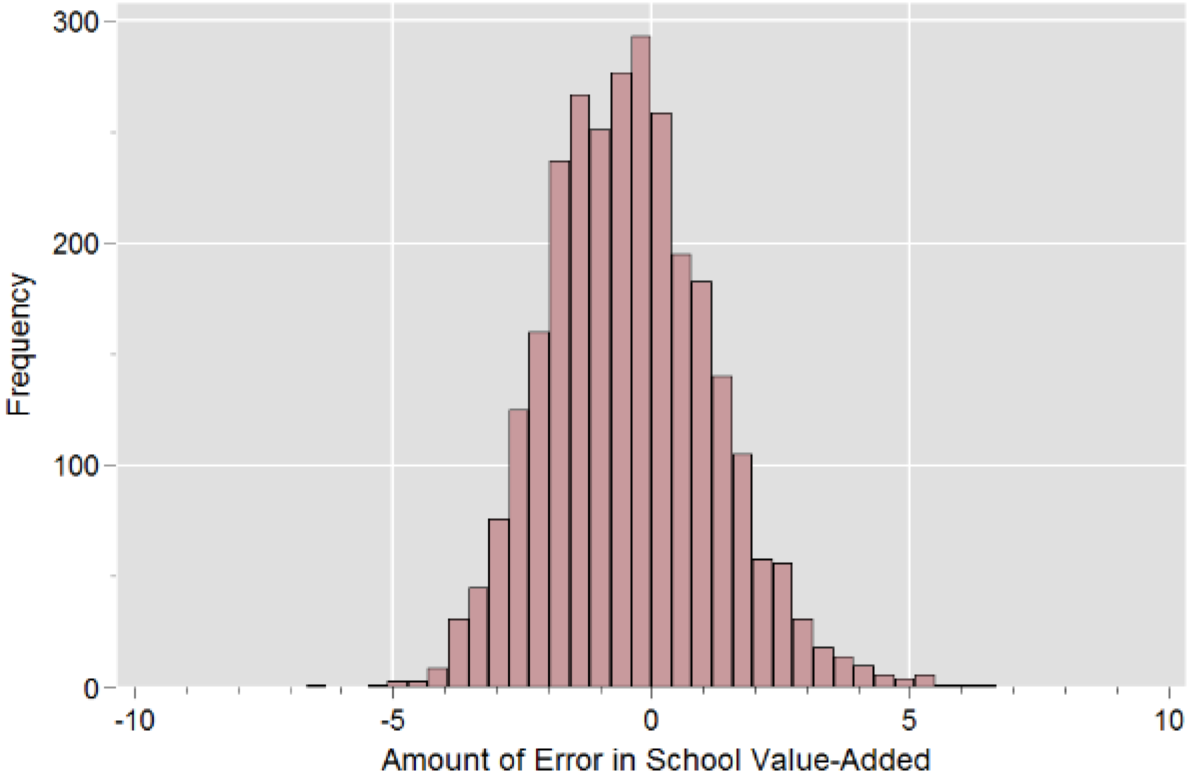
This produced the following distribution of school VA scores, from the zero starting point (see Figure 6.1.3a, below):

Figure 6.1.3a – Estimated Change in School-Level KS2-4 Value-Added Scores after Introducing a Small Pupil-Level Error in the KS2 and KS4 Attainment Scores



The right tail of the distribution stems from the ceiling in the KS2 distribution shown in an earlier section. It is entirely accounted for by 148 (selective) schools with mean intake KS2 scores of 31 and above (level 5C at KS2). When these 148 schools are removed, the distribution is as follows (see Figure 6.1.3b, below):

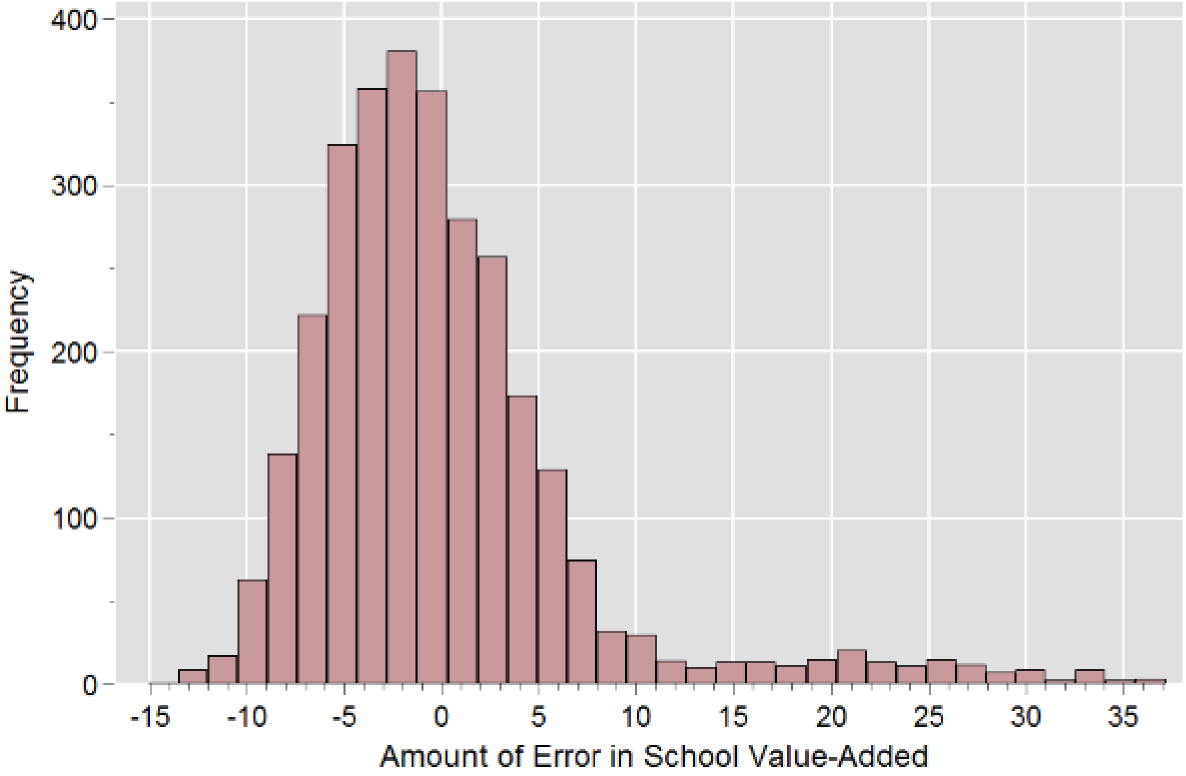
Figure 6.1.3b – Estimated Change in School-Level KS2-4 Value-Added Scores after Introducing a 'Small' Pupil-Level Error in the KS2 and KS4 Attainment Scores excluding Selective Schools



This is a very interesting result given the 'grammar school effect' highlighted in the previous section: The 148 schools with mean intake KS2 scores above 31 points is an aspect of the system – i.e. the presence of grammar schools. But in this simulation, the KS4 scores have been replaced with a *deterministic* score, so with no value-added. Yet, despite this, with the introduction of random error, selective schools (identified here as schools with a mean KS2 score above or equal to 31) still emerge as having a disproportionate number of pupils above expectations (i.e. with positive VA). The mean error at school level for these 148 schools with the introduction of a small amount of pupil-level error is 7.1.

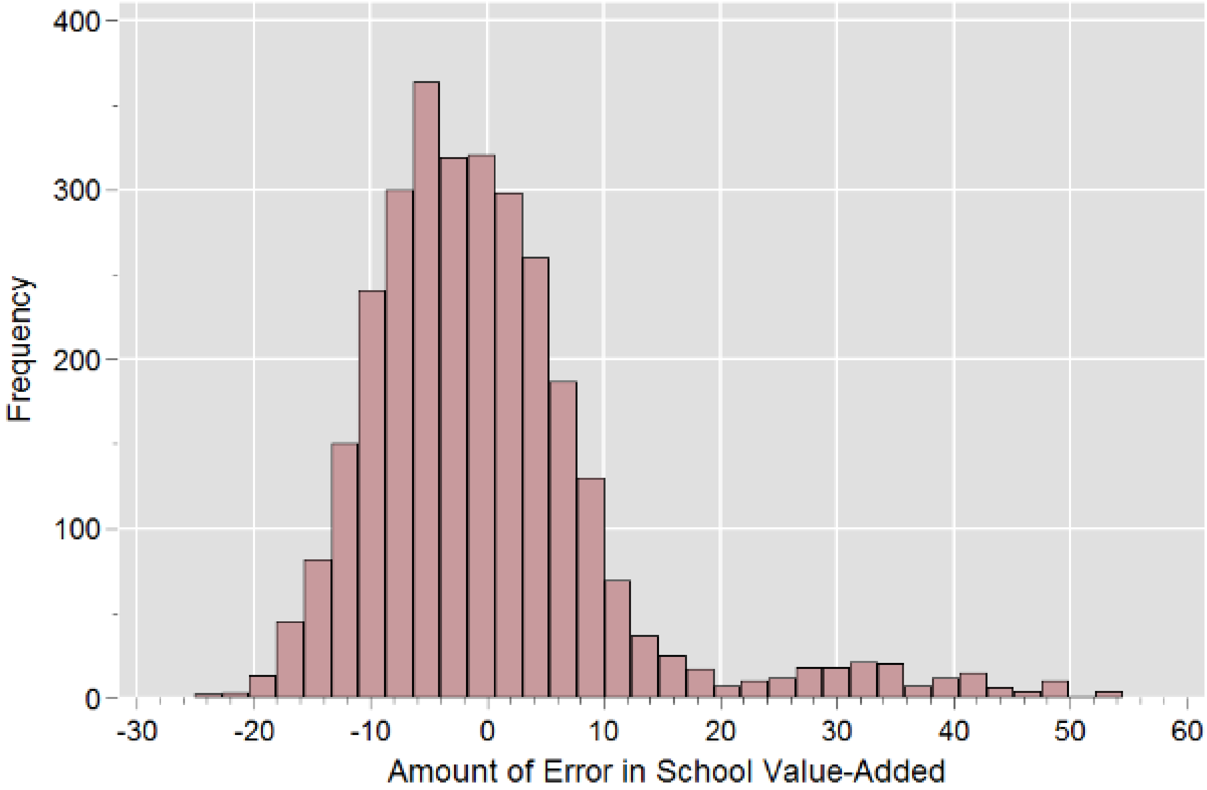
The next analysis is identical to the last but with the introduction of what are called 'medium' sized errors at KS2 and KS4. At KS2, an error distribution with standard deviation of 2 NC points (1/3 of a level, or 1 sub-level) is added to the KS2 scores. At KS4, an error distribution with standard deviation of 9 points (1.5 GCSE grades across the best 8 subjects) is introduced. The result is shown in Figure 6.1.3c, below. Again, a grammar school tail is created at the right of the distribution (mean score for the 148 schools marked as selective is 22.5, about the size of the actual grammar school 'effect'). The graph with these schools removed is not shown but, as above, is approximately symmetrical about 0 and ranges from -15 to about 15. The school-level scores are mostly between 12 points above and below zero, corresponding to an error of 2 grades per pupil at GCSE.

Figure 6.1.3c – Estimated Change in School-Level KS2-4 Value-Added Scores after Introducing a 'Medium' Pupil-Level Error in the KS2 and KS4 Attainment Scores



Finally, a large error is added to the KS2 and KS4 scores. At KS2, an error distribution with standard deviation of 3 NC points (1/2 of a level) is added to the KS2 scores. At KS4, an error distribution with standard deviation of 12 points (2 GCSE grades across the best 8 subjects) is introduced. The mean score for the 148 schools marked as selective is 34 points. This is shown in Figure 6.1.3d, below. The graph with these schools removed is not shown but is approximately symmetrical about 0 and ranges from -20 to about 20. The school-level scores are mostly between 18 points above and below zero, corresponding to an error of 3 grades per pupil at GCSE.

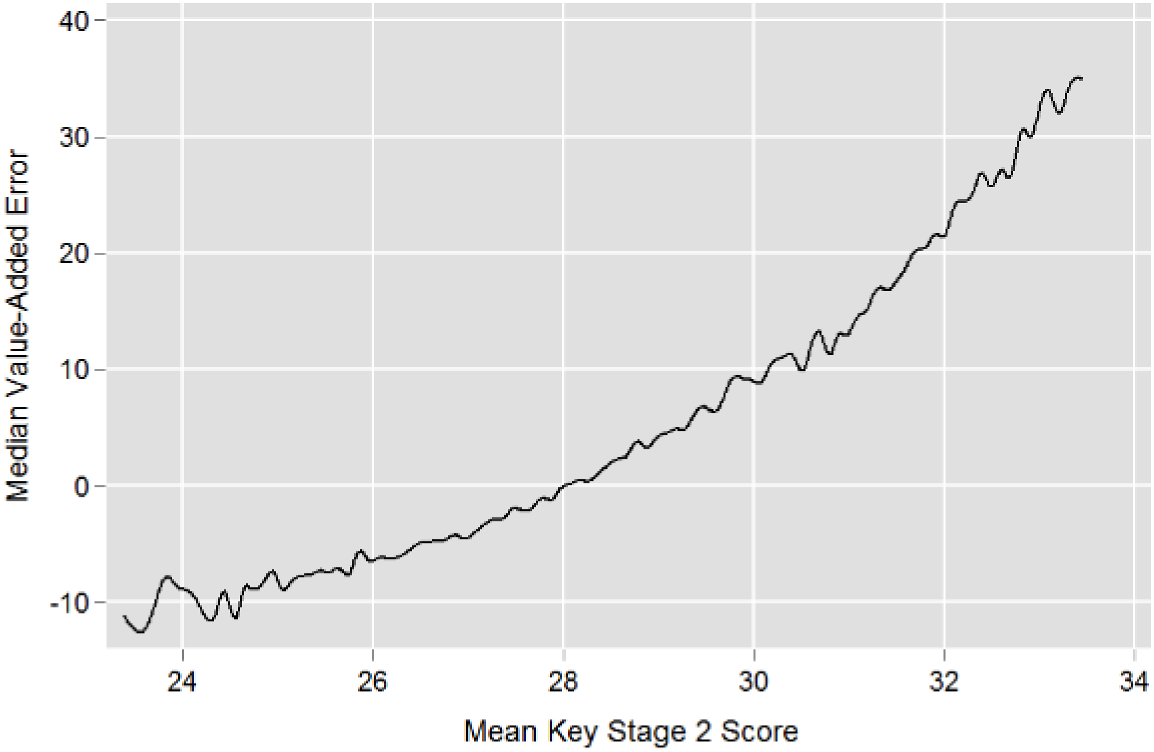
Figure 6.1.3d – Estimated Change in School-Level KS2-4 Value-Added Scores after Introducing a 'Large' Pupil-Level Error in the KS2 and KS4 Attainment Scores



These results have used three error levels to show the magnitude of error which can translate from pupil-level error to school-level scores at each. These errors can be put in context of the overall Best 8 VA school-level distribution which spans from around -150 to 150. This compares to the school-level error distributions which were approximately 5, 12 and 20 points above and below zero (see Figures 6.1.3a, 6.1.3c and 6.1.3d). The designation of these errors rates as small, medium and large is based on this researcher's expectations and could be contested. Despite this possible contention, these results suggest appreciable error rates can translate from pupil-level scores to school-level scores. This is at odds with the prevailing view that random pupil-level errors will cancel out at school-level (Reynolds et al., 2012).

It has also been found that aspects of the distribution and schooling system have caused random error to translate into disproportionate erroneous effects for selective schools. Further investigation of this tendency revealed that the median error in the school results had a clear near-linear trend with average prior attainment scores. This median school-level error by mean Key Stage 2 score is plotted in Figure 6.1.3e, below (for the medium error level):

Figure 6.1.3e – Median Error in School-Level Value-Added by Average Key Stage 2 Score



This figure suggests that the effect of random error is likely to differ due to structural features of the fitted relationship between prior and final attainment. More specifically, when a pupil has a KS2 score of about 28 (where the mean KS2 score is 27.9), the positive and negative effect of a given error size on the pupil's final score have an expected value of zero. When the pupil scores above average, there is an asymmetry between the effect of a positive and negative error such that the expected value becomes positive. When a pupil is below the mean KS2, the opposite effect is produced.

There are two main explanations for this effect, both of which appear to be present in these data. First, note that value-added estimates involve the mapping of the prior attainment scale on to the final attainment scale (where a given KS2 score corresponds to a given KS4 score). When considered alongside each other, it is not necessarily the case that units on one scale will be uniformly spaced compared to the other (consider putting a linear and logarithmic scale alongside each other, for example). When this is the case, an error on one scale will produce a different size shift on the other scale according to whether it is positive or negative. The upward trend in Figure 6.1.3e could be produced from mapping a prior attainment distribution with a lower level of kurtosis to a final attainment distribution with a higher rate of kurtosis (the latter scale would be 'squashed' in from both sides, accentuating errors in the direction of the centre of the distribution relative to the other scale). This is a possible explanation in this case by the fact that the KS2 (prior) attainment distribution has a kurtosis value of 4.5, relative to the Best 8 KS4 (final) attainment distribution of 10.4. The second explanation for the asymmetry in random error effects according to prior attainment relates to the predictive power of the value-added model. As errors were introduced into the scores, the total variance explained in the respective value-added models dropped, as would be expected (see Appendix C5). The relevance of this for this problem is that as the value-added model becomes less able to match prior to final attainment scores, predicted values in the model will tend towards the centre, where there are greater pupil frequencies. By way of explanation, in a deterministic model, a final attainment score uniquely identifies a prior attainment score. When there is error, a final attainment score corresponds to a range of prior attainment scores. This range will not have a uniform frequency and so the expected value will tend to the centre of the distribution (where there are greater pupil numbers). Again, this explanation is reflected in the data: pupil-level predictions for each of the error rates have identical means but the small, medium, and large errors have standard deviations of 55.6, 50.7 and 45.1, respectively. Put simply, when there is uncertainty, the value-added predictions err towards the mean final attainment and so more extreme results are more 'surprising' and result in greater levels of higher and negative value-added.

One final point to note about this structural error pattern is that its effect was highly consistent when analyses were conducted using a new, re-drawn random error for each level. The correlation between school-level scores for two trials of the medium error condition error was 0.93, for example. This means that the same schools can expect to be (dis)advantaged under these conditions based on the characteristics of their intake rather than this being largely driven by chance.

RQ 1.3.2 To what extent does measurement error which is random at pupil-level influence school-level KS1-KS2 value-added estimates?

This analysis follows an identical pattern of analysis as the previous analysis but concerns the KS1-2 VA scores (see Appendix C6 for model output). The specific error distributions added to the KS1 and KS2 scores are given below. As both KS1 and KS2 scores use a common scale, the same error levels were used for each. It might be that KS1 assessment, which was teacher assessed for the data used and concerns younger children, may be less reliable (and see Figure

6.1.2d in Section 6.1.2, above); but without concrete data on what the actual error rates are, the likely error rate is left the same as at KS2.

Table 6.1.3b – Random error introduced into KS1 and KS2 scores during simulation

	Error to be added at KS1	**Error to be added at KS2**
Small error	Standard deviation = 1 NC point (1/6 of a NC level).	Standard deviation = 1 NC point (1/6 of a NC level).
Medium error	Standard deviation = 2 NC point (1/3 of a NC level).	Standard deviation = 2 NC point (1/3 of a NC level).
Large error	Standard deviation = 3 NC point (1/2 of a NC level).	Standard deviation = 3 NC point (1/2 of a NC level).

The following histograms, Figure 6.1.3f, 6.1.3g and 6.1.3h plot the school value-added scores after introducing small, medium and large errors, respectively:

Figure 6.1.3f – Estimated Change in School-Level KS1-2 Value-Added Scores after Introducing a 'Small' Pupil-Level Error in the KS1 and KS2 Attainment Scores

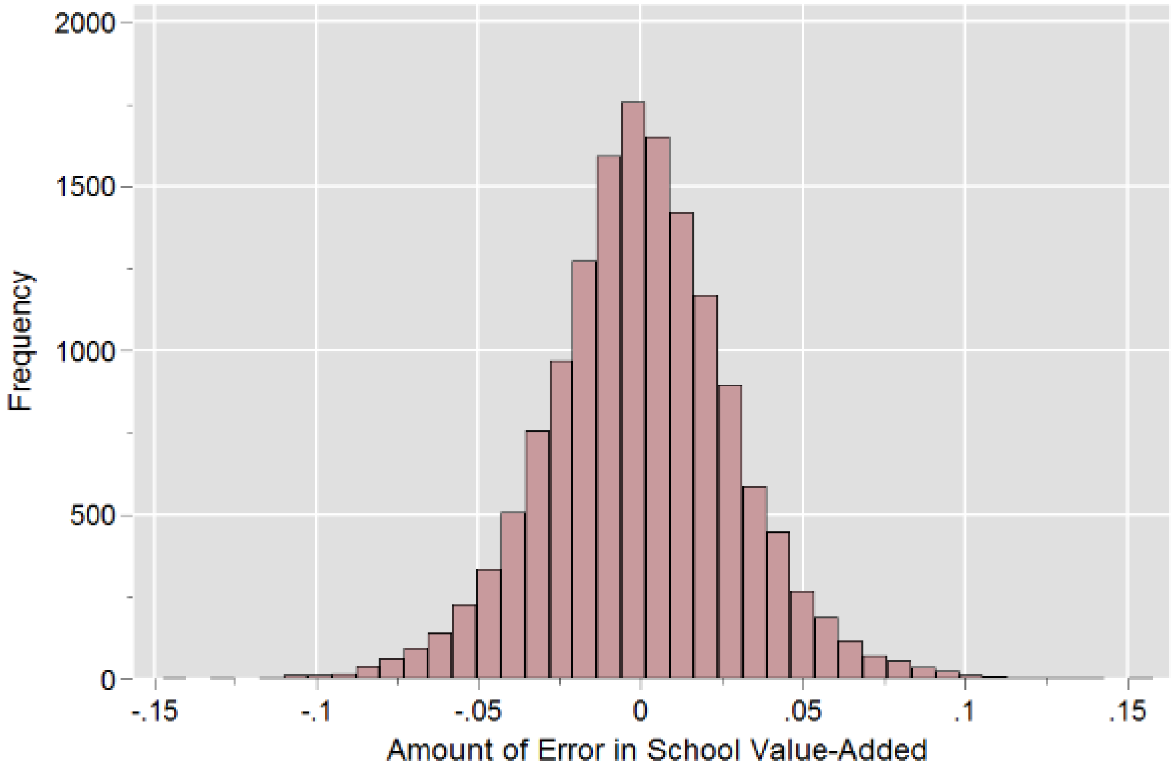


Figure 6.1.3g – Estimated Change in School-Level KS1-2 Value-Added Scores after Introducing a 'Medium' Pupil-Level Error in the KS1 and KS2 Attainment Scores

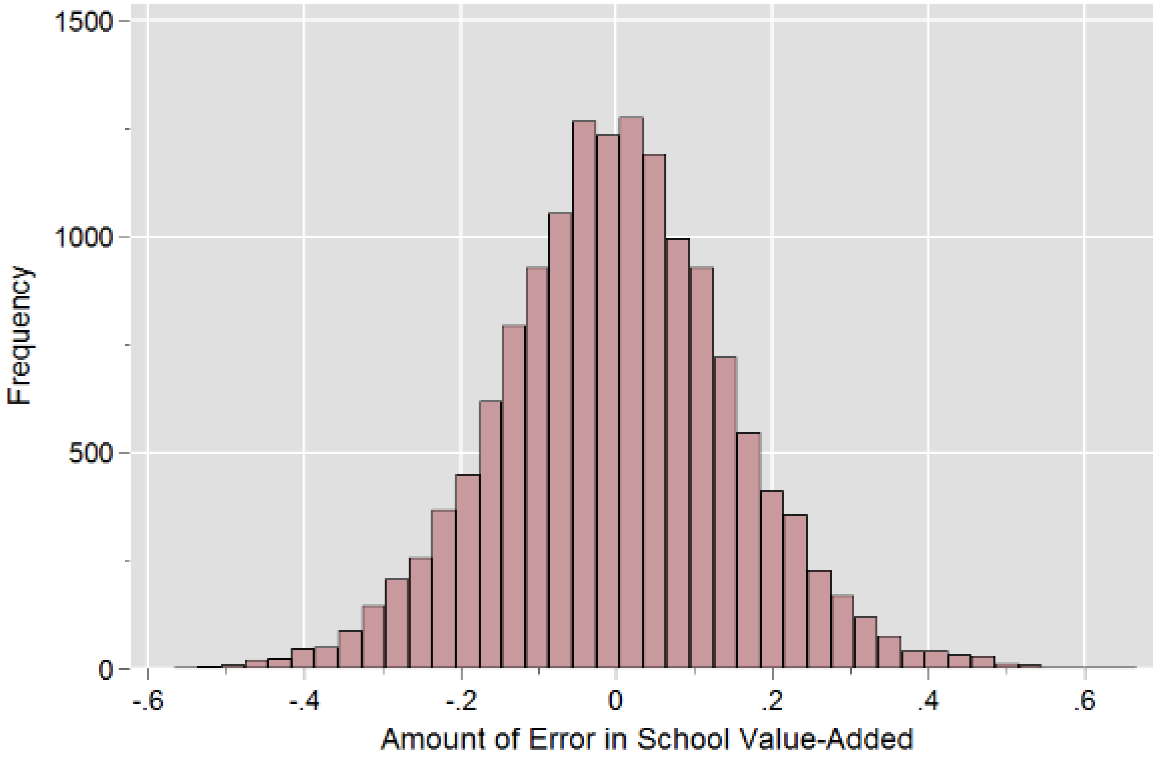
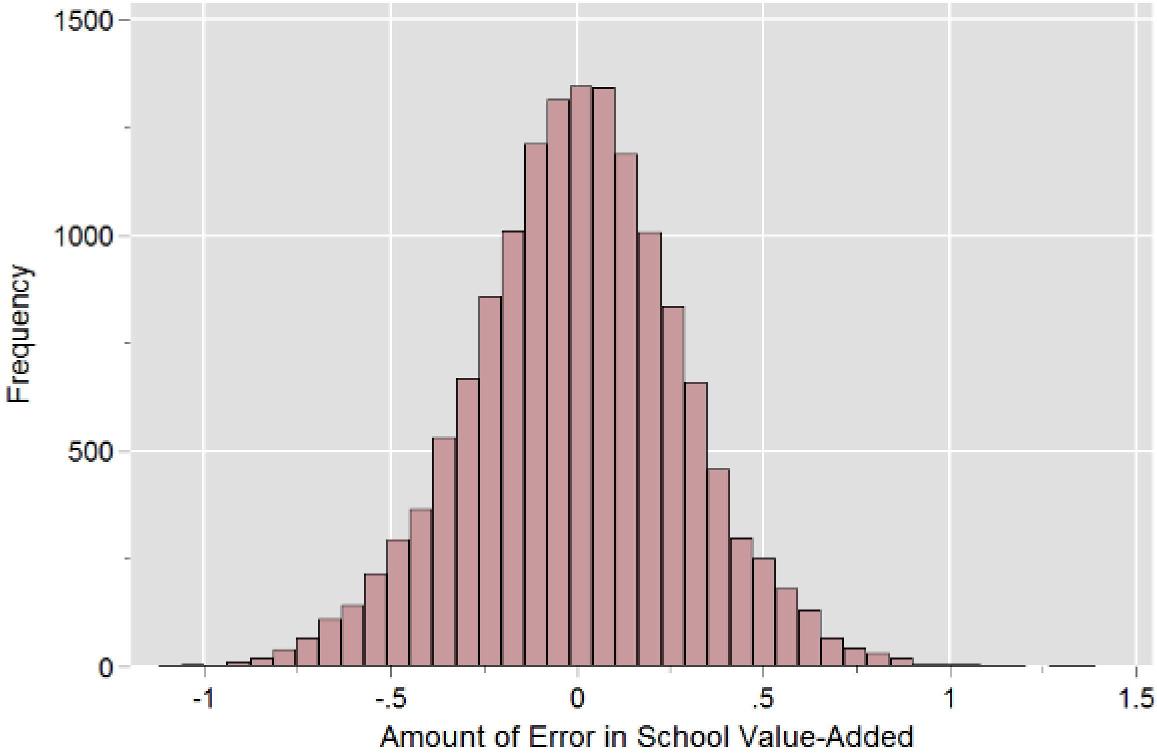


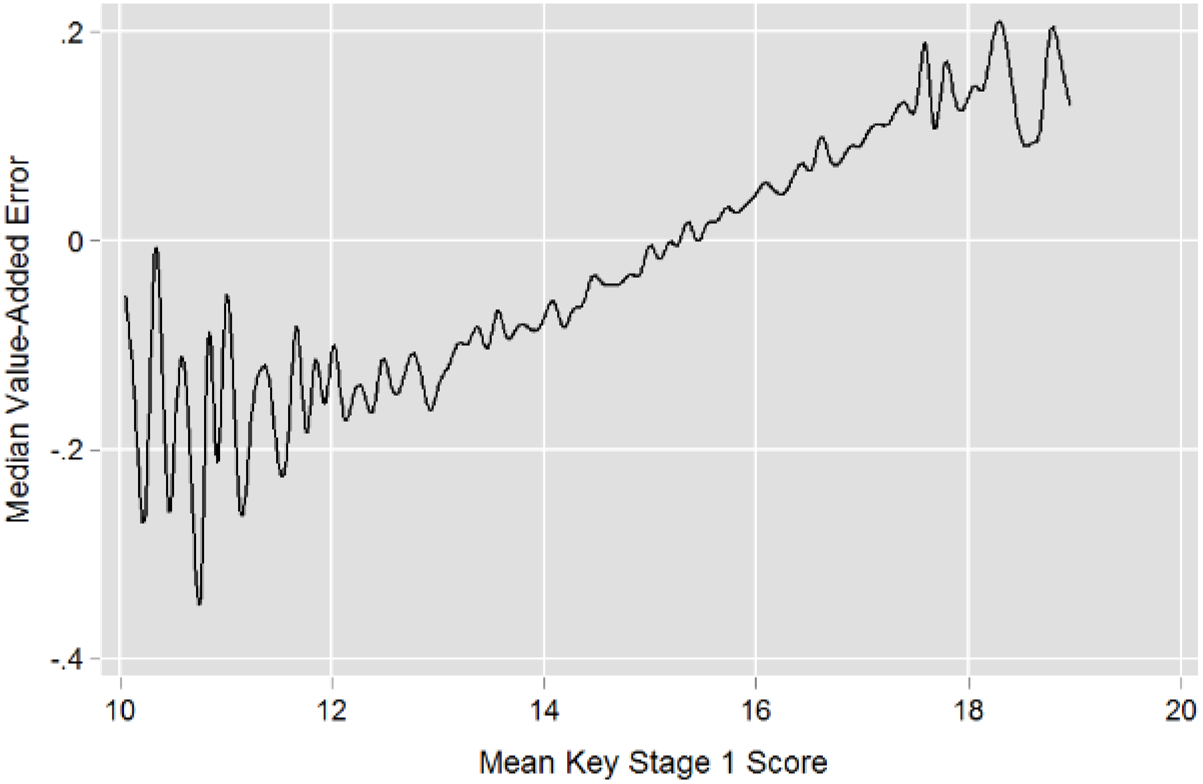
Figure 6.1.3h – Estimated Change in School-Level KS1-2 Value-Added Scores after Introducing a 'Large' Pupil-Level Error in the KS1 and KS2 Attainment Scores



These figures show smaller rates of error than the KS2-4 simulated results and no indications of a large structural difference, as discussed in the last section. The small error rate generates a school-level error histogram (Figure 6.1.3f) which spans the values of about -0.1 to 0.1. An error of 0.1 of a NC point equates to less than ½ a month's progress per pupil. For the medium error rate (see Figure 6.1.3g), the distribution spans from about -0.5 to 0.5. Half a NC point represents about 2 months' of progress per pupil at the school. Much of the distribution lies within ¼ of a NC point, representing nearer 1 month of progress per pupil. The final analysis introduced a 'large' error rate yielding school-level errors spanning from approximately -1 to 1 NC points, representing about 4 months' progress per pupil at the school (see Figure 6.1.3h). It is valuable to compare these error rates to the scale of the official KS1-2 VA where scores span from approximately -10 to 8. The substantive significance of these for schools is discussed in the next chapter.

As with the KS2-4 simulation, it was examined whether there was a structural pattern within the errors. A similar positive relationship was found, although it was relatively much smaller. For the medium error level the structural bias ranged from -0.2 to 0.2, as shown in Figure 6.1.3i, below:

Figure 6.1.3i – Median Error in School-Level Value-Added by Average Key Stage 2 Score



6.2 Study 2 - Inter-Method Reliability

6.2.1 How similar are estimates of effectiveness produced by value-added (VA), cross-sectional regression discontinuity (RD) and longitudinal regression discontinuity (LRD) designs?

This study involves comparing a CVA measure with three variations on a regression discontinuity (RD) design: a) a basic RD design using cross-sectional data (RD1), b) a RD design using cross-sectional data with added contextual factors (as 'interaction' effects) (RD2), and c) a RD design using longitudinal data (LRD). Only one previous study has compared cross-sectional (RD1) and longitudinal (LRD) applications of the RD design and none have compared VA with either of these, although one has compared RD with a statistically adjusted predication not using prior attainment scores (Cahan and Elbaz, 2000, also see Chapter 4, Section 4.4).

As described in the methods chapter (Section 5.5.1) the main intention of this study is to present results pertaining to the level of agreement between the CVA measure and the other four measures and to isolate the key differences which lead to any discrepancies with a view to gaining insight into the validity of the VA measure. As was also noted in the methods chapter, the three initial research questions concern the RD designs only. The RD design is a new innovation in educational effectiveness research and, although initial results are promising, its properties require further research. These initial questions are posed in order to first understand the properties of these RD designs.

The results presented below look both ways in terms of validity, shedding light on the viability of the RD design as an alternative method of estimating school effects as well as providing a source of comparison for VA estimates. Because of their design, there are different threats to validity for each measure of varying seriousness. See Section 4.4.4 for discussion of these differences.

The starting point for this study is to describe and give details of all four measurement designs. Although all measures are not used in all of the three initial research questions and the CVA measure is not used until the fourth and final question, it is clearer to present each measurement design up-front. The measures are introduced in the following order: CVA, basic RD (RD1), RD with contextual interaction effects (RD2) and longitudinal RD (LRD):

Details of Value-Added Measure Used in this Study:

The value-added measure used in this study is a simple contextualised value-added measure. Chapter 4, Section 4.2, described the compromise between controlling for biases and attenuating the school effect. The CVA model used controls for prior attainment, free school meals status (a measure of poverty) and gender only and is specified in Appendix D1. This is thought a good compromise between the risk of non-school factor bias and over-correction of the model. School-level averages were also considered in earlier analyses but found to have a negligible effects on overall results and are not included in what follows. Ten CVA measures were produced, one for each cohort-time combination (i to x) described in Section 5.5.2. The output from the first of these (for cohort-time i) is given in Appendix D1.

In short, the value-added measure estimates cohort performances by comparing each cohort's level of attainment to its attainment in the previous time period, adjusting this for KS1 attainment and the contextual variables.

Details of Regression Discontinuity Measures Used in this Study:

A RD model separates an age effect (pupil maturity over time) from an added-year effect (the school effect) using two or more consecutive school year groups. How this works is best illustrated in the following diagram from Luyten et al. (2009, p.155) (see Figure 6.2.1a, below):

Figure 6.2.1a – How a Regression Discontinuity Design Separated Maturity from Added-Year Effects, Diagram from (Luyten et al., 2009, p.155), Figure 2a

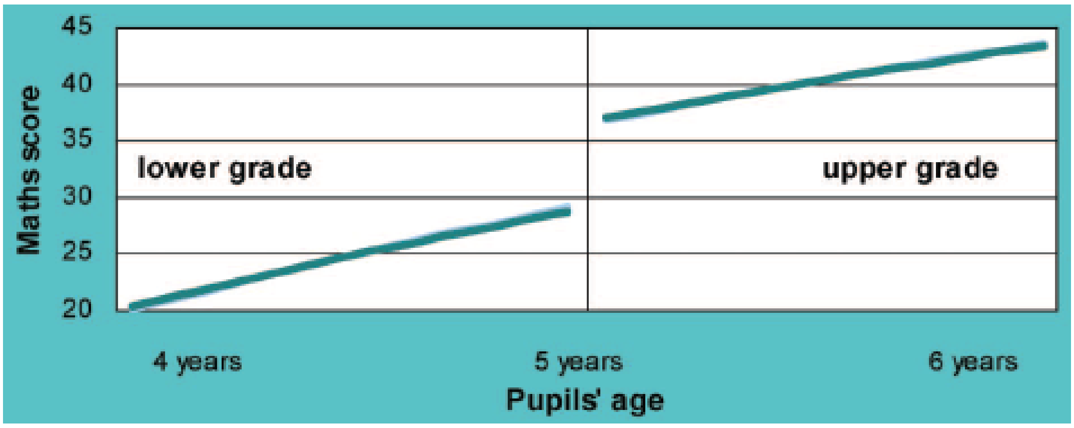


Figure 6.2.1a gives an upward-sloping regression line showing the effect of maturity on performance; there is a 'discontinuity' in this regression line at the administrative cut-off date sorting pupils into upper and lower year groups. Pupils born on the day before and after the administrative cut off are almost identical in age yet one has had an extra year of schooling. The effect of this extra year is estimated by the discontinuity in the regression line at this point.

The first RD design used (RD1) is a basic model which is based on that used in Luyten et al. (2009). This model is specified and example output is given for RD1 in Appendix D2. This model fits a regression line equivalent to that shown in Figure 6.2.1a. The only slight differences are that, first, the age variable which is recorded here in months rather than days as in Luyten et al. (2009); this is unlikely to have an appreciable impact, as discussed in the methods chapter (Section 5.5.2). Second, to avoid having to compute a number of pupil ages relative to the various cut-offs, the age within year was recorded as age within a given year rather than relative to the cut-off separating two years. This does not affect the results, only how the variable coefficients have had to be interpreted (see Appendix D2 for further details). The cross-sectional RD design essentially uses the lower year group of two consecutive cohorts in the same school as a baseline to judge the progress made by the upper cohort. It assumes, therefore, that the lower year group – being from the same school – is similar to the upper year group other than in age.

The second RD design (RD2) builds on the basic design by adding contextual variables (as in a CVA model) and terms estimating the interaction between the added year effect and the contextual variable in question (as described in Luyten, 2006). The extended RD model with contextual factors and interaction terms (RD2) can examine whether the added year effect is bigger or smaller for different groups and the association between performance and contextual factors. Several RD2 models were examined; the results of these analyses are summarised in RQ 2.1.2, below. The final model used in the comparisons at the end of this study (RQ 2.1.4) is specified in Appendix D3 and output from a selected model (i) is given. This model includes gender and free school meals (FSM) eligibility (a measure of poverty) as contextual variables and an interaction term which estimates the interaction between FSM and the added year effect. In a second step, the resulting school RD measures were then adjusted by intake mean prior attainment to ensure this model had no systematic bias (see below for further details). As with RD1, RD2 is cross-sectional and relies on the lower year group being a good comparator.

The third and final RD design is the longitudinal regression discontinuity design (LRD). This model is simply a gain score across two years (e.g. Maths Score T2 – Maths Score T1) with the age effect estimated using a RD design (as in RD1) subtracted from the overall gain. This replicates the longitudinal estimate used in Luyten et al. (2009, p.148). By using longitudinal data, yet applying the regression discontinuity design, one gets a measure of performance adjusted by maturity. Assumptions about the comparability of the lower year group (see RD1 and RD2, above) are not required as the actual cohort's baseline performance data from the previous year is used to judge progress. If performance across different consecutive cohorts in a school is inconsistent, this will result in a difference between the RD1/2 and LRD measures. While this loses the cross-sectional advantages of RD, it ensures that the added-year effect can be attributed to improvement in the cohort's performance rather than potential variability across cohorts and allows this assumption of comparability to be tested. With all measures now described, we turn to consider the results for each research question.

RQ 2.1.1 What is the effect of 1 extra year of schooling on achievement and what proportion of this is accounted for by schooling?

The first analysis in this study estimated the absolute effect of an extra year of schooling using a RD design. This was calculated in several ways, which were then compared: first, within the RD1 model looking across two consecutive year groups at a time (see Appendix D2); second, using the multiple cut-off design used in Kyriakides and Luyten (2009) and, third, using a series of linear regressions of National Curriculum (NC) level on age-within-year. All gave highly similar results. The similarity of these models is to be expected given that they are only minor variations of functional form fitted to the same data. Minor differences stemmed from the precise samples used to estimate each effect. The RD1 design, for example, calculated the age effect across two consecutive years, the linear regression calculated it for a single NC year and the multiple cut-off design made the estimate depend on the precise functional form (linear or non-linear). The functional form of the age effect was examined in the multiple cut-off design, where the age effect across the age range studied was found to be approximately linear. The results of the linear regression model are presented in Table 6.2.1a:

Table 6.2.1a – Added year effects by national curriculum year

National Curriculum Year	Constant	Age effect	Annual Progress	Constant	Age effect	Annual Progress	Constant	Age effect	Annual Progress
	2007/2008			*2008/2009*			*2009/2010*		
3	17.3	0.16	-	17.3	0.16	-	17.4	0.15	-
4	19.8	0.17	2.50	20.0	0.17	2.71	20.2	0.17	2.76
5	22.8	0.17	3.08	22.8	0.18	2.79	23.1	0.19	2.88
6	26.4	0.14	3.55	26.6	0.15	3.82	26.9	0.16	3.82
7	28.9	0.17	2.53	29.5	0.14	2.89	29.5	0.19	2.60
8	31.6	0.14	2.65	32.1	0.18	2.66	32.4	0.15	2.94
9	34.5	0.18	2.90	35.4	0.14	3.22	35.2	0.18	2.75

Table 6.2.1a can be used to calculate the mean NC level for the average pupil. This can be found by taking pupils' relative age within the year (where August=0, July=1… September=11), multiplying this by the age effect and subtracting the total of these from the annual progress made in the given year. Note that the annual progress is simply the difference between the constants in each NC year. As well as showing the utility of the regression discontinuity design to calculate (overall) absolute school effects, the results are interesting in their own right: The average progress made in a year of schooling across the whole sample was approximately 3 NC points. This is in line with the design of the NC where 3 points are expected per year. Note, however, that year 6 – a year when national examinations are taken – greatly exceeds this and other NC years tend to be slightly lower.

The overall school effect can be expressed as a percentage of the total progress: The mean annual rate of progress since the previous year is listed for each year. The overall mean age effect (per month) is approximately 0.17. This means that each year will see pupils making 2.04 points (12 x 0.17) of progress due to maturity alone. In this sense then, 2.04/3.00 (68%) of the observed improvement from year-to-year by pupils across this sample is due to pupil maturity and 32% is attributed to the general effect of schooling. This is slightly lower than the estimate for England found in Luyten (2006) of 38% and much lower than his figures for other countries studied (55-75%). Looking at English reception (age 4-5) classes, Luyten et al. (2009) found around 50% of the effect was attributable to the school.

RQ 2.1.2 Do RD school effects differ according to ability or other contextual factors?

One concern with the RD1 design which may cause a difference between RD1 results and the school CVA measure is that the RD1 measure estimates progress but does not adjust this according to context or prior attainment. While still an accurate measure of absolute effect size, comparisons across schools' relative scores will not be like-for-like and comparison with value-added estimates would be problematic. To take this possibility into account, further RD models were produced to adjust for any interactions between the added year effect and the average pupil attainment for the cohort in question or any other contextual factors. These would test whether the size of the added-year effect systematically varied according to pupil characteristics. Three contextual variables were examined: gender, free school meals (FSM) eligibility (a measure of poverty) and mean cohort prior attainment, as measured by the cohorts' key stage 1 (age 7) national examination scores.

Entering prior ability into the model proved problematic as the strength of prior attainment variables as predictors resulted in value-added-like models being produced when specifying prior attainment as per the other interaction variables. The intention was to keep the model analogous to the RD1 but ensure that there were no systematic biases which would reduce the value of the comparison with the CVA measure. To achieve this, a two-step procedure was followed where, first, adjusted measures were produced from the original RD1 measure by adding contextual and interaction terms as described in Luyten (2006). This step added the gender and free school meals variables, as was done in the CVA measure. As in Luyten (2006), the main effects of these variables as well as their interaction with the added year were examined. The specification and full results of this first step for all measures (i to x) are given in Appendix D3 and are summarised below. The cohort added-year effects which were produced using this design were then adjusted using a linear regression of the added-year effect on cohort-level mean KS1 attainment. The residual from this model was used as the final RD2 measure (used during comparisons in RQ2.1.4).

The results of the first step of the analysis can be summarised as follows: A pupil's gender being male predicted lower mathematics scores by between about a fifth and two fifths of a NC point but the interaction effect with this and the added year effect was inconsistent. FSM status predicted between about 1.5 and 2.5 NC levels lower attainment (or about a year's progress) as a main effect. It also had a substantial interaction effect of about 0.2-0.4 NC points, or about 1-2 months' lower progress (per year) than pupils not eligible for free school meals. This suggests that not only are pupils who are eligible for FSM about a year behind their peers on average, they also fall further behind each year. The interaction effect between FSM and progress was fairly consistent across all measures with the exception of the measure concerning progress from year 5 to 6, where FSM had a smaller and more inconsistent effect. This is most likely due to the influence of the key stage 2 national examinations in this year.

The second step of the analysis, examining the relationship between the adjusted RD estimate (Step 1) and cohort mean performance gave inconsistent results. It might be that controlling for prior attainment in a second step resulted in the contextual variables from the first step acting as a proxy for prior attainment due to multicollinearity. This may have inflated the estimates from the first step as well as causing inconsistent results in the second. This should be kept in mind when interpreting the associations between the contextual factors and attainment or progress. As the intention is to create an unbiased measure which can be compared with a CVA measure, however, this is not held to be especially problematic.

RQ 2.1.3 To what extent does the effect of 1 extra year of schooling vary between schools?

To address this point, overall school effects across the whole sample have been examined. We now look at the relative school effect. As noted in Appendix D3, the RD designs are calculated within a multi-level model and one of the outputs are school-specific deviations from the overall school effect (i.e. relative school effects). Similarly, the LRD estimates can be mean-centred to clearly compare the size of differences in rates of progress by school. Correlations between all four measures are compared in the final research question (RQ 2.1.4). Before this, the distributions of each measure are examined. Table 6.2.1b, below, gives summary statistics for the distribution created using each measurement, including the range and standard deviation.

Table 6.2.1b – Estimated school effects on attainment for each measurement design

Measure	Obs	NC Year	Std. Dev.	Min	Max	Measure	Obs	NC Year	Std. Dev.	Min	Max
LRD_i	271	4	0.9	-3.6	3.3	CVA_i	271	4	0.7	-2.2	2.2
LRD_ii	271	5	0.9	-3.3	3.2	CVA_ii	271	5	0.7	-2.2	2.6
LRD_iii	260	6	1.1	-3.4	3.2	CVA_iii	260	6	0.9	-2.9	3.1
LRD_iv	69	8	1.5	-3.8	4.7	CVA_iv	69	8	1.3	-3.2	3.0
LRD_v	68	9	1.4	-3.4	3.3	CVA_v	68	9	1.2	-3.3	2.8
LRD_vi	225	4	0.8	-2.2	3.1	CVA_vi	225	4	0.7	-1.9	2.7
LRD_vii	226	5	0.9	-3.6	2.3	CVA_vii	226	5	0.8	-2.8	1.9
LRD_viii	212	6	1.1	-3.0	3.4	CVA_viii	212	6	0.9	-2.7	2.4
LRD_ix	52	8	1.2	-3.8	2.5	CVA_ix	52	8	1.0	-3.3	2.1
LRD_x	49	9	1.3	-2.7	4.6	CVA_x	49	9	1.2	-2.8	4.4
RD1_i	271	4	0.4	-1.3	1.0	RD2_i	271	4	0.3	-1.3	1.1
RD1_ii	271	5	0.7	-1.9	1.7	RD2_ii	271	5	0.6	-1.6	2.0
RD1_iii	260	6	0.6	-1.6	2.4	RD2_iii	260	6	0.6	-1.6	2.5
RD1_iv	69	8	0.8	-1.9	2.2	RD2_iv	69	8	0.7	-1.9	1.7
RD1_v	68	9	0.9	-3.3	1.9	RD2_v	68	9	0.8	-3.5	1.8
RD1_vi	225	4	0.5	-1.4	1.3	RD2_vi	225	4	0.5	-1.2	1.0
RD1_vii	226	5	0.4	-1.2	1.3	RD2_vii	226	5	0.4	-1.3	1.1
RD1_viii	212	6	0.5	-1.4	1.5	RD2_viii	212	6	0.5	-1.4	1.7
RD1_ix	52	8	0.8	-1.6	2.3	RD2_ix	52	8	0.6	-1.3	1.9
RD1_x	50	9	1.1	-3.5	3.9	RD2_x	50	9	1.1	-3.4	4.4

These results show large differences in the average annual rate of progress for each cohort. These are most clearly seen in the LRD. There were large differences in the school average rates of progress compared to the expected progress due to age of about 1-2 NC points (net of maturity effect). Some cohorts made as much as a year's more or less progress than the expected rate. The adjusted models produced estimates less extreme than this with the RD giving the least extreme differences between rates of progress between schools.

RQ 2.1.4 How similar are estimates of effectiveness produced by value-added (VA), cross-sectional regression discontinuity (RD) and longitudinal regression discontinuity (LRD) designs?

The key intention of this study is to compare value-added measures with other measures of the school effect. Table 6.2.1c, below, shows the correlation between the contextualised value added measure for each study year and national curriculum year combination and the 3 measures created using variations on a regression discontinuity design.

Table 6.2.1c – The correlation between the value-added measure and four regression discontinuity measures for the corresponding study/national curriculum year.

	Study Year	NC Year	RD1	RD2	LRD
CVA_i	2	4	0.41	0.42	0.92
CVA_{vi}	3	4	0.50	0.58	0.94
CVA_{ii}	2	5	0.55	0.69	0.93
CVA_{vii}	3	5	0.50	0.45	0.96
CVA_{iii}	2	6	0.39	0.40	0.95
CVA_{viii}	3	6	0.45	0.51	0.96
CVA_{iv}	2	8	0.56	0.53	0.92
CVA_{ix}	3	8	0.65	0.54	0.93
CVA_v	2	9	0.28	0.28	0.97
CVA_x	3	9	0.64	0.63	0.97

These correlations can be summarised as follows: the CVA and LRD measures have high to very high correlations. The RD designs generally yield moderate correlations with the CVA measure but in some cases correlations are as low as 0.28. Correlations between LRD and RD1 (not shown) range from 0.26 and 0.65 with a mean of 0.46; this correlation between school effects for individual schools is lower than the figure found in Luyten et al. (2009) of 0.71, although this was for English 4 and 5 year-olds in 18 schools. The implications of these results are discussed in Chapter 7.

6.3 Study 3 - Stability over Time

6.3.1 RQ 3.1 - How stable is the current English value-added measure across several years?

This study examines the stability of value-added measures over time. As discussed in Chapter 4, the stability of value-added scores can be viewed as indirect evidence about their validity. Where stability is in line with what would be expected for changes in performance over time, this suggests the measure is valid. If results over time are more volatile than can reasonably be attributed to changes in school performance, this suggests measurement invalidity.

The first question in this study concerns the current official school VA scores in England and so replicates previous studies of the stability of the English CVA scores (Leckie and Goldstein, 2011, Gorard et al., 2012) using the current (2011-2014) VA model. The analysis is also extended to consider the stability of the primary level scores. The results presented in this section are all pairwise correlations between school value-added or attainment scores. All value-added measures refer to the official value-added measure in the given year and the results concern all state-maintained, mainstream schools in England. We start by looking at correlations between school value-added scores and school 'raw' performance scores at primary level, given in table 6.3.1a, below:

Table 6.3.1a – Pairwise correlations over time in primary school value-added and unadjusted attainment measures

Primary Level							
School Value-Added				Unadjusted Performance (APS)			
	1 year earlier	2 years earlier	3 years earlier		1 year earlier	2 years earlier	3 years earlier
2014	0.61	0.46	0.35	2014	0.66	0.61	0.56
2013	0.60	0.45		2013	0.66	0.60	
2012	0.59			2012	0.66		

(school n ranges from 13,473 to 14,454)

At primary level, unadjusted performance correlations over time are moderately stable. The value-added correlation between the current and previous year is roughly similar to the unadjusted correlation. However, the value-added correlations fall quite sharply when comparing value-added scores 2 and 3 years apart. Primary school value-added performance is hardly related to performance 3 years earlier.

The secondary-level results are given in Table 6.3.1b, below:

Table 6.3.1b – Pairwise correlations over time in secondary school value-added and unadjusted attainment measures

Secondary Level							
School (Best 8) Value-Added				**Total Average Capped point score**			
	1 year earlier	2 years earlier	3 years earlier		1 year earlier	2 years earlier	3 years earlier
2014	0.56	0.49	0.45	2014	0.79	0.78	0.78
2013	0.79	0.68		2013	0.90	0.86	
2012	0.79			2012	0.91		

(school n ranges from 2792 to 3076)

These data show a discontinuity in the school scores between 2013 and 2014 in both measures where the 2014 measures show considerably lower associations between earlier measures. This is most likely due to government reforms to which qualifications are counted as equivalents within the total capped point scores (and so the Best 8 VA measure). The stability of like-for-like raw attainment scores is more clearly seen by examining the average capped point scores for GCSEs only (i.e. without GCSE equivalents being taken into account in the measure). This is shown in Table 6.3.1c, below:

Table 6.3.1c – Pairwise correlations over time in secondary school value-added and unadjusted attainment measures

Secondary Level							
School* (Best 8) Value-Added				**Total Average Capped point score (GCSEs only)**			
	1 year earlier	2 years earlier	3 years earlier		1 year earlier	2 years earlier	3 years earlier
2014	0.56	0.49	0.45	2014	0.94	0.92	0.90
2013	0.79	0.68		2013	0.96	0.94	
2012	0.79			2012	0.96		

*(n ranges from 2792 to 3076)

The Best 8 VA measure uses the total capped point score and so there is some disconnect as what is counted as a GCSE equivalent has changed. It is likely therefore that the 2012-2013 results are more representative of typical conditions. The results show that, as with the primary data, the unadjusted secondary correlations are more stable over time than the value-added correlations. It is valuable to compare these VA correlations with the correlations found in previous research concerning the former CVA measure. The correlations presented here for the VA measure on the high end of the correlations found by Gorard et al. (2012) for the CVA measure of between 0.58-0.79 and 0.48-0.67 for 1 and 2 years apart, respectively. This suggests that the stability of secondary value-added is higher using a VA model than with the CVA model. The correlations are only moderate however and likely to be largely due to the reintroduction of the observable biases revealed in study 1. Relatively more stable contextual influences remaining in the data will have shifted the unstable CVA scores towards the more stable but more biased raw scores. Moreover, as discussed in Chapter 4, there is considerable disagreement over whether correlations of around 0.6-0.8 can be considered acceptable in this context. This is discussed further in the discussion chapter (Chapter 7).

6.3.2 RQ 3.2 - Is the rate of stability in value-added scores associated with school performance?

It is difficult to know to what extent the instability seen in the results above can be attributed to changes in school performance. This analysis attempts to get some indication of what is causing the observed changes by considering the extent to which stability differs in schools in different circumstances. The results presented below concern whether stability differs according to the prior performance of schools. The reasoning is as follows: It is quite likely that low performing schools will go to great lengths to improve performance. This will result in considerable changes in effectiveness over a short period of time. Conversely, high performing schools are more likely to seek to replicate previous performances and this may be reflected in more stable scores. This conjecture suggests there is value in examining stability according to school performance to examine whether there are structural differences in the stability of school performances. If levels of stability are consistent across the performance range, this suggests that instability is something linked to the measure itself. If levels of stability systematically differ at different levels of performance, although this does not rule out this being a problem of measurement, it is certainly suggestive that changes in school performance are related to changes brought about by structural aspects of the school system (i.e. there being considerable pressure for low performing schools to improve performance relative to higher performing schools).

To examine whether stability relates to original school performance in this way, schools are sorted into quintiles (i.e. 5 groups of equal size) based on their 2011 VA performance. The stability analysis, as above, is then repeated for these groups. The estimates are presented in Table 6.3.2a, below:

Table 6.3.2a – Pairwise correlations over time in primary school value-added and unadjusted attainment measures by performance quintile

Primary Level								
		School Value-Added			**Unadjusted Performance (APS)**			
		1 year earlier	2 years earlier	3 years earlier		1 year earlier	2 years earlier	3 years earlier
All Schools	2014	0.61	0.46	0.35	2014	0.66	0.61	0.56
	2013	0.60	0.45		2013	0.66	0.60	
	2012	0.59			2012	0.66		
Highest Quintile	2014	0.58	0.40	0.23	2014	0.70	0.65	0.57
	2013	0.55	0.33		2013	0.69	0.63	
	2012	0.44			2012	0.68		
2nd Quintile	2014	0.50	0.32	0.11	2014	0.65	0.62	0.58
	2013	0.47	0.10		2013	0.65	0.59	
	2012	0.15			2012	0.64		
Middle Quintile	2014	0.51	0.29	0.03	2014	0.62	0.56	0.51
	2013	0.43	0.03		2013	0.61	0.54	
	2012	0.08			2012	0.58		
4th Quintile	2014	0.53	0.31	0.03	2014	0.61	0.53	0.51
	2013	0.45	0.03		2013	0.59	0.55	
	2012	0.08			2012	0.58		
Lowest Quintile	2014	0.58	0.39	0.14	2014	0.61	0.54	0.47
	2013	0.53	0.16		2013	0.63	0.53	
	2012	0.25			2012	0.58		

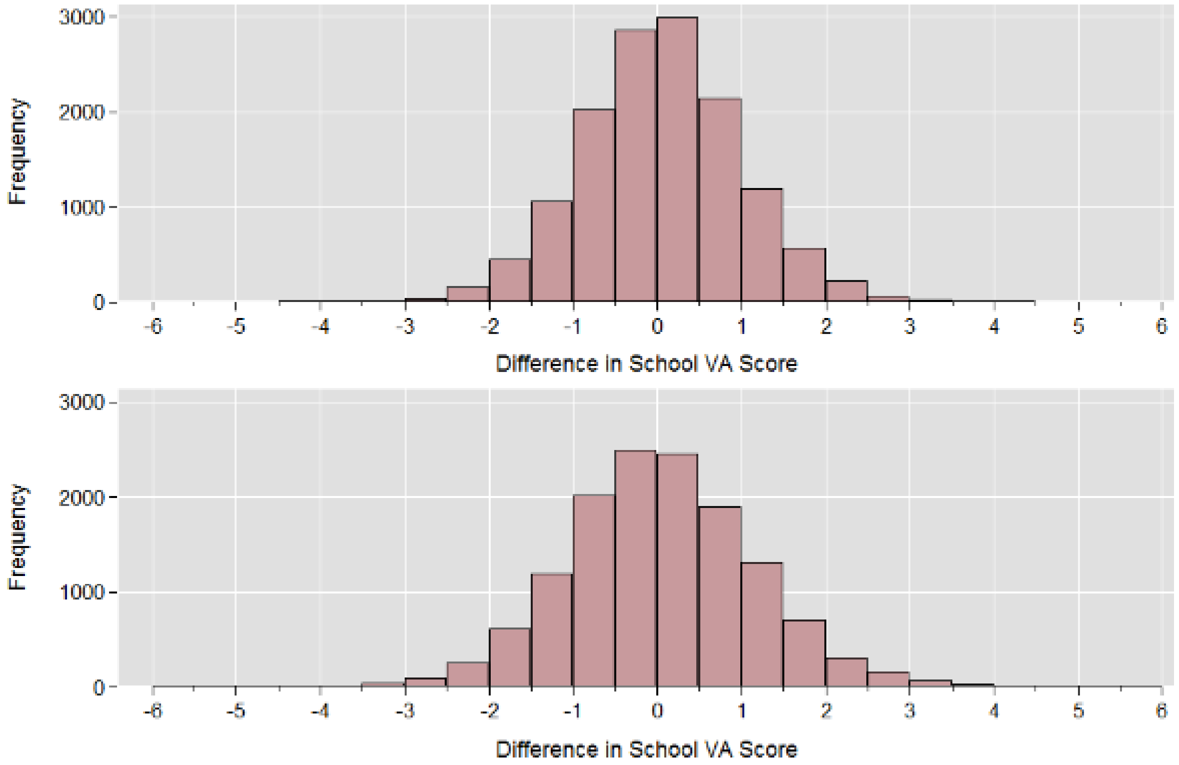
Quintile-correlations are based on between 2483 and 3075 schools

Table 6.3.2a shows very low rates of stability over 2 and 3 years and only moderate stability over 1 year. There appears to be a large change between 2011 and 2012 in the middle three quintiles in the distribution resulting in there being very little correlation between VA scores across these two years. This change is not seen in the APS scores which strongly suggests that the difference stems from changes at KS1 or in the value-added model in question.

Note that lower correlations would be expected when the distribution is broken up in this way as the differences in the scores would be larger relative to the variation in scale (now split into 5 parts). Also note that the range of scores covered in each quintile will vary such that central quintiles will have a slightly smaller range and yield lower stability estimates. As a result, the interpretation of these estimates is not strictly comparable to those using the whole sample: the actual stability of individual schools has not changed, only the scale on which it is compared. A more important result is that the level of stability is broadly consistent across the performance range, with little differences between higher or lower performing schools.

A perhaps more informative way of looking at the results is to look at the absolute size of the differences in school VA score over time. Figure 6.3.2a, below, shows the size of the difference in estimated primary school value-added scores in the official data of 1 and 2 years. The figures are in National Curriculum points, where 2 points equates to differences of about 8 months' worth of progress per pupil at the school.

Figure 6.3.2a – Distribution of 1 and 2 year Differences in Primary School Value-Added Scores



One final estimate was calculated to give an indication of stability. For each school with results in all of 2011 to 2014 (n= 13,135), the change in VA in 2011 to 2012; in 2012 to 2013; and in 2013 to 2014 was calculated. For each school, the minimum, maximum and mean changes were recorded. Finally, the mean of these three variables were calculated. The mean minimum change was 0.3 points or just over a month of progress per pupil. The mean maximum change was 1.2 points or just under 5 months of progress per pupil. The overall mean of each school's mean change was 0.7 points or just under 3 months' worth of progress per pupil. These three figures give a clear indication of the minimum, maximum and typical changes which can be expected in school VA scores on an annual basis over the course of 4 years.

This primary-level analysis was then repeated at secondary level. Table 6.3.2b gives the pairwise correlations in secondary school value-added and secondary school performance (using the total average capped point score, GCSEs only) over the period from 2011 to 2014:

Table 6.3.2b – Pairwise correlations over time in secondary school value-added and unadjusted attainment measures by performance quintile

Secondary Level								
	School (Best 8) Value-Added				Total Average Capped point score (GCSEs only)			
		1 year earlier	2 years earlier	3 years earlier		1 year earlier	2 years earlier	3 years earlier
All Schools	2014	0.56	0.49	0.45	2014	0.94	0.92	0.90
	2013	0.79	0.68		2013	0.96	0.94	
	2012	0.79			2012	0.96		
Highest Quintile	2014	0.40	0.37	0.19	2014	0.96	0.94	0.93
	2013	0.66	0.44		2013	0.97	0.96	
	2012	0.63			2012	0.97		
2nd Quintile	2014	0.41	0.25	0.15	2014	0.95	0.94	0.92
	2013	0.57	0.17		2013	0.96	0.95	
	2012	0.18			2012	0.97		
Middle Quintile	2014	0.35	0.18	0.04	2014	0.93	0.91	0.90
	2013	0.55	0.13		2013	0.95	0.93	
	2012	0.23			2012	0.95		
4th Quintile	2014	0.42	0.26	0.14	2014	0.89	0.86	0.81
	2013	0.55	0.17		2013	0.93	0.90	
	2012	0.20			2012	0.93		
Lowest Quintile	2014	0.46	0.32	0.27	2014	0.90	0.87	0.83
	2013	0.64	0.32		2013	0.93	0.90	
	2012	0.46			2012	0.93		

Quintile-correlations are based on between 490 and 606 schools

As with the primary results, breaking down the results into quintiles produces markedly lower correlations. As noted above, a reduction might be expected given that changes are now being given relative to a smaller range of schools. As with the primary results, there are large changes between 2011 and 2012 and a small tendency for greater levels of instability in the middle of the distribution. Again, the actual differences in school VA scores over 1 and 2 years were calculated. The distributions of these differences is shown in Figure 6.3.2c, below. Twenty four points on the Best 8 Score equates to about 4 GCSE grades per pupil at the school across the 8 GCSEs (plus bonus) included in the measure.

Figure 6.3.2b – Distribution of 1 and 2 year Differences in Secondary School Value-Added Scores

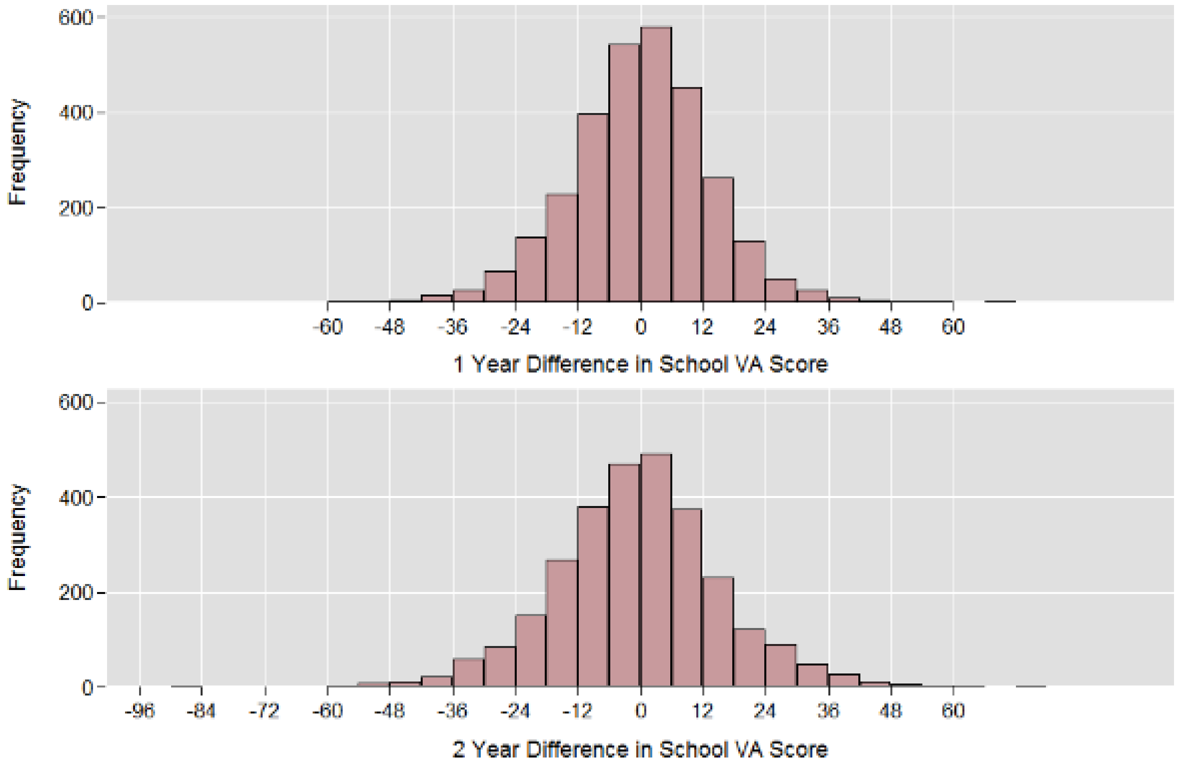


Figure 6.3.2b shows that within a single year, the performance of schools can be as much as 6 GCSE grades per pupil higher across the Best 8 measure and in some cases higher. Most schools see changes between 0 and 24 points higher or lower. When looking over 2 years, the range is not substantially increased, but the proportion of schools with changes over 12 points above or below is increased. One final estimate was calculated to give an indication of stability. For each school with results in all of 2011 to 2014 (n= 2792), the change in (Best 8) VA in 2011 to 2012; in 2012 to 2013; and in 2013 to 2014 was calculated. For each school, the minimum, maximum and mean changes were recorded. Finally, the mean of these three variables were calculated. The mean minimum change was 4.7 points or about ¾ of a GCSE grade per pupil. The mean maximum change was 21.1 points or about 3½ GCSE grades per pupil in a given year. The overall mean of each school's mean change was 12.1 points or 2 GCSE grades per pupil. These

three figures give a clear indication of the minimum, maximum and typical changes which can be expected in school VA scores on an annual basis over the course of 4 years.

6.3.3 RQ 3.3 - How stable is the contextual value-added performance of a given cohort over time?

School value-added scores, as examined above, refer to the final KS2/4 performance of specific cohorts. When one looks at stability across years, one is comparing the performance of successive cohorts which have passed through the school. It is possible that differences in cohorts are a cause of the instability observed in this study. If the value-added performance of a cohort was highly consistent over time, this would suggest that instability in school value-added scores was largely driven by differences in cohorts. If the stability of cohort performance was also low, this suggests that other sources of difference acting on and across cohorts are the source of school VA score instability.

This final research question examines the stability of the performance for a given cohort over successive years. This is possible as the ‘Making Good Progress’ study (see Chapter 5, Section 5.3.3) collected data for National Curriculum years 3 to 9 and ran for three calendar years (also see Table 5.5.2a, Section 5.5.2 for frequencies for each cohort in each year of the study). It is therefore possible to estimate performance score correlations for specific cohorts across 2, and in some cases, 3 years. A simple CVA measure was created, regressing the teacher-assessed mathematics scores (see the methods chapter, Section 5.3.3 for further details) on (exam-assessed) prior attainment at the previous key stage, this prior attainment score squared, gender and free school meals status (a measure of poverty). This model is specified in Appendix E1 where example output is also given. During analysis, a VA model without the contextual variables was also produced but as this gave highly similar results, these are not reported. The correlations below at primary level are based on the results of 251-271 primary schools in 2008 and 2009 and 207-226 in 2010. Secondary level correlations are based on the results of 67-71 schools in 2008 and 2009 and 48-51 in 2010. Correlations between cohort CVA performance across two years and, where data were available, across three years are presented below in Table 6.3.3a:

Table 6.3.3a – Pairwise correlations over time in cohort CVA performance

Cohort Letter	Initial year	Initial NC year	1 year earlier	2 years earlier	No. of Cohorts	Mean pupils per cohort
			Primary level			
B	2010	4	0.65		225	53.8
C	2010	5	0.66	0.52	226	57.7
	2009	4	0.61		272	54.6
D	2010	6	0.57	0.44	212	59.9
	2009	5	0.73		272	56.6
E	2009	6	0.55		260	56.3
			Secondary level			
F	2010	8	0.62		52	225.5
G	2010	9	0.69	0.62	49	233.7
	2009	8	0.46		70	223.0
H	2009	9	0.43		68	222.2

NB: Cohorts A and I were only in the sample for 1 year

These results show moderate stability of cohort performance over time, even when considering results only 1 year apart. There is some indication from the three data points looking at scores separated by two years that correlations across two years are approximately 0.1 lower than for those 1 year apart. These results will be considered alongside those in the next study, looking at the consistency between different cohorts within a given school.

6.4 Study 4 – Cohort Consistency

6.4.1 RQ 4.1 - How consistent are value-added estimates of performance across cohorts from within a single school in a single year?

When looking across time, it is difficult to know whether differences relate to the limitations of the value-added measure itself and its ability to separate school effectiveness from intake characteristics, from genuine changes in school effectiveness across time or from fluctuations in the examination system or some other change taking place over time. By looking within a single year, differences caused by change over time can be separated from differences present in a single time period. This final study uses teacher-assessed levels for National Curriculum years 3 to 9 over three calendar years from the MGP progress dataset (see Chapter 5, Section 5.3.3) to estimate the consistency of value-added scores for different cohorts within a school at a given point in time. This involved producing a CVA measure of relative progress from NC year 2 to 3, NC year 2 to 4, NC year 2 to 5, NC year 2 to 6 for NC years 3, 4, 5 and 6, respectively (i.e. the CVA since KS1 for each cohort is estimated). Similarly, at secondary level, KS2 scores were used and a CVA score was calculated to give the relative progress for each cohort in each school since their KS2 attainment. This CVA measure is the same as that used in Study 3, RQ 3.3 and is specified in Appendix E1, which also gives example output from one of the models. This was all done for three study years (2008-2010), yielding 3 estimates of each NC year correspondence. The primary-level results are given below in Table 6.4.1a and are organised so that the correlation between the CVA for a cohort and the cohort 1, 2 and 3 years below it is given in the respective columns where available:

Table 6.4.1a – Pairwise correlations of cohort CVA scores 2008-2010

		Primary Level		
Study Year		1 Year Lower	2 Years Lower	3 Years Lower
2008	NC Year 6	0.29	0.13	0.12
2009		0.48	0.27	0.19
2010		0.51	0.25	0.29
2008	NC Year 5	0.46	0.28	
2009		0.52	0.39	
2010		0.59	0.39	
2008	NC Year 4	0.43		
2009		0.49		
2010		0.51		
Mean		0.47	0.28	0.20

The correlations at primary level are generally very low to moderately low. This means that knowing the CVA performance of Year 6 (11 year-olds) in a primary school relative to similar pupils in other schools reveals very little about the performance of Years 3 or 4. Correlations of 0.4, 0.3 and 0.2 correspond to 16%, 9% and 4% variance common to both years, respectively. Even the correlation of consecutive years' performance of around 0.5 is not high; the CVA performances of year 6 explains just 25% of the variance in year 5 performances. The level of (in)consistency is most clearly seen on a scatter plot showing the CVA performances of year 6 and year 5 for each of the schools in the sample, see Figure 6.4.1a, below). The y=x line shows perfect consistency between years. Also recall that 2 NC points equates to an average of 8 months more/less expected progress for pupils in the cohort.

Figure 6.4.1a – School CVA performance for NC Year 5 and Year 6 in 2010 (r=.51)

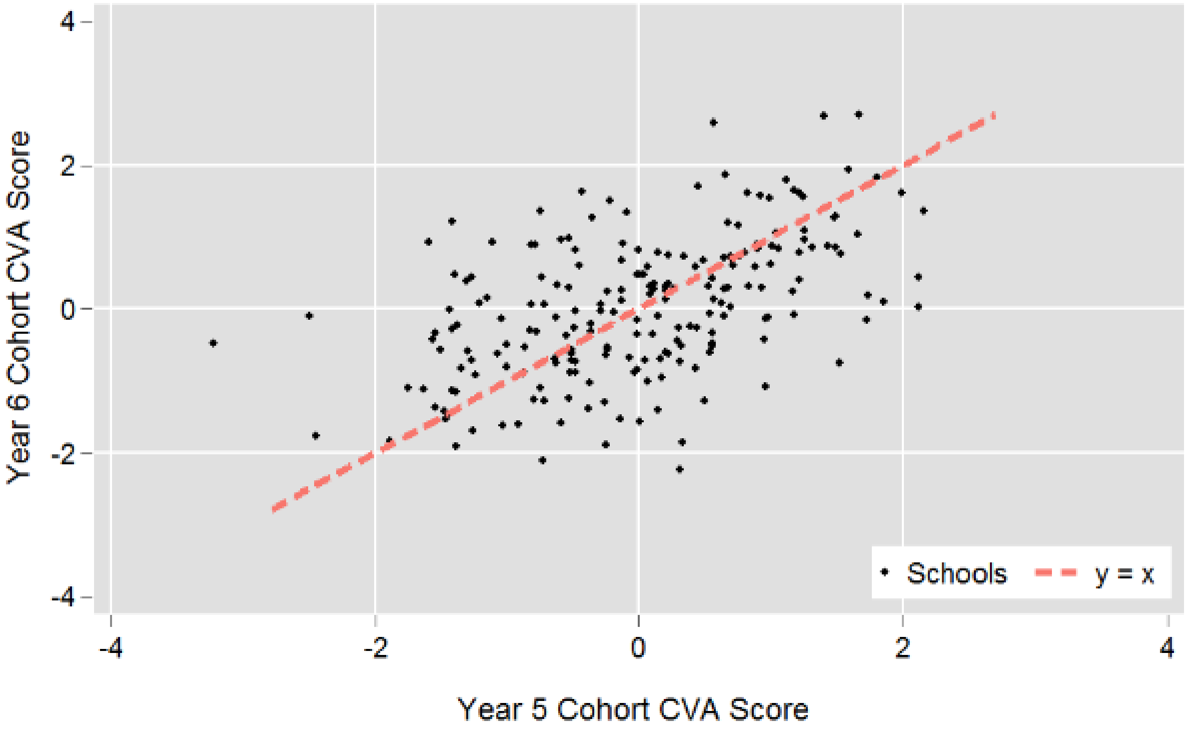


Table 6.4.1b, below, gives the equivalent results at secondary level:

Table 6.4.1b – Pairwise correlations of cohort CVA scores 2008-2010

Secondary Level			
Study Year		1 Year Lower	2 Years Lower
2008	NC Year 9	0.55	0.51
2009		0.75	0.44
2010		0.37	0.35
2008	NC Year 8	0.58	
2009		0.69	
2010		0.65	
Mean		0.60	0.43

There are a smaller number of schools at secondary level. There are also a number of outliers in the data. Given the moderation activities over the 3 calendar years of the study and the drop in school numbers in the final year, the 2009 results are likely to be the most robust estimates, suggesting mean correlations of 0.72 and about 0.45 for cohorts 1 and 2 years apart, respectively.

Both primary and secondary results, particularly the former, give a bleak picture of consistency in cohort performances within schools. It is likely that the assessments being conducted by teachers substantially contributed to this inconsistency (and instability in RQ 3.3 in Study 3). Despite this, the results should be considered alongside the low levels of stability found across years in value-added measures from the previous study. The results in Study 3, RQ 3.1 and 3.2 examined the official value-added measures which are based on standardised examinations; in that case, inconsistent teacher assessment cannot be the source of the instability. So while teacher-assessment is likely to have produced greater inconsistency, these results nevertheless suggest large differences in value-added performance between cohorts. Taken on their own these differences are sufficiently large to be a highly plausible driver of the year-to-year instability in the official VA measures demonstrated in the previous study, as different cohorts pass through the school. However, the instability of performance for cohorts across time in the last section suggests that instability is a more general problem with the measures which unobserved differences in cohorts cannot entirely account for. These explanations will be considered in the discussion chapter along with consideration of the implications of this instability for data use in schools.

6.4.2 RQ 4.2 - How consistent is performance within cohorts?

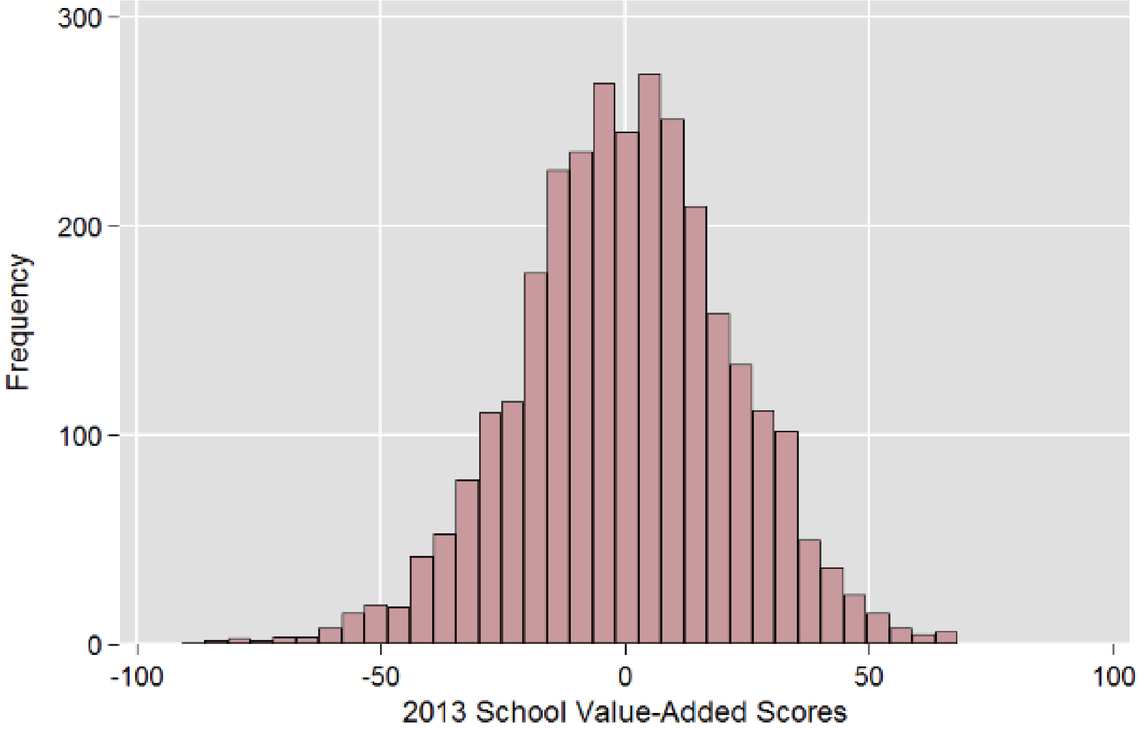
Currently, the DfE break down school value-added scores according to ability groups and for disadvantaged pupils. This gives some idea of the extent to which performance is consistent within a school’s cohort. This question takes this one step further to look at the distributions of pupil-level scores within schools to gauge the extent to which one can expect a mean score for a cohort or for groups within a cohort to be consistent with pupil-level scores within it.

This section looks at the level of consistency within the value-added performance of cohorts. This relates to the general problem that the overall school VA score is a mean value and so disguises the range of VA scores received by individual pupils. If a school were to receive a school-level value-added score of 20, for example, what range of scores can be expected from the individual pupils within the school? This section answers this question.

Secondary Level

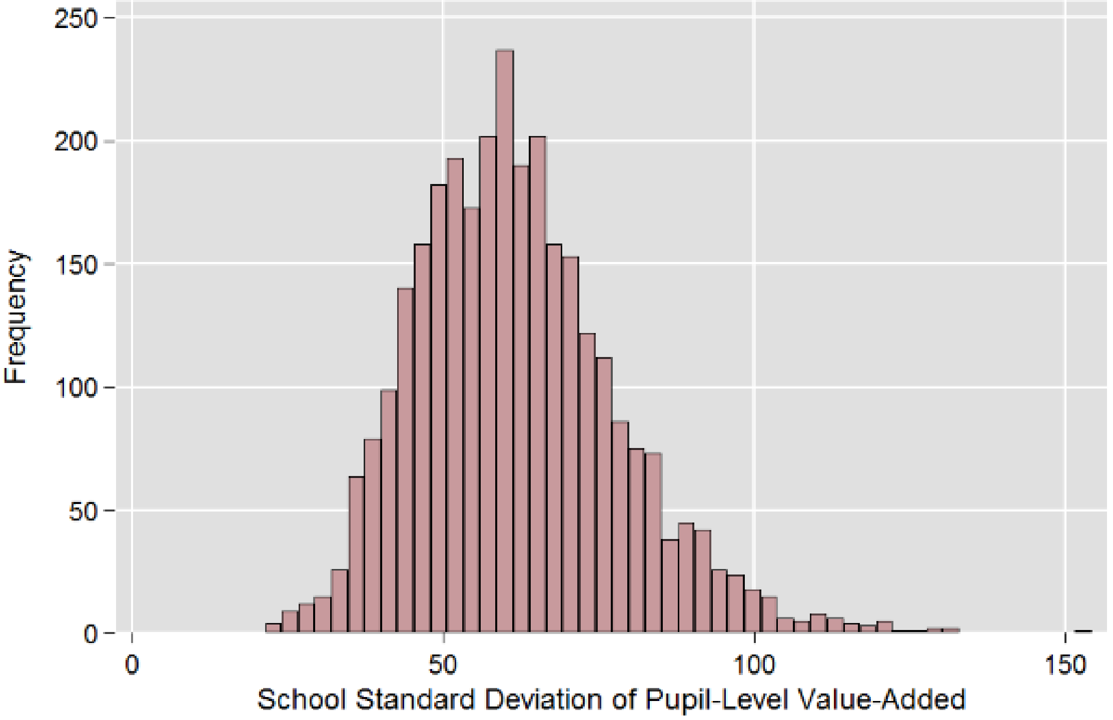
To provide context for the results that follow, the 2013 school-level VA score distribution is given in Figure 6.4.2a, below:

Figure 6.4.2a – 2013 Secondary School Value-Added Score Distribution



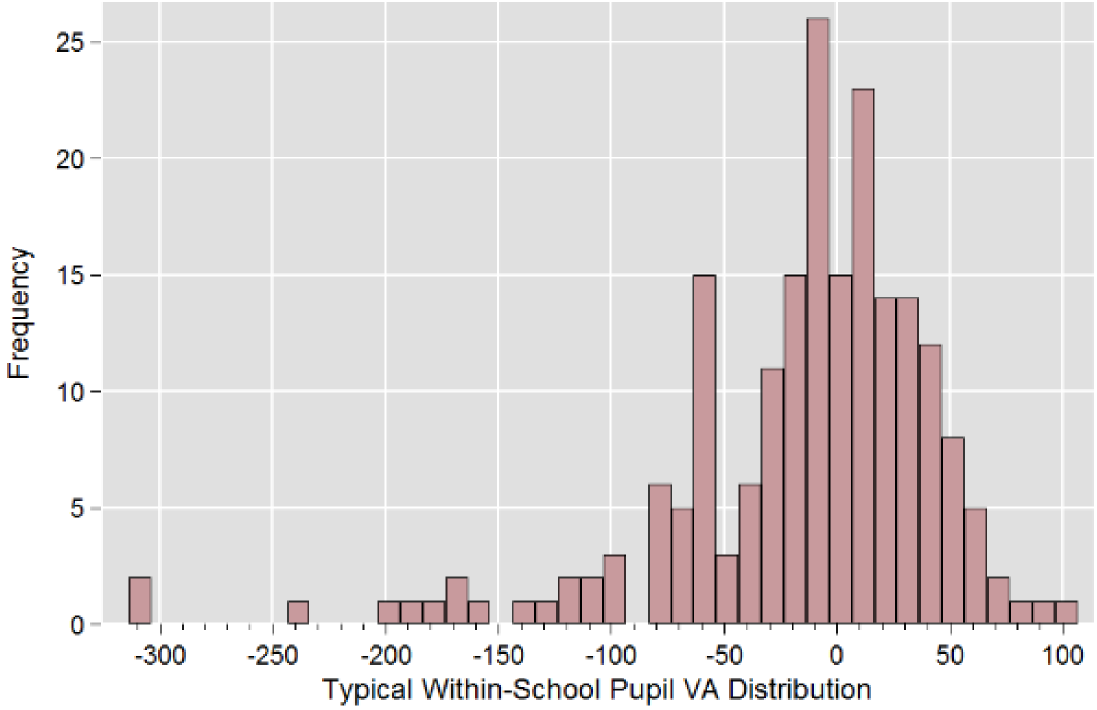
The extent to which pupil-level scores varied within these school-level scores was examined by estimating the standard deviation of pupil value-added scores within the 2013 pupil-level NPD data for each school, as shown in Figure 6.4.2b, below:

Figure 6.4.2b – Secondary School Standard Deviations of Pupil-Level VA Scores 2013



The (unweighted) mean school standard deviation score was 61.8. This means that the typical school value-added score disguises a pupil-level VA score distribution spanning over a hundred points either side of the mean result. To illustrate this by way of an example, a school with a pupil-level VA standard deviation of 61.8 (i.e. the mean standard deviation) was identified. This school had 201 pupils and a school-level VA score of minus 15.3. This school VA score of minus 15.3 suggests that this is a low performing school. This full pupil-level distribution this mean score is based upon is given in Figure 6.4.2c, below.

Figure 6.4.2c– Pupil-Level VA Distribution for a Typical Secondary School (School-Level VA score of -15.3)



As is clearly seen in this example, a school value-added score is a fairly unrepresentative mean of a vast range of pupil-level scores within the school. In this case ranging from about minus 300 to 100. Several other distributions for individual schools were examined during analysis. This example was chosen due to being the closest to the mean standard deviation of pupil-level value-added scores and its representativeness of the many other examples examined during analysis.

Primary Level

This analysis was repeated at primary level. Similarly, the starting point for this was the examination of the overall school-level VA distribution. This is given in Figure 6.4.2d, below.

This is immediately followed by the distribution of school standard deviations in pupil-level VA scores, given in Figure 6.4.2d, below:

Figure 6.4.2d – 2013 Primary School Value-Added Score Distribution

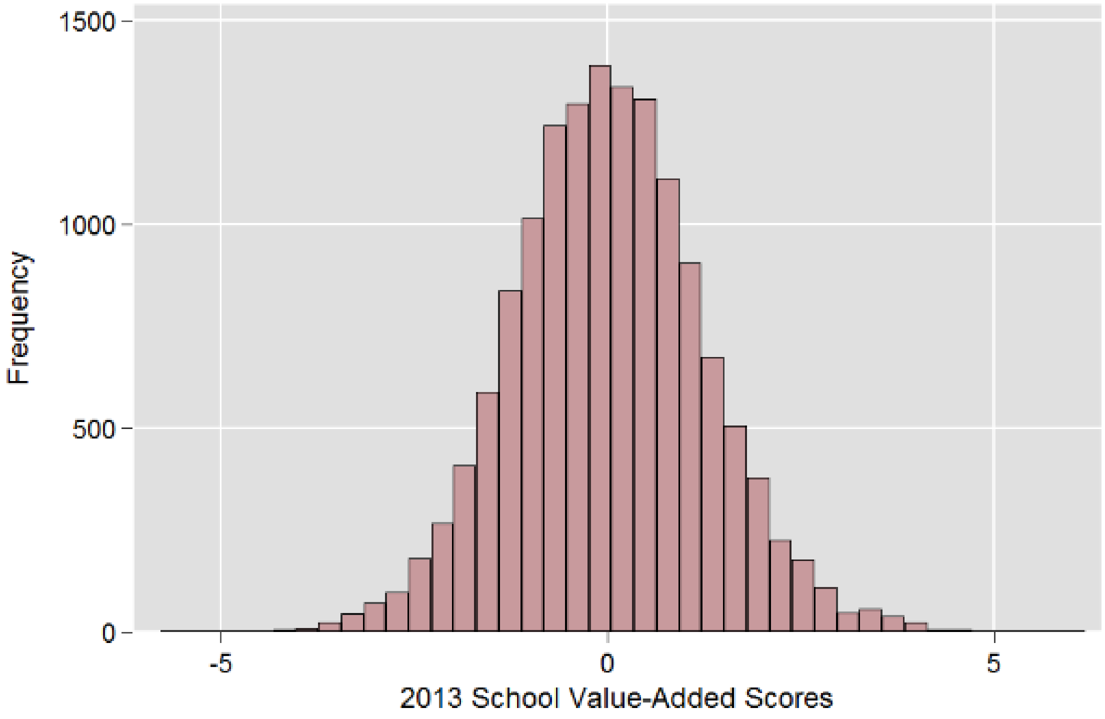
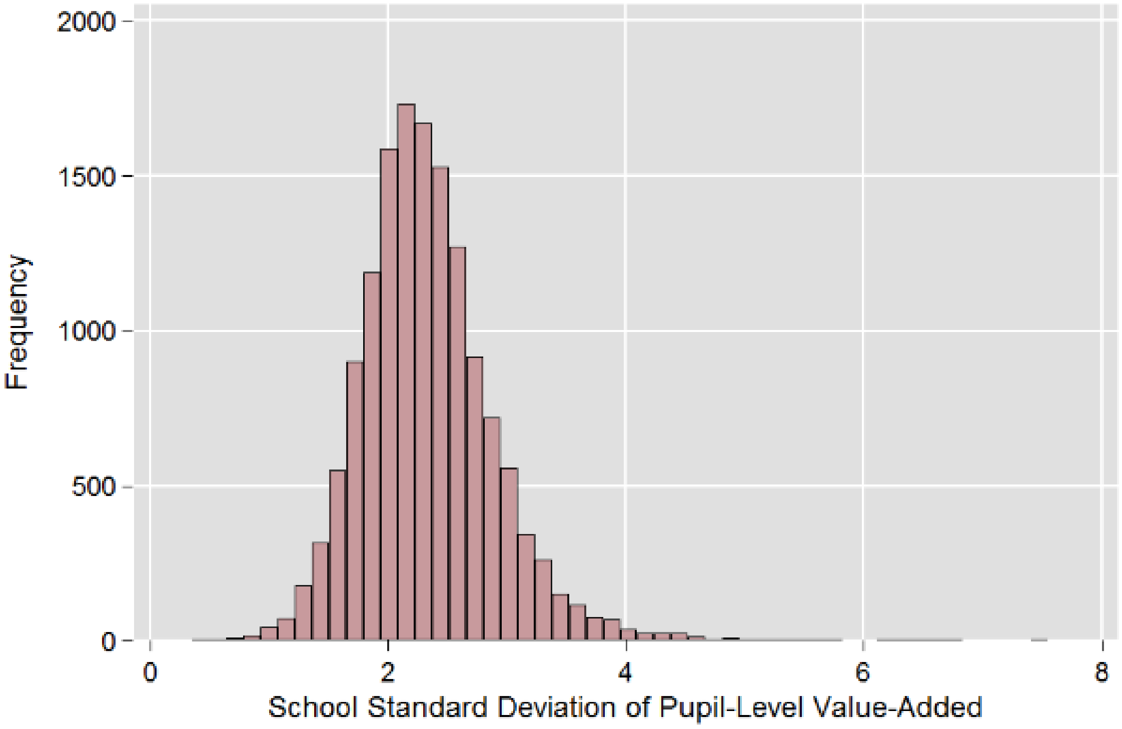
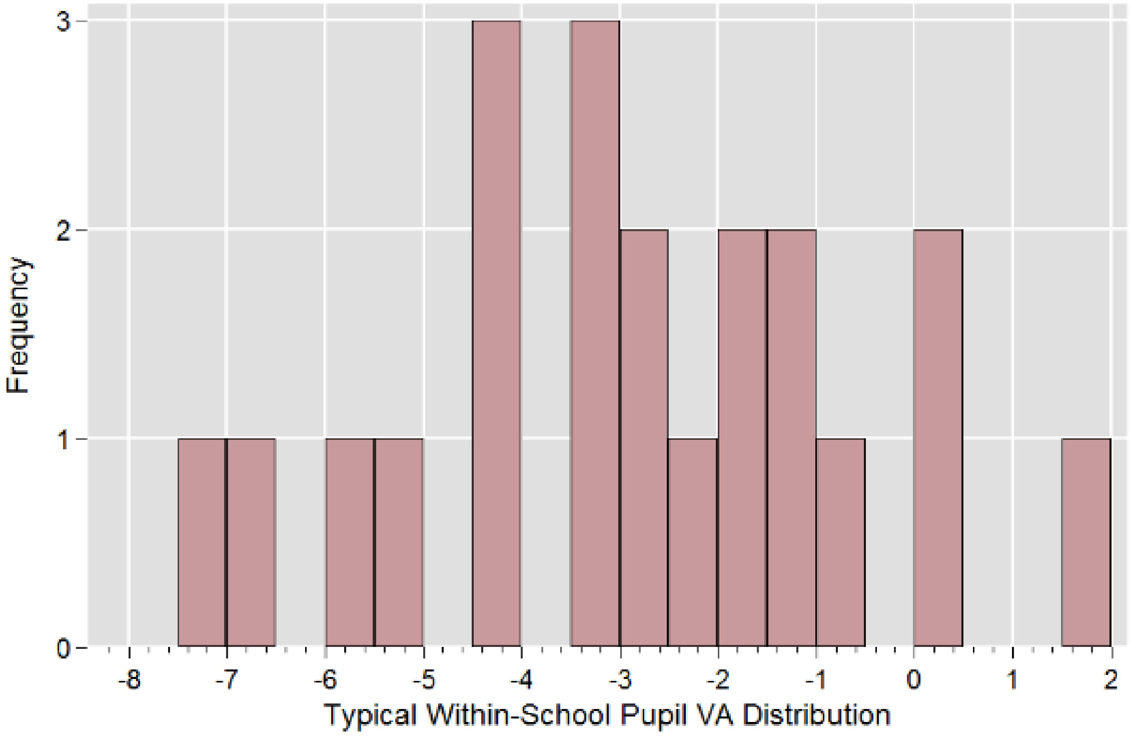


Figure 6.4.2e – Primary School Standard Deviations of Pupil-Level VA Scores 2013



The mean standard deviation was 2.33 NC points. The school closest to this was chosen as an illustration. This school had 21 pupils and a school-level VA score of -2.9.

Figure 6.4.2f – Pupil-Level VA Distribution for a Typical Primary School (School-Level VA score of -2.9)



According to the overall school VA distribution Figure 6.4.2d, above, this school has very low value-added of -3. The pupil-level distribution shows that pupil VA performance ranged from -7.5 to 2 national curriculum points.

6.4.3 RQ 4.3 - Does within-cohort consistency vary by mean school performance?

This final section of study 4 examines the relationship between the within-cohort consistency (see last section) and school performance. Similarly, the aim is to estimate the range of school scores within school cohorts across the performance range and so the extent to which a high school VA, for example, reflects the performance of all pupils within the cohort. At both secondary and primary level, 6 performance scores were selected to represent very low, low, fairly low, fairly high, high and very high levels of school performance. To do this, the overall school VA distribution was used as a guide to what scores relate to particularly low and high performance. The schools closest to these performance scores were selected and their pupil-level distributions (as above) are compared on a series of graphs showing typical pupil-level distributions across the school VA performance range. As these are scores for individual schools, chosen as described above, these are given as illustrative rather than representative examples of pupil-level consistency. It would be possible to explore the entire pupil-level distribution for all schools within a performance band but it is thought more informative to look at the variation that can be expected for individual schools.

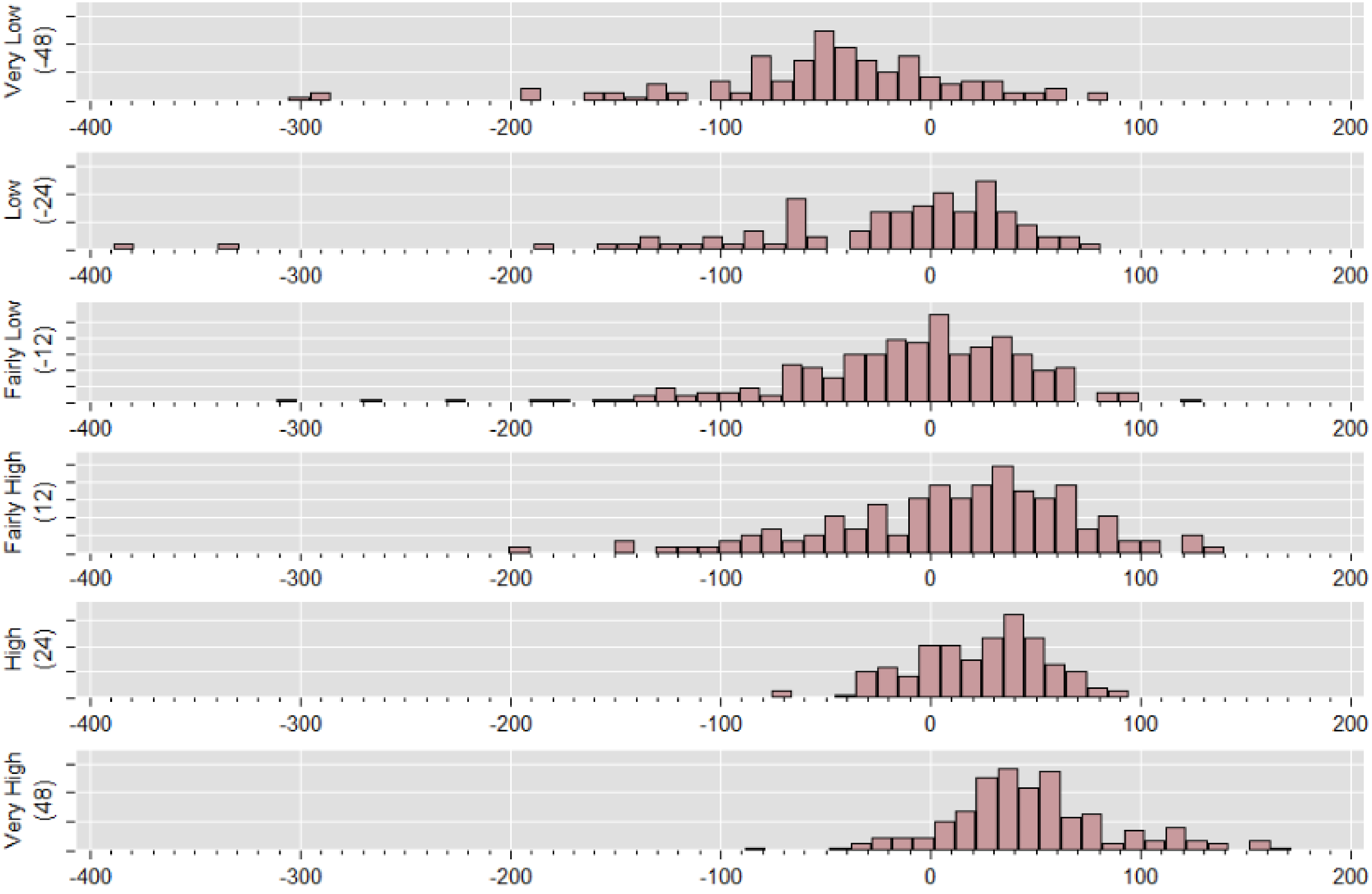
Figure 6.4.3a, below, shows pupil-level value-added distributions for six selected schools across the performance range to give an indication of the spread of pupil-level VA scores which can be expected. Summary statistics of these are given in Table 6.4.3a, below.

Table 6.4.3a – Summary Statistics of the Pupil-Level VA Distributions for Selected Secondary Schools across the Performance Range Shown in Figure 6.4.2a

School	Mean VA	Pupils	Std. Dev	Min VA	Max VA
Very Low	-48.2	137	64.6	-305.2	79.0
Low	-24.0	89	75.9	-388.7	78.0
Fairly Low	-12.0	232	60.5	-310.0	123.2
Fairly High	12.0	143	58.9	-200.4	135.2
High	24.0	137	32.0	-75.4	90.9
Very High	48.1	175	40.6	-88.1	163.6

Table 6.4.3a and the corresponding figure, (Figure 6.4.3a, below) suggest that there is a large range of pupil scores across the school performance range. This is to be expected by what is known about the extent to which differences lie within rather than between schools (Reynolds, 2008) and the calculated level of statistical uncertainty in the school-level scores (Leckie and Goldstein, 2011).

Figure 6.4.3a – Pupil-Level VA Distributions for Selected Secondary Schools across the Performance Range



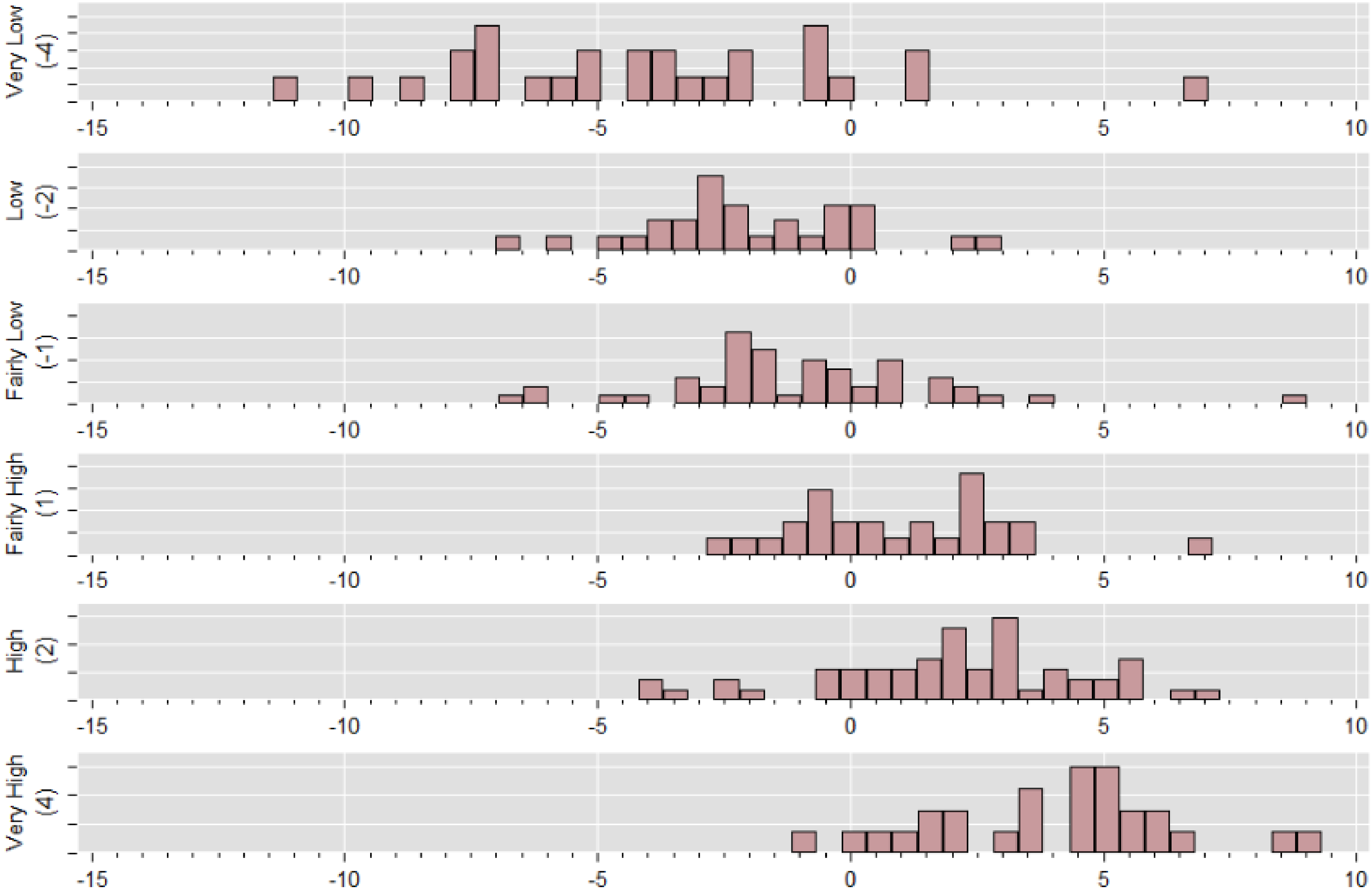
As can be seen, even schools identified as very high attaining on the basis of the school VA score had pupils with performance the equivalent of 10-15 GCSE grades lower than would be expected across the Best 8 subjects. Conversely, the lowest performing schools had pupils achieving this size of VA score above the statistical expectations.

This analysis is now repeated at primary level. Table 6.4.3b, below, gives the summary statistics for the 6 schools chosen. Figure 6.4.3b shows the pupil-level VA score distributions on a series of histograms. As above, the schools were chosen as being the closest to school VA scores chosen for each level of performance using the distribution. The only exception was the 'high' performing school. The closest school to a VA score of 2 had only 6 pupils. The next closest, with 54, was selected as this was more typical of other schools in this performance vicinity which were inspected. As with the secondary results, the primary school pupil-level VA score distributions show that school VA scores disguise a large range of pupil VA performances.

Table 6.4.3b – Summary Statistics of the Pupil-Level VA Distributions for Selected Primary Schools across the Performance Range Shown in Figure 6.4.2b

School	Mean VA	Pupils	Std. Dev	Min VA	Max VA
Very Low	-4.0	27	4.0	-11.4	6.9
Low	-2.0	28	2.2	-7.0	2.9
Fairly Low	-1.0	49	2.7	-7.0	8.8
Fairly High	1.0	27	2.1	-2.8	6.9
High	2.0	54	2.6	-4.2	6.9
Very High	4.0	27	2.4	-1.2	8.9

Figure 6.4.3b – Pupil-Level VA Distributions for Selected Primary Schools across the Performance Range



7. Discussion

7.1 Chapter Introduction

This chapter discusses the results from Chapter 6 and their implications for value-added in general and for specific areas of use, as reviewed in Chapter 3.

The first main section (Section 7.2) discusses the results in relation to areas of literature reviewed in Chapter 4. The section concentrates on what can be said about the validity of value-added in general and discusses general methodological and technical issues which are informed or raised by the present results.

A central conclusion of this thesis (see Chapter 8) is that validity is often best considered in relation to interpretation and use of a measure rather than just being considered as a property of a specific measure. This position draws on recent thinking in validity theory stressing the importance of considering validity (and validation) in far broader terms than the technical validity of specific measures (Newton and Shaw, 2014). The opening subsection of Section 7.3 (Section 7.3.2) addresses the theoretical grounding for situating consideration of validity in relation to interpretation and use, first in general and then specifically in relation to value-added. Although the general consideration briefly reviews some existing literature, it was clearest to present this immediately before the original material relating the general theory to value-added rather than separate these between the literature review chapter and this one. This has the added advantage of putting the key grounds for considering validity in terms of interpretation and use, as well as the major considerations for doing this, directly before the discussion sections on the implications of the findings for particular uses.

The theoretical material is followed by a series of subsections (Sections 7.3.3, 7.3.4 and 7.3.5) which combine a) the theoretical understanding of validity in relation to interpretation and use; b) information on the use and interpretation of value-added measures across research, policy and practice presented in Chapter 3; and c) the results presented in Chapter 6 to discuss the implications of the results for specific uses of value-added in research, policy and practice. With the exception of the research context, particular uses are considered in relation to the current English system where the findings are particularly relevant.

7.2 General Implications of the Results

7.2.1 Section introduction

Section 7.2 links the results presented in Chapter 6 with the pre-existing literature reviewed in Chapter 4. It is organised along similar lines to clearly identify the contribution of the results to the issues reviewed in Chapter 4. There are five main sections, each of which corresponds to a section or sub-section within Chapter 4, as follows: Bias (Section 4.2.2), Measurement Error (4.2.3), Stability (4.2.4), Consistency (4.2.5) and the Regression Discontinuity Design (4.4).

7.2.2 General Implications of the Results

Bias

The results in Study 1 (RQ 1.1) highlighted the importance of contextual variables for creating unbiased measures. This is a longstanding finding (Teddlie and Reynolds, 2000). There is, however, value in updating and replicating estimates in new datasets, especially those with particular importance such as the NPD. Estimates presented suggest that about 35% of the variance in the secondary school value-added scores can be explained using a small number of theoretically important contextual variables. The figure was lower at primary level where 10% of the school-level value-added was explained by contextual factors. This may also be related to the fact that the KS1 baseline was from within the same school. Where this is the case, one-off effects of pupil contextual factors will be accounted for within the baseline, although ongoing effects will not. Where the baseline is from a previous school and there is subsequent sorting to secondary schools by pupil characteristics, contextual variables may be acting as proxies for unobserved pupil characteristics.

The distinction between within-school value-added and across-school value-added measures is also likely to have bearing on the issue of school-level compositional effects. Study 1 found that not controlling for prior attainment at school-level did not wholly prevent school-level associations between prior attainment and value-added. At secondary level, the correlation was small and positive; at primary level it was small and negative. A negative compositional effect in the primary age range was also found by Televantou et al. (2015); who also found that controlling for measurement error made the negative effect more marked. See below for further discussion about the possibility of positive biases on compositional effects through measurement error in the next sub-section.

One result which is maybe surprising (although it is hinted at in recent research such as Burgess (2014)) is the tendency for pupils with English as an additional language (EAL) to get higher value-added. Again, the value-added evidence does not in itself provide clear explanations for this and there are many possible explanations such as pupils' English language competency catching up with their main language; favourable characteristics or practices of immigrant populations; the benefits of bilingualism; or perhaps greater school effectiveness for these populations (in which case this is not a bias). Further research as to why this effect is found would be valuable for the specification of value-added models in particular as well as the obvious wider value for educational research.

One particularly striking result from Study 1 relates to the finding that the negative relationship between free school meals (FSM) status and performance varied according to the proportion of pupils in a school who were eligible. Again, narrowing down the cause of this effect would be of great value. It might be that socio-economic indicators operate poorly when used across socio-economic regions and populations (such as between urban and rural) or there may be some educational or social factor at play with regards to group composition.

Measurement Error

As reviewed in Chapter 4, Section 4.3.3, Reynolds et al. (2012, p.8) argue that errors 'tend to be randomly distributed' and so have little impact on the school-level scores, whereas Gorard (2011a, p.18) argues that this is unlikely and that 'it is unfair and unethical' to assume this is the case. The remarkable result from Study 1 is that *even if* pupil-level errors were random (across pupils, time and schools), these can still translate into serious school-level errors. Moreover, these errors were systematic, raising concerns about being unfairly advantageous for schools with more able intakes and raising the possibility that the grammar school effect could stem entirely from this effect (and so be entirely spurious). This general problem was more severe in value-added simulations for KS2-4 than KS1-2, suggesting that the problem is worse when prior and final attainment measures are poorly matched.

The finding supports those in RQ1.1 relating to positive school-level correlations between prior attainment and value-added, again suggesting that these are linked with measurement error in a systematic fashion to produce a spurious grammar school effect. It is also in line with the findings of other studies such as Harker and Tymms (2004), Televantou et al. (2015) and He and Tymms (2014) who show that measurement error can lead to systematic biases favouring groups of higher-ability pupils (see Chapter 4, Section 4.2.2). By looking at this problem in this new way (i.e. using the simulation) new explanations have been suggested as to the mechanism by which random pupil-level error can lead to systematic school-level error within value-added calculations. The present study seems to be looking at the same problem from a new angle and drawing a similar conclusion: Errors are likely to have a systematic impact on school value-added scores, especially in relation to 'phantom' compositional effects.

It is the systematic nature of the effect and the explanations for this (as suggested in the above papers and in Chapter 6, Section 6.1.3) which suggests that this is a more general problem within the value-added methodology rather than a specific problem in the English data, where it appears to be particularly acute in the KS2-4 data. If this is a more general problem and applies outside of the particular measures studied, the assertion (Reynolds et al., 2012, p.8) that errors 'tend to be randomly distributed' and so are 'unlikely to be systematically different in different schools' could be wrong on both counts.

Stability

The results in Study 3 estimated primary level stability to be 0.6, 0.46 and 0.36 for scores 1, 2 and 3 years apart respectively. These results are not dissimilar to previous estimates (see Chapter 4). This level of instability in primary-level value-added scores means either that value-added is not providing a valid measure or that effectiveness is not a stable property of primary schools (Marks, 2014). This finding has now been confirmed over time, across different systems and using numerous datasets and is squarely at odds with the use of school value-added for high-stakes accountability. Whether primary school value-added can be used for low-stakes purposes such as monitoring or school self-evaluation is considered further in Section 7.3.

At secondary level, the degree of stability is somewhat higher. The present results estimated stability of English secondary VA scores over 1, 2 and 3 years of 0.8, 0.7 and 0.45, respectively. It is likely that the latter result is depressed due to changes in the qualifications points system. In line with previous research (Gray et al., 2001), the value-added scores were considerably more stable than the raw scores (at secondary level). This comparison suggests that ignoring contextual variables reintroduced biases back into the measure, made them closer to the raw scores and thereby made VA more stable over time than contextualised value-added measures (see Chapter 4, Section 4.2.2).

What should be made of these rates of stability at secondary level? Secondary-level scores are certainly more defensible than primary-level scores as a meaningful measure of (changing) school performance over time. A correlation of 0.8 corresponds to just under two thirds of the variance in school VA scores being explicable by the previous year's score. A correlation of 0.7 across two years means that just under half of the differences in a given year's schools' VA scores are explicable by results 2 years earlier. The difficulties of interpretation were discussed in Chapter 4 in relation to stability specifically (Section 4.2.4) and value-added inference more generally (Section 4.3.2). There seem to be considerable differences in interpretation of correlations in this region. It is maybe best to consider stability in terms of the size of the difference which can be expected in the underlying measure. Results in Chapter 6, Figure 6.3.2a estimated a mean change in secondary school value-added of 12.1 points or 2 GCSE grades per pupil. This seems large as a typical change, but would not be surprising in a single given case. It is difficult to infer much about validity from these as the scores are not sufficiently volatile so as to rule out a meaningful school effect. Unlike the primary-level results, these secondary figures do not obviously support either the critics or proponents of value-added. One important note, however, is that the level of stability is substantially inflated by the failure to control for contextual factors in the English VA measures.

The final consideration relating to stability which has been examined in the results is the extent to which a given cohort's performance is stable over time. The results suggest that, rather than cohorts' relative performance being steady over time, there appears to be year-to-year change. This is interesting to consider in light of the issue of omitted variable bias. Suppose that unobserved (but stable) non-school factors had a considerable influence on the value-added scores and the differences between cohorts were largely explained by these. If this was the case, we would expect 1) a fair degree of stability in the relative performance of cohorts in different schools over time yet 2) instability in school value-added as different cohorts 'pass through' the key stage years. Moderate stability suggests that this is not the case and that stable unobserved differences are not the main driver of year-on-year instability in the school scores. Rather, this result suggests a more general problem of instability. This line of thought is returned to after looking at the results from the consistency study.

Consistency

A problem with looking at stability over time, as above, is that a degree of change is expected. When considering whether 0.7 is a sensible degree of correlation for VA scores separated by two years (above), it is difficult to identify measurement unreliability from change in school performance. The fourth and final study in this thesis is of great value because of this problem. It is also of significant value as this issue ‘has not been adequately researched’ (Teddlie and Reynolds, 2000, p.118) so there are only a very small number of studies from which estimates of consistency can be based.

The results of Study 4 suggest that consistency of value-added scores for different year groups at a given point in time is low. The correlation in performance for cohorts only 1 year apart was moderate; the performance of cohorts 2 or 3 years apart is barely related. At secondary level, consistency was somewhat higher but still suggests that VA scores cannot be safely generalised from a single cohort to the school at large. This is an important issue as the school performance tables present annual results based on one cohort with the implication that these give the performance of the school rather than merely the cohort which most recently left (Goldstein, 1997). These results support those of Mandeville and Anderson (1987) who found ‘discouragingly small’ correlations and characterise consistency between grades 1 through 4 as ‘very unstable’ (p212 & 203, respectively). Also, in line with the stability results, primary level correlations are too low to support the view that school effectiveness is a meaningfully consistent effect as well as the view that value-added is a reasonably precise measure of it. Secondary-level results do not lend themselves to unequivocal conclusions, especially as the available data were only for national curriculum years 7 to 9.

The results at primary level show that the results of consecutive NC year groups are more similar than those 2 or three years apart. What does this tell us about what is driving the results? If this instability were the result of random fluctuations in cohort characteristics or random measurement error, we would expect the correlations in performance between cohorts to be similar irrespective of how many years apart they are. Consider some possible explanations: First, it might be that cohort characteristics gradually change over time as the local neighbourhood changes or as the school’s reputation changes and a different intake is attracted. Second, this may reflect the NC assessment scheme: what it means to be good at maths, for example, slowly changes throughout the age range. This might make closer cohorts have a higher correlation in relative performance given more similarity in the measure. Third, this may be something to do with the likelihood of cohorts receiving the same quality of education. It is more likely that year 6 and year 5 in a school had the same year 3 teacher, than year 6 and year 4. With staff changes over time, as year groups are further separated, they are less likely to have had common inputs (in terms of teachers and curriculum) at various stages. The third explanation has the most intuitive appeal, although given the low correlations for teacher-effectiveness found in Chapter 4 (Section 4.2.5) it is maybe more accurate to understand this as various factors (pupils, teachers, curriculum and myriad complex cultural and pedagogical factors) coming together to create a more or less favourable learning environment for a period of time which consecutive year groups are more likely to share.

Regression Discontinuity Designs

The results in Study 2 suggest that the cross-sectional RD design is unsuitable for use to compare the relative effects of individual schools because it relies on the assumption that a smaller group of pupils within the same school (but in the lower cohort) are a suitable baseline for estimates. Volatility between cohorts' performances (as in Study 3 and 4) make the cross-sectional design vulnerable to differences between cohorts and these are likely to confound any differences in school effects sufficiently to make estimates at this level valuable. In contrast, the VA design was found to produce results which are almost identical to actual differences in levels of progress (recorded using the longitudinal regression discontinuity design). This result alone is valuable as it suggests that VA is capturing differences in pupils' performance (although note this does not it itself make such variation causally attributable to schools).

Despite these difficulties at school-level, results suggest that RD remains a powerful design for larger cross-sectional studies where this problem will not be apparent and absolute measures of performance are desired. The RD design produced very clear and consistent estimates of how performance varies by maturity and with schooling across the national curriculum years studied. Indeed, this focus on system-level performance and the factors influencing school effectiveness over a large sample have been the more common use of the RD to date (see Chapter 4). Unlike VA, the RDD produces estimates which are not relative and so can be used to identify changes in system-level or area-level performance over time. The RD study (Study 2) has, for example, identified large differences in the amount of progress made in different National Curriculum (NC) years, results which have implications to the system and to VA given that it uses NC year 6 results as a baseline.

7.3 The Validity of Value-Added in Relation to Interpretation and Use

7.3.1 Introduction

Looking ahead, Chapter 8 evaluates all of the available evidence to reach a position on the core research question. In short, the general position on the validity of school value-added reached is that school value-added is seriously and fundamentally flawed as a measure. It is not, however, concluded to be either meaningless or valueless. The task of making beneficial use of the evidence is held to rest heavily on the particular interpretations and uses of (specific) value-added evidence in a given context. Moreover, identifying specifically what constitutes valid interpretation and use in a given area is a difficult but important question in its own right.

Section 7.3 makes important steps towards this conclusion. The first sub-section (Section 7.3.2) is designed to achieve two goals: first, to provide a short theoretical grounding for positioning validity (or validation) as a practical matter involving consideration of interpretation and use as well as the properties of the measure. The second goal is to develop this theoretical grounding to align it more specifically with value-added validation. Both of these aims support the remaining subsections as well as work towards the concluding chapter. The three remaining subsections (Section 7.3.3, 7.3.4 and 7.3.5) discuss the implications of the results for policy, practice and research respectively.

7.3.2 Theoretical Basis for Considering Validity in Relation to Interpretation and Use

General Theoretical Grounding

As reviewed in Chapter 3, there are numerous different uses of value-added evidence across and within research, policy and practice. Even within particular areas, there were considerable differences in the attitudes and interpretations of the measures (see Chapter 3, Section 3.5.2). As well as differences in the type of value-added evidence (and models) used, even the exact same evidence raises many potential distinctions: Researchers, policy-makers, school inspectors, parents and school leaders, for example, can all potentially look at the same value-added evidence (e.g. a school's VA score) and make different interpretations and put this information to different uses. Moreover, it is clear that different demands could be made on the validity of the evidence, different levels of professional and technical expertise can be brought to bear during interpretation and there are markedly different consequences for getting it 'right' and 'wrong'. If value-added evidence is to have any effect – positive or negative – it is through interpretation and use. One is tempted to conclude that it is more valuable to identify which interpretations and uses are 'valid' given what is known about the validity of value-added evidence rather than to try and generalise about whether the evidence is sufficiently valid to be of value in general. Note that this idea shifts the meaning of validity from something inherent to a measure to a broader concept which also encompasses its interpretation and use, positioning validation of these as an important endeavour (Newton and Shaw, 2014).

While this conception of validity is far from new (Messick, 1987, Brennan, 2013, Newton and Shaw, 2014), it has received considerable attention in recent years, in large part because of the clear statement and development of the 'argument-based approach to validation' in Kane (2013, p.1). Kane's paper is 'the most complete and clearest discussion yet available of the argument-based approach to validation' and 'with the contributions of Kane (and others) we now have a practical, useful scaffolding that provides ways to frame and address claims about test score interpretations and uses' (Brennan, 2013, p.73 & 81). As this scaffold is precisely what is required to consider the validity of value-added in relation to interpretation and use, this section presents key ideas from this paper before adapting the general ideas to the context of value-added. This presentation is necessarily selective and summary. The basic conception of validity behind the argument-based approach to validation is as follows:

> "Validity is not a property of the test. Rather, it is a property of the proposed interpretations and uses of the test scores. Interpretations and uses that make sense and are supported by appropriate evidence are considered to have high validity (or for short, to be valid), and interpretations or uses that are not adequately supported, or worse, are contradicted by the available evidence are taken to have low validity (or for short, to be invalid). The scores generated by a given test can be given different interpretations, and some of these interpretations may be more plausible than others.
>
> (Kane, 2013, pp.3-4)

As well as shifting the focus to interpretation and use, this conception takes validity to be a 'matter of degree' which 'may change over time as the interpretations/uses develop and as new evidence accumulates' (Kane, 2013, p.3).

The argument-based approach to validity identifies two types of argument: first, the *interpretation/use argument* (IUA). An IUA makes claims based on the information captured within a particular test. If these claims are modest, little empirical support is needed; where more ambitious claims are made, more evidence of greater quality is required to support them (Kane, 2013, p.21). Second is the *validity argument* which 'provides an overall evaluation of the claims in the IUA' (Kane, 2013, p.14) on the following terms:

> "The proposed interpretations and uses are valid to the extent that the IUA is complete and coherent, that its inferences are reasonable, and that the assumptions supporting the warrants for these inferences are either inherently plausible or are supported by adequate evidence."
>
> (Kane, 2013, p.14)

Within the validation argument, it is often important to make distinctions between the validity of interpretations and use: rejecting a use does not necessarily invalidate the underlying interpretation and accepting an interpretation does not validate a use (Kane, 2013, p.46).

One aspect of this conception of validation which is particularly pertinent to value-added relates to the evaluation of precision, uncertainty and generalisability of the evidence. As Kane notes, test scores are often used to make generalisations across tasks, raters or context. Similarly, value-added scores are often used to make generalisations over time, subjects or pupil groups. The stability and consistency evidence presented in Chapter 6 has, to this point, largely been positioned as indirect evidence on validity. It can now be seen in a new light: evidence on the inconsistency of performance between cohorts presented in Study 4, for example, can be used to critically evaluate IUAs which rely on generalisations about school performance across and within cohorts. IUAs which accurately reflect the large range in pupil performances found may be regarded as valid (subject to other concerns) and IUAs which over-generalise what is highly inconsistent performance evidence can be said to be invalid.

Another issue which has an important bearing in the context of test scores in addition to value-added is that of error and uncertainty. Kane explains that because the assumptions underlying various claims within IUAs are ‘hardly ever exactly true’, we must ‘build some slack into the system’; without this – i.e. where ‘absolute confirmation’ is demanded for all claims – all IUAs ‘would be shot down immediately’ (Kane, 2013, p.3):

> … typically we [create slack in the system] by postulating the existence of errors (random and systematic) of various kinds. The explicit recognition of uncertainty makes the interpretations viable, but it also makes interpretations a bit fuzzy and decisions a bit tentative. As is the case in evaluating scientific theories, we never achieve certainty but we can achieve a high degree of confidence in certain interpretations and uses of test scores.
>
> (Kane, 2013, pp.3-4)

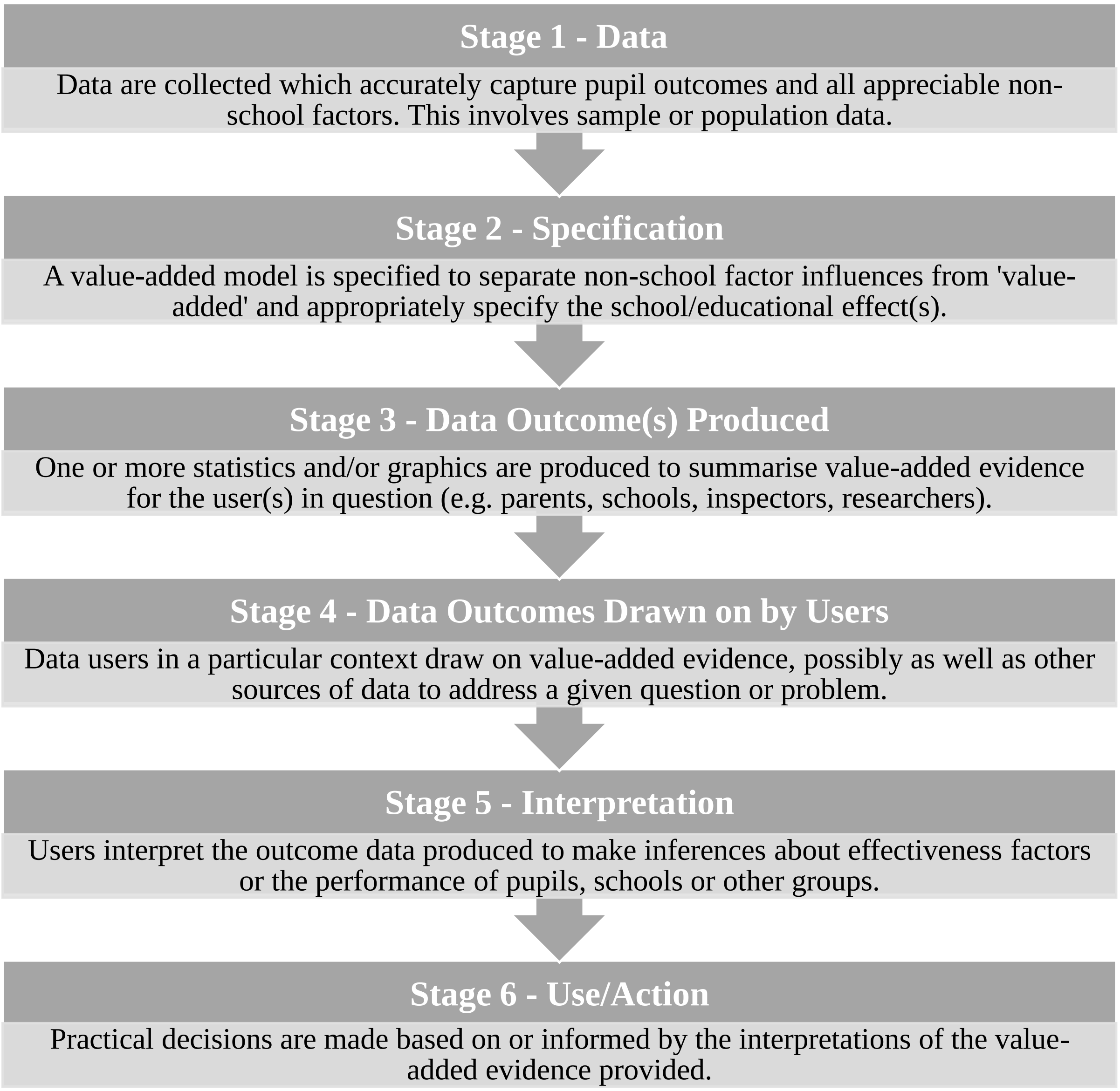
Chapter 4, Section 4.3.3, discussed the appropriate way of understanding uncertainty, bias and error in value-added estimates. As with the issues of consistency and stability, we can now reposition this issue in light of this broader conception of validity. Rather than solely seeing the debate as working towards a greater understanding of the validity of the measure, we can also cast confidence intervals and any other approaches or information designed to aid interpretation in terms of their role in promoting valid IUAs. It is of some concern, then, that Section 4.3.3 concluded that sampling error (statistical significance, confidence intervals etc.) are, at best, limited and misleading measures of uncertainty. It is valuable to connect the issues of interpretation and uncertainty with the issue of bringing technical and professional expertise to bear on the available evidence, as discussed in the concluding section to Chapter 3 (Section 3.5.5). This is discussed at greater length in relation to league tables and practical use of value-added in Section 7.3, below.

The conception of validity captured by the argument-based approach to validation and the consideration of IUAs and validity arguments underpins the remaining sub-sections within Section 7.3 and leads toward the thesis’ overall conclusions. One final task before addressing specific areas of application, is to look at the validation of value-added specifically and outline a number of considerations which ought to be taken into account in order to validate claims based on value-added evidence.

Theoretical Framework for Validating the Interpretations and Uses of Value-Added Evidence

Throughout this thesis many distinctions have been made concerning value-added evidence and its uses. This section brings all these distinctions together to identify which are relevant for validating the interpretation and use of value-added evidence. This is not intended to be wholly comprehensive or discuss each issue at great length but to give an outline of the major factors which ought to be considered. The starting point for this is to define the process to be validated. The key aspects of this are summarised by the following diagram, giving the series of steps which must be taken in order to create and use value-added evidence, see Figure 7.3.2a below:

Figure 7.3.2a – Overview of Process for the Use of Value-Added Evidence



Stage 1 - Data

Data are collected which accurately capture pupil outcomes and all appreciable non-school factors. This involves sample or population data.

Stage 2 - Specification

A value-added model is specified to separate non-school factor influences from 'value-added' and appropriately specify the school/educational effect(s).

Stage 3 - Data Outcome(s) Produced

One or more statistics and/or graphics are produced to summarise value-added evidence for the user(s) in question (e.g. parents, schools, inspectors, researchers).

Stage 4 - Data Outcomes Drawn on by Users

Data users in a particular context draw on value-added evidence, possibly as well as other sources of data to address a given question or problem.

Stage 5 - Interpretation

Users interpret the outcome data produced to make inferences about effectiveness factors or the performance of pupils, schools or other groups.

Stage 6 - Use/Action

Practical decisions are made based on or informed by the interpretations of the value-added evidence provided.

There are some difficulties in following this right through to the sixth stage, the actual impact of the specific uses, as evaluating the efficacy of any action is likely to require consideration of a separate body of evidence and is, in many respects, a separate matter. Nevertheless, it is useful to retain this end point for two reasons (also see Kane, 2013, and Sireci, 2015: pp. 1-10). First, it is important to know the likely consequence of any interpretation of value-added evidence in any given school system even if it is difficult to know the precise effects of this. Second, the interpretations are not always made explicit and so an evaluator may wish to 'work backwards' from an action to identify the interpretations which are implicit within it. If the action and the data outcome are known, it is possible to consider the implicit interpretation,

Looking at each stage of the process in Figure 7.3.2a, the following is a list of factors which should be considered within a value-added validity argument:

Table 7.3.2a – Factors Impacting on Validity by Stage (see Figure 7.3.2a)

A. Data Quality

1. *The outcome measure*: Important considerations include the level of measurement error, content validity, presence of missing data, granularity, degree of isomorphism with the prior attainment measure and the possible presence of ceiling or floor effects.
2. *Non-school factor measures*: It is important to have collected suitable measures of all non-school factors (see next section). The accuracy, missingness, format and completeness of these variables should be considered.
3. *Effectiveness factors*: If the model is designed to estimate the effect of an effectiveness factor, a suitable measure must be obtained.

B. Specification and Modelling

1. *Selection of non-school factors*: Judgement is needed to identify which variables it is appropriate to control for. Inclusion of too many contextual factors will attenuate the value-added scores (Willms, 2003, OECD, 2008). Inclusion of too few will result in bias. There is no exact solution to this (Visscher, 2001). The choice of contextual variables will also depend on the ultimate use of the measure (e.g. whether the measure is used by parents for school choice or evaluation of school performance) (Raudenbush and Willms, 1995).

2. *Modelling of relationships*: It is important to ensure that the relationship between the outcome measure and all other measures is specified appropriately using a suitable technique. Consideration of non-linear functional forms to capture non-linearity will be required as well as considering whether models can produce unbiased estimates across the outcome and prior attainment measure distributions (Burgess and Thomson, 2013b).
3. *Identification of Problems*: During the modelling process various problems common to econometric models (such as multicollinearity, non-normality, heteroscedasticity) should be identified. The analyst should also consider potential problems such as missing data, measurement error and (possibly spurious) compositional effects.

C. Output and Presentation

1. *Level of aggregation:* A key consideration is the extent to which outcomes should be presented broken down by pupil groups and outcomes or, conversely, further aggregated by creating averages across outcomes or years. Data can be presented at more than one level.
2. *Provision of contextual data:* This includes VA estimates for other time periods, outcomes or contexts, as is deemed valuable. Issues of consistency and stability should be apparent to aid interpretation.
3. *Reporting of problems:* Potential problems should have been identified during the steps in areas A and B, above. This may need to be reported with the measure to enable informed interpretation.
4. *Reporting of uncertainty:* Similarly, some indication of the level of uncertainty should be provided (see below).
5. *Comprehensibility and detail:* There are various options for the form in which the data are presented. Data can be presented graphically, as individual statistics or in tabular forms.
6. *Accessibility and availability:* Consideration will be needed about where and how the data are made available and to whom they are accessible.

D. Use Context and Users

1) *User value-added knowledge and expertise*: Users of data will have varying degrees of understanding of value-added measures. Information may be required about the particular and some degree of professional and/or technical knowledge may be required (Visscher and Coe, 2003).

2) *User context knowledge and expertise*: Value-added evidence can be brought together with other sources of performance information (OECD, 2008). Some understanding of the school context will be of value to meaningfully interpret value-added evidence.
3) *Other sources of information*: The value-added evidence may be compared to other sources of performance information to support performance judgements. These sources of information may be more or less available (and high-quality) in different use contexts.
4) *Measure context and consistency information*: A key source of information is from other value-added outputs. Users may or may not look at estimates over numerous outcomes and years if the information is available (see last section).

E. Interpretation

1. *Data attitudes*: Literature about attitudes to data use was reviewed in Chapter 3. Attitudes such as unengaged, technist, sceptical and heuristic were discussed.
2. *Causal attribution*: Relating to the previous point, value-added scores can be interpreted with or without causal attribution to schools or with some interim position. If no causation is attributed, the only question is whether the differences observed are genuine differences rather than measurement error or bias. If causation is attributed, the user must additionally identify the causes of the observed differences in VA performance.
3. *Confidence*: All interpretations can be held with varying degrees of confidence. Confidence in terms of a range of possible performance but also the strength of support which the best estimate of performance can provide to any claim.
4. *Generalisations*: Value-added interpretations, like test score interpretations (Kane, 2013), frequently involve generalisation of the specific data to a more general concept. Can, for example, the VA outcome for one test, one outcome or one cohort be generalised to the school more generally?

F. Uses/Actions

1. *Consequences*: Interpretations matter due to the actions which follow. Where interpretations are mistaken, actions based on them may be harmful to a greater or lesser degree. These consequences may not be symmetrical (with positive/negative consequences being of differing likelihood or importance) and may not be linear (e.g. when accountability systems have cut-off points for intervention based on VA performance).

These twenty one considerations all have bearing on an interpretation-use argument (IUA) (Kane, 2013) which draws on value-added measures. This is not an exhaustive list but outlines many of the key concerns. This takes the process of creation, interpretation and use of value-added measures from start to finish. Depending on the context, different concerns will be more pertinent, particularly across markedly different use domains such as research or accountability use, for example. Following Kane (2013), an IUA should be constructed using the available information, using considerations such as those above as a guide. Kane (2013) describes the approach from this point as follows:

> "The validity argument then can evaluate the warrants in the IUA and the assumptions on which they depend. Some assumptions may be accepted a priori or be based on analyses of procedures (e.g., sampling assumptions). Some assumptions … may be accepted on the basis of experience, but any questionable assumptions will require new empirical evidence to be considered plausible. Strong claims (e.g., causal inferences or predictions of future performance in different contexts) typically would require extensive empirical support. The most questionable assumptions should get the most attention in the validity argument. For highly questionable assumptions, it is useful to consider several parallel lines of evidence."
>
> (Kane, 2013, p.14)

In the context of value-added, the assumptions requiring more extensive empirical support and parallel lines of evidence relate to causal attribution to schools attached to 'high-stakes' consequences. The more specific and the more high-stakes a causal attribution, the greater the evidence required. In practice, sufficiently strong value-added and supporting evidence may not be possible. Where causal attribution is not made or is only tentative, the main concern is the degree of precision. Estimates of the degree of measurement error in the underlying attainment tests (e.g. see Newton, 2013) and how error rates can translate to school-level within value-added calculations, such as provided in this thesis (Study 2), may inform a decision about the degree of precision which is warranted in any interpretations of value-added evidence.

As well as the stakes placed on causal attributions, one other important distinction to highlight is the unit of analysis (C1). Biases stemming from unobserved factors or measurement error are highly likely to prove problematic for certain groups of schools but, when looking across a large number of schools, this might not always disturb the substantive conclusions (but see Coe, 2009). It is also useful to note that even concerns within the same 'level' can be markedly different in the demands made of the data available. Consider the following concerns, which all relate to school-level scores:

1. Comparing value-added scores for one or more schools on an individual basis.
2. Comparing value-added scores for theoretically important groups of schools (e.g. comparing the performance of selective schools to non-selective schools).
3. Examining the overall distribution of school-level value-added scores for a particular outcome of interest.
4. Estimating the relationship between overall distribution of school-level value-added scores and a number of measured characteristics of these schools (perhaps in the context of a multi-level model).

Each of these in different contexts will raise a different emphasis on the various threats to validity for value-added methods and quantitative methods more generally. When working with larger samples and larger units of analysis, for example, measurement bias and omitted variable bias will be a greater concern than random measurement error or 'sampling' error. Similarly, teacher-level or system-level analysis may lead to the diminution or amplification of various concerns, all of which have an impact on the validity of IUAs in this area.

With this broader conception of validity in mind and the theoretical framework outlined above, we now turn to consider the implications of the present and pre-existing results for the areas of application reviewed in Chapter 3. Sections 7.3.3, 7.3.4 and 7.3.5 (below) consider the implications of the results for interpretations and uses in particular areas. Each one raises and discusses several issues which are particularly pertinent to a specific area or somehow different from the general discussion above. Note that these sections do not actually seek to construct IUAs or validation arguments for them - this would require more space and context-specific evidence than is available. What they aim to do is to raise relevant issues which would need to be considered during validation, as raised by the present and pre-existing results. These sections feed into what are viewed as numerous inter-connected validity debates (validating particular

IUAs) across areas of research, policy and practice. These debates are not entirely separate and may be thought to occupy a position along a general continuum of interpretations/uses - maybe according to the strength of evidence required. Yet, as has been argued above, neither is it held to be the case that the particular circumstances are unimportant.

7.3.3 Implications for Policy Use of Value-Added Evidence

General Implications

As was described in Chapter 2, on the development of value-added measures, value-added measures play an important role in policy given their claim to identify more or less effective schools. The imminent adoption of the 'Progress' value-added measures at secondary level (DfE, 2014b) and primary level (DfE, 2016) as the headline measures of school performance could be viewed as the end point in a long series of policy moves over several decades and a manifestation of the general direction of thinking behind these. Specifically, a great deal of educational policy making is based on a quasi-market-based approaches to school improvement where market and regulatory forces are employed as an approach to bring about school improvement. Many of these mechanisms, to work effectively, require the identification of more and less effective schools. There are numerous concerning results presented in this thesis and in the pre-existing evidence reviewed in this area which have serious implications for any regulatory or quasi-market-based framework which requires a valid measure of relative effectiveness that can be used across contexts in this way.

The following sub-sections consider the role of school effectiveness measures in various policy areas, how to design and make best use of value-added measures to ensure that the IUAs are valid. This section begins by considering general issues with the quality of the available data (see Stage 1 of Figure 7.3.2a, above) which has relevance across several policy areas.

General Implications Relating to the Quality of the English Data (Stage 1, Figure 7.3.2a)

The results presented in Section 6.1.2, Study 1, concerned observed rates of missingness in the NPD. The results suggest that missing data rates are fairly low in the NPD for most of the variables considered. This evidence presented, however, is likely to be misleading without further consideration. For instance, recall that about 1% of pupils were found to have been missing the FSM variable in Study 1. While low overall, these rates could prove problematic given that rates of missingness are heavily concentrated by school type and in individual schools; moreover, pupils missing the FSM code are apparently a super-deprived group rather than being typical of either FSM-eligible or FSM-ineligible groups (see Gorard, 2012b, on both these points). Moreover, the opt-in system for FSM status registration has been widely criticised (Freddie, 2015) and raises further concerns about likely rates of misclassification. Even if the problems with observed levels of missingness and classification were addressed, one cannot assume that the variable is sufficient to capture the effects of poverty. This was discussed in Section 4.2.2 and the results shown in Chapter 6, Figure 6.1.1e, raise serious concerns about the observed FSM-attainment relationship in relation to school FSM proportions. The FSM variable is illustrative of the difficulties within the available data, even in high-quality sources such as the NPD. There are also concerns relating to variables of theoretical importance which are missing entirely. However serious, these problems are currently academic given the policy decision to ignore contextual variables entirely (see below). Be this as it may, the most concerning problem with missing data found was in relation to attainment data; specifically, the rates of missingness in the KS1 data (around 1 in 20). These are concerning levels, especially given that it is highly likely that pupils who are KS1 missing data will be concentrated by school type, area or school.

This problem is further compounded by the examination of the attainment measures in RQ 1.2.3 (see Figure 6.1.2d). KS1 results do not appear to be a robust measure of performance. It is likely that this problem is not isolated to the English examination system, relating to the difficulties of creating robust tests for children of this age. Note, however, that it is possible to create robust tests for children aged 7 and younger (Tymms et al., 2004). KS1 assessment is likely to be a major source of the issues of consistency and stability found within primary-level school/cohort value-added scores (see Section 7.2.2, above). English KS1-2 value-added scores have been found to be biased by the omission of contextual factors, highly unstable (despite this omission), poorly reflective of school performance more generally (i.e. inconsistent across cohorts) and based on highly problematic underlying attainment measures.

One positive is that there were indications that the data were improving over time and may have improved further since 2012 (the KS1 extract examined). It is possible, therefore, that these problems may be gradually ameliorated by improvements in the data over a number of years. This is something which is likely to continue up to some practical limit of what it is possible to capture using numerical social science data (see Chapter 4, Section 4.2.2). Also, this may take some time as any improvements to baseline results will not be realised until pupils reach the endpoint for the value-added measure. Further research will be needed to identify whether continued improvements in the data have indeed taken place and thereby reduced these problems.

While the KS1 attainment measure has been mentioned specifically, Study 1 (RQ 1.2.3) also found sizable ceiling and floor effects within the KS2 (and other) data. These have been a long-standing concern in the English data. Kelly and Downey (2010), for example, point out that approximately one third of the pupils in the CVA pilot got the top grade at KS2 (also see Tymms and Dean, 2004). Problems with the KS2 data are also problematic for the KS2-4 measure, which uses the KS2 results as a baseline. Other results which raise concerns are the results in Study 2 that found that rates of progress in year 6 (when the KS2 results are taken) were an anomaly, with far higher rates of progress on average than all other surrounding years. It is simplistic to assume that the KS2 results cannot be artificially inflated by test preparation or that the predictive power of the KS2 results is independent of the extent to which scores reflect test-specific preparation or an unusually intensive 'cramming' of the content prior to the examination (Stobart, 2008). As with other problems, it cannot be assumed that these practices average out at school-level. Secondary schools are likely to be considerably (dis)advantaged by the approach taken to test preparation in their feeder primary schools.

Implications for English Performance Tables

Publication of value-added scores in the school performance tables is the most problematic use of value-added measures in the English system (high-stakes uses of teacher-level value-added could be considered to be more problematic if other systems internationally were considered). As a result, this particular use is discussed at greater length than others.

There is already a body of research in this area considering various aspects of the publication of measures of school performance (e.g. Foley and Goldstein, 2013, Dearden et al., 2011a, Leckie and Goldstein, 2011, Allen and Burgess, 2011, Visscher, 2001). Visscher (2001) reported that while experts were unanimous on the limited value of unadjusted examination scores, they raised numerous issues for public school performance measures: these include the level of uncertainty in school scores; the difficulties in meaningfully ranking schools; the 'mean-masking' problem (i.e. where mean scores do not reflect the performance of all pupils, such as disadvantaged students) (Teddlie et al., 1995, Wilson et al., 2008); problems of student mobility; the limitations of relative measures and issues of interpretation, understanding and usage (Visscher, 2001). As noted, there is a large amount of literature on the provision of performance information and it is not the intention here to repeat this; rather, the aim is to consider a number of specific issues raised within or informed by the present results.

Contextual Variables and Model Specification (Stage 2, Figure 7.3.2a)

First let us consider the issue of the omission of contextual variables in the English current VA and Progress 8 specification. The findings show a number of predictable associations between characteristics of schools' intake and the value-added scores, seriously undermining the claim that the VA is a fair and valid measure of school effectiveness. This situation is hard to justify given that variables used are readily available in the NPD and could be taken into account, as in the former CVA measure. The politicisation of this methodological issue appears to stem from several sources: first, a confusion between statistical and pedagogical expectations; second, the view that it is the standards and expectations set by policy-makers and regulatory bodies which are the key 'driver' of performance; and, third, the political value of the policy-change as a 'gesture' to signal the government's ethos. The real choice in taking contextual variables into account is whether one wishes to make the measure identify and reward schools which have been able to overcome difficult circumstances or to punish schools who are unable to entirely do so. English policy makers have chosen the punitive option. It may well be true that schools respond to disincentives related to being identified as a failing school more than incentives linked to being identified as successful in a difficult context. However, the credibility of the measure as a fair measure of performance is reduced by such blatant associations with intake characteristics.

It is worth making a clear distinction between statistical and pedagogical expectations. The latter are cultural and can reflect any level of aspiration, whereas the former merely reflects the status quo and, crucially, changes as the situation does. If schools in challenging areas were able to emulate the success of schools that have been able to counter the negative effects of educating in such an environment, the measure would adapt to reflect the reduced link between intake characteristics and attainment. Moreover, adjusting expectations according to prior attainment but not contextual factors is to misunderstand the correlational nature of the exercise. Why do we not also consider it 'wrong in principle' (DfE, 2010, p.68) that we have lower expectations of pupils who have performed poorly in earlier key stages, especially given that these differences are strongly related to social class, gender and race in the first instance? But, similarly to the principle about ignoring contextual factors, ignoring prior attainment would make little sense if we were genuinely interested in making fair and informed judgements about school performances in their given context. It is also important to stress that school value-added scores have not even been found to be independent of prior attainment (see Study 1), so not even this can be claimed. While this particular problem is less serious and may not have been expected by policy-makers, it is similarly difficult to justify ignoring it. If schools are not to be judged according to the prior attainment of individual pupils, why are they judged according to average prior attainment of their pupils? Similarly, the grammar school 'effect' is highly consistent and the evidence strongly points to this being due to measurement error rather than a genuine effect (see Section 7.2.2). Even if this was thought to reflect greater effectiveness, given the systematic nature of the effect, the claim that value-added measures are levelling the playing field when judging school performance is seriously undermined.

Conveying Differential, Unstable and Inconsistent Effectiveness (Stage 3, Figure 7.3.2a)

Next, let us consider the issue of differential school effectiveness and how this should be reflected in the data outcomes. In relation to consistency across pupils, there are several levels of generalisation which take place in the current performance tables: from individual pupils to ability bands (low, middle and high), to the entire cohort, to the school's performance (at least implicitly). The results clearly show that single measures seeking to generalise performance are problematic. The question of how to present these differences has been considered in previous research (Allen and Burgess, 2011, Dearden et al., 2011a). A key message of these, as reflected in the English performance tables, is that it is valuable to present results for different ability groups. The argument goes that this provides an incentive for schools to focus on the performance of all pupils and that parents can better assess whether a school is suitable given the attainment of their child. Similarly, the new Progress measures are designed to reflect the progress made by all pupils. The present results question whether this is sufficient. Even when broken down by attainment groups, there is considerable variation in pupil outcomes. Much information is lost when presenting value-added outcomes as cohort or attainment group averages, a 'mean masking' problem (Stringfield and Herman, 1996, p.168). It is worth pointing out that, despite the apparent enthusiasm for doing so, there is no technical requirement to summarise the output of value-added estimates for groups or cohorts using a single number

(OECD, 2008). One simple solution to this is to create pupil-level value-added scores and present the proportion of pupils scoring in a number of performance bands. This can be further broken down by sub-groups if desired, as shown in Table 7.3.3a, below:

Table 7.3.3a – Example of an alternative to presenting mean scores in light of high levels of inconsistency

	Pupil Performance Bands				
	Greatly below expectations	Below expectations	Broadly as expected	Above expectations	Greatly above expectations
Overall	4%	25%	50%	15%	6%
High attaining	Etcetera ...	...	...	...	...
Medium attaining	...	...	...	...	...
Low attaining	...	...	...	...	...
Pupil Premium Eligible	...	...	...	...	...

There are numerous advantages to presenting value-added outcomes in this way: first, all pupils matter to a greater degree than in an overall mean score. When using group means, a good school can afford to 'let down' small numbers of pupils for this to be masked in the average scores (Wilson et al., 2008). Second, this is simple to interpret and allows the bands to be chosen to communicate pupils’ scores in relation to expectations: what is known about measurement error and the substantive significance of the differences on the outcome scale can be used to set appropriate thresholds for what constitutes being above and below expectations. Third, this approach clearly communicates the differential nature of school effectiveness across different groups and the extent of within-school variation. This has been an ongoing problem with the presentation of value-added scores and the difficulty balancing complexity and interpretability (Kelly and Downey, 2010, Allen and Burgess, 2011). Fourth, this prevents the need for confidence intervals which are essentially presenting the same information but in a way which creates technical barriers to understanding and is highly misleading (see Chapter 4, Section 4.3.3). Fifth, these tables can be extended to consider any groups of pupils (as shown) and also different outcomes. The evidence reviewed in Chapter 4 suggested a moderate correlation across academic outcomes and almost no link between academic outcomes and wider outcomes of schooling. This inconsistency poses serious challenges for the interpretation of value-added scores (Sammons, 1996). Differences relating to outcomes, pupil groups, cohorts and stages of schooling appear to be an important part of an adequate understanding of the measurement of school effectiveness (Thomas, 2001, Chapman et al., 2015). It is important that these differences are apparent whether they are interpreted as evidence of differential effectiveness or as reflecting unreliability, error or bias in the measure (see Chapter 4, Section 4.3.2). This complexity renders crude accountability judgements between overall school performance of little value. Presenting these differences as above is considerably clearer than as a series of means with confidence intervals or, worse, not at all.

The other difficulty which must be addressed is the instability in the results. Again, it is possible to present results over several years in the same place using tables such as in Table 7.3.3a. Another option is to create measures which average scores over several years which smooth over some of the volatility, as was suggested in the initial reports looking into VA (Fitz-Gibbon, 1997) and is currently practised in Wales. This would prevent over-interpretation of short term fluctuations and a more long-term view of school performance. It is concluded here, however, that it is better to expose instability and inconsistency than mask it in smoothed figures. A similar line of thinking could be applied to differences between outcomes and across pupil groups. Hiding the instability and inconsistency of the measures is certain to lead to highly mistaken interpretations of the measures.

One problem which changes in presentation, such as the one suggested, cannot address is the lack of consistency between performance of a given cohort and other year groups within the school at a given point in time (see Study 4). This is especially problematic as the results presented in the performance tables concern a cohort no longer at the school, having completed the final examinations the year before (Leckie and Goldstein, 2009). This inconsistency at a point in time is difficult to show by way of further or alternatively presented data as concurrent performance data for different cohorts in a school are not collected.

Conveying Uncertainty and Preventing Misunderstandings (Stage 3, Figure 7.3.2a)

It is maybe defensible to suggest that the limitations of value-added measures could be understood by professionals and the data used cautiously within the context of first-hand experience and other data sources (see below). When publishing this information to the public, however, the interpretation and use of the information is out of the hands of the creators of the measures, and so misinterpretation and misunderstanding by some – given the numbers involved – is a practical certainty. It is unlikely that the concerns about the validity of value-added examined throughout this thesis can be adequately communicated to a lay audience to any great degree. This has been a common concern in the literature on the publication of performance indicators. There have been repeated calls for ‘health warnings’ to be included in school performance tables (e.g. Visscher, 2001, Foley and Goldstein, 2013). If valid interpretations are to be reached from value-added data presented in league tables, some understanding of how value-added measures work and the difficulties with the method is required by the user. This is a considerable challenge given that VA measures are "- for all but statisticians - obscure and mysterious" (Allen and Burgess, 2011, p.254). At the very least, the league tables should urge extreme caution and point out the major areas of difficulty (such as causal attribution and measurement error).

Currently the only indication that the estimates have any degree of uncertainty is the presentation of confidence intervals. As was discussed in Chapter 4 (Section 4.3.3), however, this is inadequate and even inappropriate. It is doubtful that the difficulties of interpretation and uncertainty (statistical or otherwise) can be adequately accommodated within a technical framework and then expressed within a measure of ‘confidence’ for public or professional consumption. Biases stemming from unobserved non-school factors, non-random measurement error, policy-decisions to ignore contextual factors, inconsistency across outcomes, missing data and many more threats to validity considered in Chapter 4 and examined in the results simply do not enter into the calculations. Yet, DfE guidance describes confidence intervals as ‘the range of scores within which each school’s underlying performance can be confidently said to lie’ (DfE, 2014a, p.9). This guidance is highly misleading and should be amended. A real danger is that the limited test provided by confidence intervals will distract from the numerous non-technical difficulties discussed and lead to misplaced ‘confidence’ in the results. It might be that confidence intervals, far from urging greater caution and awareness of uncertainty are actually performing a rhetorical function of putting ‘statistically significant’ results beyond professional dispute or debate. At best, all confidence intervals can achieve is to give some indication of the level of inconsistency in the pupil-level results. This is a very generous interpretation and, given that alternatives are possible to convey inconsistency (see above), there is little to recommend the use of confidence intervals even for this limited purpose.

Another major area for potential misinterpretation, aside from the threats to validity, is more basic and relates to the scale. As noted by Leckie and Goldstein (2011, p.209), concerning the former CVA measure, “no attempt [was] made to communicate to users the units in which CVA scores are measured.” The new Progress measures are an opportunity to fix this problem. It is essential that small differences between schools are not over-interpreted because of a failure to understand the scale of the differences. This is another problem which could be addressed through presenting the results as shown above. One final point to mention is the practice of adding a large positive number to the school VA scores. This is highly misleading and makes it very difficult to understand that it is a relative measure. If users have not understood that value-added is relative, this is a problem to be fixed rather than disguised. Happily, indications are that this will be addressed in the new measures and the national baseline will be left at zero (DfE, 2016).

School Inspection

The results suggest that inspection judgements should not be heavily based on value-added data. Nevertheless, this section considers a positive case for their use within the inspection process. As was suggested in the above discussion, the beneficial use of value-added may be possible in a professional or specialist context in which there is an awareness of the various threats to validity and understanding of how to use the measure as part of an informed and contextualised discussion. That is to say that the limited evidence provided by value-added, along with an understanding of the major threats to validity, are brought together with many other sources of evidence to reach the most valid inspection judgements possible (as recommended in Evans, 2008). Given that this is held to be one of the more credible uses of value-added (in which valid IUAs are more likely to be possible) the general problem which must be solved and the role of value-added within this is discussed in this section.

The basic difficulty with using value-added to inform decisions about school performance is the actual use that can be made of uncertain evidence. At one extreme, we could simply ignore value-added evidence given that it is liable to mislead us. But this will throw away potentially useful information. At the other extreme, we could take all value-added evidence at face value and use it as a primary basis for judgements. What is the best approach to steering between these two courses to make valid use of the information available? Chapter 3 discussed differences in attitudes towards data use by practitioners. This discussion is also highly relevant to school inspectors who, like school leaders, need to reach valid judgements about school performance. The general approach advocated here could be described as data-informed rather than data-driven (Murray, 2013). Another way to put this is to draw on distinctions reviewed in Chapter 3, Section 3.5.2, between a literal, 'technicist' use of data (i.e. data-driven) and a provisional, 'heuristic' approach (i.e. data-informed) (Saunders, 2000). It is the data-informed, heuristic position which is most likely to successfully navigate the numerous threats to the validity of value-added evidence.

It is important to understand what value-added scores are: namely, unexplained differences in performance with an appreciable level of measurement error. This understanding should be the starting point for any use of value-added scores. Many, if not the majority, of observed differences between school value-added scores are likely to reflect measurement error and non-school differences rather than genuine differences in performance. As a result, it makes more sense to use value-added to a) identify particularly ineffective schools for further inspection and b) as a very rough boundary for what is credible in an inspection judgement about effectiveness. For the vast majority of schools this will indicate that the differences between school performances are fairly small and most schools are performing as would be expected given their intake. Indeed, a key danger is to overplay the magnitude of any differences (see last section) and to fit differences in performance into pre-existing conceptual frameworks about what causes good or bad performance. Before drawing firm conclusions about differences in school effectiveness, inspectors should ask, 'is this difference sufficiently large to rule out pupil intake differences and measurement error as a cause?' The general answer to this will be 'no' and this should be reflected in more tentative judgements.

One final point to note is that inspections are not currently carried out on an annual basis. An inspection report remains current for a number of years. Yet, the value-added scores presented on the performance tables have been found to exhibit considerable levels of instability. Both of these cannot be simultaneously valid: either the Ofsted reports only have a 'shelf-life' of 1-2 years, or the changes in performance indicators in the performance tables should be disregarded until the longer-term trend is clear. If changes in value-added are considered to reflect changes in performance, there is a case for annual amendments/updates to be made to the most recent inspection report to reflect this.

System Monitoring

Use of value-added scores in a system monitoring role is readily defensible. This use can combine expert interpretation with low-stakes but potentially valuable uses such as conducting further investigations or identifying potential problems. What proportion of schools will have performances clearly outstanding from the overall distribution and so flagged for subsequent analysis or action is debatable, but it seems sensible that a monitoring process should take place to identify such cases, however high the threshold is set. Use of value-added as a system monitoring tool is supported by researchers such as Goldstein et al. (2000, p.27), who conclude that value-added is best used in a 'formative' rather than a 'judgemental' role (also see Foley and Goldstein, 2013). Support for this use of value-added is unlikely to attract much contention, although the specifics of this monitoring may be of interest. It is worth briefly mentioning several issues and considerations here.

The major limitation of a value-added-based monitoring system is that only relative performance estimates are produced. Based on the results in Study 2, the regression discontinuity design, coupled with national samples of performance across all NC years would be a suitable approach to providing such absolute performance estimates. The key factors to achieve this would be a sufficiently valid attainment measure (which can be used across more than one year) and a sufficiently large and representative sample to generalise about levels of performance across the national system.

7.3.4 Implications for Practice

Practical use of value-added is an area which is very difficult to comment on in specific and concrete terms given the relative scarcity of research evidence in this area. As discussed in Chapter 3, there are many different data services available to schools. Information is not readily available on many of these. Much of what is known about practical use of value-added data is presumably in the form of the (possibly tacit) understanding of experts who are experienced with the use of data in practical educational settings (but see Kirkup et al., 2005, Demie, 2013). As a result this section briefly discusses issues related to using value-added data from a more general perspective, considering the importance of combining technical and professional expertise to make judgements and how to use value-added for diagnostic rather than evaluative purposes. In terms of the framework in Figure 7.3.2a, above, the focus is on the final stages of the overall process relating to context, interpretation and use. A thorough consideration of IUAs in practice would require scrutiny of the underlying data, modelling decisions and the output provided, as has been done for the official English data above. Although note that many data services are based on the centrally collected English examination and census data (e.g. RAISEonline) (see Chapter 3).

Initially, let us consider the use of value-added to get an indication of the overall performance of the school. Where overall judgements are sought, much of the discussion above about inspection judgement applies: care is needed when interpreting uncertain data and over-interpretation of too small, too uncertain or too few data points should be guarded against. The main area of difference in terms of the IUA outlined above relates to differences in users of the measure. It may be possible for inspectors to be selected and trained such that a relatively strong grasp of the value-added method and its limitations could be widespread (but see Waldegrave and Simons, 2014). It is not clear whether this will be possible for all schools and so a loose conjecture is that technical understanding amongst school leaders will be ‘patchier’ and depend on the leaders in question. Value-added indicators are difficult for non-experts to understand and interpret (Visscher, 2001), particularly more complex CVA measures (Kelly and Downey, 2011a) and so there are some technical obstacles to interpretation. This difficulty relates to the extent to which a value-added score can provide ‘self-contained’ evidence and to what extent other sources of evidence should be relied on. Trying to place too much pressure on uncertain measures of a complex phenomenon and trying to treat all problems as technical problems creates considerable tensions, promoting ever-more complex measures to more adequately reflect performance and the concomitant separation of the technical experts creating the measure and the educational experts making use of the measure. This problem is described very clearly in Kelly and Downey (2010):

> "[If what is measured] is accepted as being only a small part of the *education* they receive in school, and if 'the school effect' is anyway accepted as being relatively small, one must ask whether the obfuscation that results from the complexity of ever more accurate measures is worthwhile when ever-fewer people can understand and interpret the results… Despite [value-added's] complexity, even for an academic audience, it represents in some ways an inappropriate *over-simplification* of the nature of school performance...not [capturing] the differential effectiveness of schools across the range of prior attainment and across the various sub-groups."
>
> (Kelly and Downey, 2010, p.192 & 195)

It may be welcome, then, that the new Progress measures are based on a methodology which is relatively simple to understand, despite the problems with biases this brings. Such biases however are of greater or lesser concern depending on the use of the measure. Where the measure is used for high-stakes, evaluative purposes, these biases are unacceptable. When value-added data are treated merely as unexplained differences in performance (and this is what they technically are) one avoids mistaken causal attributions based on *a priori* assumptions about what value-added scores are measuring (i.e. school effectiveness). This puts the user in a position to ask what has brought about the changes. Bias is less of a concern so long as non-school factors are permitted as part of the answer (which might not be the case in ‘no excuses’ cultures assuming outcomes are fully under the control of schools). In terms of reaching valid interpretations of the data, school leaders are in a particularly good position as they can be expected to have an intimate knowledge of the context as well as being able to access a greater range of data sources to aid with decisions about school performance. Although identifying the cause of unexpectedly higher or lower performance is a difficult question, school leaders and teachers may be in the best position to answer it. In other words, providing value-added data to raise rather than answer questions (see Chapter 3, Section 3.5.2) is something which puts considerably less burden on the validity of the measures and should be encouraged.

One final point to consider in this section relates to the findings concerning rates of progress by year group and the quality of the English attainment data. The results have indicated considerable problems of reliability and inconsistency across and within cohort as well as over time. Given this, it would be unwise to place great weight on value-added evidence. The use of value-added (or, worse, raw progress data) to heavily inform decisions about teacher performance is highly concerning, especially as such decisions are linked with teacher pay (Evans, 2008). Moreover, other related results, such as the finding in Study 2 that rates of progress are consistently related with NC year group, should raise further questions about crude comparisons of progress and value-added to inform decisions about teacher or cohort performance (see Amrein-Beardsley, 2014, for further information on the use of value-added for teacher evaluation).

7.3.5 Implications for Educational Effectiveness Research

Introduction

This final subsection within Section 7.3 discusses the implications of this research for educational effectiveness research. It was concluded in Chapter 3 that school-level use of value-added was a minor aspect of the current research. Where school-level scores were studied it was generally for the purposes of a methodological study often related to policy uses. Moreover, given that the data analysed were an administrative dataset and one based on teacher-assessed data within English practice, the direct implications of the main empirical results are predominantly for policy and practical use of value-added. Despite these points of difference, there are many results and contributions in this thesis which have clear relevance for educational effectiveness research (EER). Many of the issues discussed in Section 7.2.2, for example, have general relevance to applications of the value-added method and so have implications for EER. These are not repeated here; rather, the aim of this section is to discuss implications of results of the methodological study in Chapter 3 and advance the discussion of the issue of interpretation in Chapter 4 (Section 4.3 in particular). What follows is discussion of three concerns which have been raised but not fully addressed so far: first, the difficulty of understanding the properties of the school effect; second, the difficulties of inference and justification, particularly in relation to causal claims; and, third, the proper approach to addressing validity problems.

Understanding the School Effect

Issues of interpretation are vital and were discussed at length in Chapter 4, Section 4.3. This brief section raises some final questions about school effects and their measurement.

First let us consider the magnitude of the school effect itself. Reynolds et al. (2012) reply to reoccurring criticism of EER that there is over-emphasis on schooling (rather than personal, social and cultural factors which influence attainment for example) by referring to effects reported in more recent research:

> "[More recent research] employs the most sophisticated methodological and statistical methods shows much higher 'effects of schools' than the early 12–15% of variance explained that the critics highlighted. Guldemond and Bosker (2009), for example, show school-level variance explained 30% to 50% of variations between students, and Luyten, Tymms, and Jones (2009) over 33%, both figures considerably in excess of earlier estimates and both similar to family background effects."
>
> (Reynolds et al., 2012, p.6)

Consider what it means to say that the schools effect on performance when examining two points in time ('traditional' value-added) are relatively small yet school effects on the growth trajectories across time are 'sizable' (Guldemond and Bosker, 2009, p.255). On what basis can we claim that the meandering trajectories found between two points represent effect? What have we 'explained' in our new formulation of the school effect, capturing more of the instability between these two points? The more variables one adds to a model and the greater functional flexibility allowed in the growth curves, the greater extent that the variance between these points can be explained (Gorard, 2011a). This is a very hollow use of the word explain. And it is a very misleading use of the word effect. How meaningful is it to say that performance in one school zigzags up and then down while in another it zigzags down and then up if the tendency is to reach a very similar end point? This confusion is built into the terminology of random 'effects', which tell us very little about what is affecting what, only where the unexplained variance lies. It might not even do this. If teachers were sorted to schools by effectiveness such that there were no differences between teachers in schools, but large differences in the groups of teachers between schools, this would – when one partitions variance to school and teacher level – be recorded as a school effect. All we can see is where the differences lie, not what is causing them. Yet such assumptions underpin most of what is known in EER. On what basis do Bosker and Scheerens (1989, p.747) assume that performance differences across cohorts (grades) 'are in fact teacher effects'? Also consider the prevailing view that EER has 'demonstrated not only that teacher effects tend to be larger than school effects but also that, in combination, teacher and school effects could account for a substantial proportion of the variance in student outcomes'(Reynolds et al., 2014, p.204). How are the words *account* and *effect* being used here? If school effects are small but teacher effects are large, this seems to lead us to conclude that teachers greatly matter but that it is rare for any one school to have large numbers of effective teachers. If there were a large number similarly effective teachers within a school, this would show up as a school effect. Surely the most effective schools in the country would contain teachers of consistent effectiveness. In this way, the school-level distribution suggests a boundary on the teacher effect as well as the school effect. Similar to the discussion of the school effect above, breaking this down into more levels and more time periods does not increase the substantive size of the effect once the inconsistency and instability has smoothed out and is analogous to the claims that there are 'larger' school effects in growth models. Correlations for the stability of teacher-level value-added scores are typically in the region of 0.2-0.4 (see Chapter 4, Section 4.2.5). Including these fluctuations in the statistical model may allow us to *account* for more variance, but whether these differences represent effects or indeed have any practical value is another matter.

Discussion of effects in terms of variance explained are similarly misleading in relation to school effects. It must be kept in mind that the intra-class correlation simply gives the proportion of the unexplained variance which is situated at a given level. As one explains variance by entering further variables in the model, accounting for fixed effects or measurement error, for example, the amount of variance to be explained falls. We are just cutting sections out of the 'pie' and then comparing the relative size of the remainder. The intra-school correlation will increase if a greater proportion of the newly explained variance is at pupil-level and it will decrease if a greater proportion is at school-level. In either case, the absolute size of the school effect is not getting any bigger. It is unhelpful to be moving between a vernacular understanding of a school effect and the specific school effect as given by the intra-class correlation. As with stability (see Section 7.2.2), it is often more informative to look at data rather than statistics. In the English data, VA and replica CVA measures produced school-level value-added distributions spanning from approximately -50 to 50, where a score of 48 represents a difference of 8 GCSE grades higher per pupil. If this represents the extent of school influence, this is certainly meaningful. But without these scores being stable, it is hard to argue that this is wholly the case. Given that they are model residuals, how can we know what they are? This brings us back to the issue of justification.

Inference and Justification

Problems of justification abound within EER. As was explored in Chapter 4, Section 4.3.2, "we simply do not know if [the residuals used in a value-added analysis] are error or effect" (Gorard, 2010, p.757). As was discussed, the favoured approach is to draw on evidence of consistency and stability and present a case based on 'circumstantial evidence' (Rutter, 1983, p.12). With any given evidence-base, one can only endeavour to draw the most plausible conclusion and express this with appropriate caution for the circumstance in question. The key decision, in my view, is choosing the acceptable standard of evidence required to make a claim of a given strength. If one was to take the default position that most differences represent error, this is likely to result in a large number of false negatives. At the other extreme, if one tended to assume that differences represent effects, the result will be numerous false positives.

Much rests on whether the scores are considered to be causal. As noted in Chapter 3, when reviewing educational effectiveness research papers, authors put very different amounts of emphasis on the strength of correlational results. There was generally some acknowledgement of the problem in all cases. The question is whether these caveats are sufficient. What is the role of researchers in demanding higher standards of evidence from policy-makers (Goldstein and Woodhouse, 2000) and to what extent should they demand it from themselves? Should policy recommendations be based on evidence with 'medium levels' of internal validity, as in Chapman and Muijs (2013, p.359)? It is important to recognise how fundamental the problem of causation is to the value-added method (Coe and Fitz-Gibbon, 1998). This thesis contains example after example where correlational evidence has placed severe limitations on the task of identifying the cause of observed differences. The whole framework is potentially very misleading in terms of what causes what. When we look at the factors associated with effective teaching in Muijs and Reynolds (2000, p.288) and find, for example, that a composite measure of effective teaching containing measures of behaviour management and interactive teaching, for example, ‘explain’ the value-added differences in pupil performance, how do we know that these outcomes are caused by teachers? Could components such as ‘classroom climate’ be caused as much by the pupils as the teacher (Willms, 2003)? These fundamental difficulties of inference do not prevent the authors proceeding to present output from 3 multi-level models and 6 complex structural equations models. Every variable considered raises the same problem: the ‘effect’ of gender, to take another example, may tell us the general relationship between (a crude measure of) gender on attainment and almost nothing on what causes it. As the new handbook notes (Chapman et al., 2015, p.96), we do not yet have enough understanding of why these differences are present. Is it enough to simply caution the reader to the problem, look for stability and consistency and continue to draw causal conclusions? This is a longstanding and unresolved problem:

> “Some school effectiveness researchers have acknowledged the absence of evidence about causality (e.g. Scheerens, 1992, p. 71; Gray et al., 1995, p. 221; Reynolds & Stoll, 1996, p. 104), but the impression often gained - even where the issue is mentioned - is that it is something of a technicality, rather than a fundamental flaw in the methodology of school effectiveness research.”
>
> (Coe and Fitz-Gibbon, 1998, p.427)

Chapter 3 found marked differences between the two educational effectiveness journals studied, it seems each community of researchers has responded to this fundamental issue in a very different manner. Only one example of an experimental study was found in the School Effectiveness and School Improvement journal, whereas the vast majority of empirical papers in the Journal of Research on Educational Effectiveness were based on active experimental or quasi-experimental designs. This is all therefore related to a wider debate about methods, causality and the role of theory and design in demonstrating this; these are issues addressed at length in methodological textbooks in EER (Creemers et al., 2010). For the reasons stated above, it is my view that many of the ‘advances’ discussed in Creemers et al. (2010) are, to some degree, advances down a correlational cul-de-sac in which the value of multilevel models, structural equation models etcetera are overplayed.

Technical, Practical and Theoretical Problems

An over-riding theme of the present results and the pre-existing evidence reviewed is that educational data are imprecise, incomplete and uncertain. Trying to bring more of this complexity within a statistical modelling framework carries costs of comprehensibility without solving the fundamental problem (above). This would maybe be acceptable if more complex models were making considerable advances dealing with issues of unobserved differences (Coe, 2009), problems of interpretation and causal attribution.

I do not agree with Muijs et al. (2011) that the weaknesses of measurement and conceptualisation in EER studies is 'primarily a technical problem with a technical solution.' While technical solutions and technical analyses of specific issues (e.g. Televantou et al., 2015) allow new insights on certain issues, the assumption that the key problems are technical, to be addressed post hoc by way of complex analysis is mistaken (Gorard, 2007). The construal of uncertainty in statistical terms is a clear manifestation of this mindset. The difficulty this presents for the field is that wider questions essential to its health but not amenable to technical analysis are downplayed. Moreover, the technical approaches are often a barrier to examination of less tractable issues. Whether sampling error is the only or even an appropriate approach to modelling uncertainty does not appear to be a subject of discussion, it is simply the field's modus operandi. Even researchers often have little understanding of distinctions such as model-based and design-based inference (Snijders and Bosker, 2011) and the conversation around philosophical positions about the proper way to conceptualise complex and probabilistic phenomena is narrowed and obscured by terminology such as 'infinite populations', 'super-populations' and 'virtual populations'. The use of the word 'population' if one means probabilistic generation mechanism is misplaced and misleading (see Chapter 4, Section 4.3.3). The mutual accusations between critics and researchers in the field of lacking basic understanding are not surprising when such issues and terminology are so obscured. Narrow and misleading use of words like confidence, uncertainty, populations, sampling error (and effect, explain, account – see above), need to be justified and the conceptual frameworks drawn upon (such as model-based inference) made explicit and transparent. It might well be possible for a precise technical use of these terms, distinct from vernacular usage, to prevail in the research community. I do not think, however, that researchers can ignore misinterpretation by researchers not within the core of the area or outside users such as practitioners and policy-makers wishing to draw on research evidence.

In sum, it is reasonable to conclude that value-added methods are capturing school effects and meaningful differences worthy of continued study. But it is unreasonable to assume that all which falls outside the technical modelling framework is a trivial concern; this is the McNamara (or ‘quantitative’) fallacy (Syverson, 2008): i.e. measure what can be easily measured, disregard that which cannot be easily measured or give it an arbitrary quantitative value and presume that what cannot be measured is not important or does not exist. Are differences between groups of teenagers reducible to a dozen variables? Even the most spectacularly complex models will inevitably fall short of capturing more than a crude model of reality (see Chapter 4, Section 4.2.2). The evidence suggests that more complex analyses are not the solution to dealing with many of the issues considered in this thesis. Continuing down a technical, model-based route using ever more complex models to solve these problems is to confuse accuracy and precision: these may refine the estimates, but major problems relating to unobserved differences will remain, potentially leaving the estimates some way off and causal conclusions weak (Coe, 2009). This may just have to be accepted as an inevitable methodological limitation if it wasn’t for the fact that there are design-based approaches to dealing with such unobservable differences and reaching robust causal conclusions. Experimental studies such as Antoniou and Kyriakides (2011) are a rare exception in the School Effectiveness and School Improvement journal. As Rutter (1983, p.12) noted, although ‘circumstantial evidence’ does often suggest causal effects, ‘firm and unambiguous evidence on causation can come only from experimental studies in which school practices are deliberately changed.’ The aim of the vast majority of educational effectiveness research is to reach causal conclusions about what works (Teddlie and Reynolds, 2000). It is valuable to continue to generate correlational evidence for this insofar as there are no alternatives and this can lay the groundwork for more robust designs; but it is also surely time to submit some of the effectiveness factors which have been identified to a robust causal test (Gorard, 2007) and see ‘firm and unambiguous evidence’ as the desired end product of the research process.

8. Summary and Conclusions

8.1 Chapter Introduction and Thesis Summary

8.1.1 Chapter Introduction

This is the final chapter in this thesis. The first section (Section 8.1) summarises the thesis up to this point. It begins by recalling the core question and the approach taken to addressing it (Section 8.1.2). This is followed by an overview of the key results from the four empirical studies and some summary comments highlighting other contributions made by this thesis to the overall topic (Section 8.1.3). The final subsection (Section 8.1.4) briefly discusses the main study limitations and areas for further research (Section 8.1.4).

The final section of the chapter and of the overall thesis (Section 8.2) presents the final conclusions. Drawing on the extensive discussion of validity in Chapters 4 and 7, Section 8.2.1 gives a concise answer to the core research question, ‘Are school value-added measures valid measures of school effectiveness?’ This is followed by Section 8.2.2 which summarises the implications of the results by way of a number of recommendations. The final section (Section 8.2.3) gives some concluding remarks on the overall problem which has been examined and the contribution that this thesis has made to understanding it.

8.1.2 Thesis Summary

This research has examined the validity of value-added measures of school effectiveness. Value-added models are extensively used in educational research, accountability systems and policy-making to estimate school performance (see Chapter 3). Value-added evidence is used to inform and underpin myriad research findings, policy decisions and high-stakes school performance judgements. It is cause for concern, then, that there are theoretical and empirical grounds to doubt the validity of school value-added scores as being unbiased estimates of the causal effect of schools on their pupils (Coe and Fitz-Gibbon, 1998, Gorard, 2010, Marsh et al., 2011, and see Chapter 4). School value-added is, by design, capturing the unexplained (residual) differences in pupil attainment after controlling for differences between pupils. One can never be sure what causes these residual differences and the extent to which they can be casually attributed to schools (Gorard, 2010). School value-added scores are known to be somewhat unstable, inconsistent across outcomes and subject to known biases. The value-added design, however, provides little clear indication of the seriousness of these problems and what their underlying causes are. As a result, interpretations of the evidence generated using a value-added method are underpinned by many assumptions about what constitutes error, bias and effect within the data. The extant evidence base does not provide a definitive test of these assumptions. These difficult interpretative decisions are made in the context of highly complex statistical models, for which the analyst must make many difficult technical and non-technical decisions to produce the value-added estimates.

There have been recent debates (Reynolds et al., 2012) concerning how best to understand error, bias and uncertainty within value-added estimates. These debates have not resulted in an overall consensus and there are many outstanding points of contention and unanswered questions which remain. The contested issues within these debates are in part empirical and in part theoretical. This distinction is reflected in the scope, focus and organisation of this thesis which, on one hand, presents empirical results from original analyses and/or based on an under-utilised data sources (i.e. the Making Good Progress data, see Chapter 5); on the other hand, the thesis has looked to reframe the overall problem to give a broad and nuanced answer the core research question. Throughout the thesis there are crucial sections which situate the problem in practical contexts, looking at the various uses of value-added (Chapter 3), reconsidering how value-added evidence is interpreted and how to deal with uncertainty (Chapter 4), identifying the key aspects of validity which require consideration when validating claims made using value-added evidence (Chapter 7) and considering the implications of the pre-existing and newly presented evidence across a range of practical applications (Chapter 7).

The key contribution of this thesis to this debate - its ‘value-added’ – is that it brings together, updates, advances, synthesises and evaluates a large range of evidence and many fundamental methodological ideas within what is a large area of enquiry. This range and depth of study is what is held to be required to advance what is a complex, longstanding and unresolved question.

8.1.3 Key Results and Other Thesis Contributions

The four empirical studies presented and discussed in Chapter 5, 6 and 7 constitute the major empirical contribution of this research. Each of the four empirical studies addressed a particular problem: bias and error, inter-method reliability (against a quasi-experimental design), stability over time and consistency across cohorts. Using two large and relatively high-quality data sources allowed each study to examine specific validity problems and address questions arising from the literature review. The results have particular relevance for English policy and practice given the data sources used. There were also many more general findings due to, first, the fundamental nature of many of the issues examined and, second, the likelihood that other research and international policy contexts will face similar issues related to data quality, model specification and use.

Below are the ‘headline’ empirical findings of this thesis. Further details of more specific results underpinning these are given in Appendix F1. Also, note that there were numerous more fine-grained findings discussed in Chapter 6 and 7 which form or add nuance to these summary results.

Table 8.1.3a – Headline Thesis Results by Study

Key Findings of Study 1 - Bias and Error

1. English School VA scores, despite the method controlling for prior attainment at pupil-level, are not wholly independent of prior attainment at school-level.
2. Failure to include contextual variables in the more recent English VA measures has resulted in a number of substantial and systematic biases related to intake characteristics. These are consistent with factors found in school effectiveness literature.
3. There are several specific problems within the National Pupil Database that pose serious threats to the validity of school value-added measures. These relate to both measures of attainment and other contextual variables
4. A simulation of pupil-level random measurement error suggests that even random measurement error will translate to substantial school-level errors. This is especially the case for KS2-4 value-added and is likely to apply to VA measures more generally.

Key Findings of Study 2 - Absolute School Effects and the Regression Discontinuity Design

5. Progress is heavily patterned by year group in the English system. Year 6 results (as used in KS1-2 VA and KS2-4), for example, are considerably higher than other years.

6. Value-added is successfully capturing differences in progress between pupils. At least to the extent to which the underlying measure of performance can accurately capture these.
7. The regression discontinuity design is not suitable for comparing the effectiveness of different schools on an individual basis.
8. The regression discontinuity design shows considerable promise as a way of monitoring or comparing systems or large groups of schools.

Key Findings of Study 3 - Stability over Time

9. Secondary school value-added scores are moderately stable over time. Rates of stability have been inflated by failure to account for contextual variables.
10. Primary school value-added scores have moderate to very low stability over time. Correlations drop very quickly when at performance looking 1, 2, and 3 years apart.
11. Instability is not strongly linked with initial poor performance. It is a general characteristic of the value-added scores across the performance range.
12. Cohort performance over time has moderate levels of stability.

Key Findings of Study 4 - Consistency within and across Cohorts

13. Consistency between the performances of different KS3 year groups in the same school at a point in time is moderate.
14. Consistency in the performance of different KS2 cohorts in the same school at a point in time is moderate for adjacent cohorts and low to very low for cohorts 1 and 2 years apart.
15. School value-added scores mask very large differences in pupil performance for schools across the performance range.

8.1.4 Study Limitations and Areas for Further Research

Limitations of the scope (Chapter 1), data (Chapter 5), analysis and results (Chapters 6 and 7) have been discussed within the relevant sections in the thesis. The conclusions reached above are held to be justified given the evidence presented in Chapter 6 and material reviewed in Chapter 4. There is always a difficulty of over-generalisation and over-drawing of conclusions from insufficiently indeterminate evidence; this must be offset against being too equivocal and not reaching practical conclusions or dismissing potentially valuable evidence due to concerns over robustness. The reader can judge whether the conclusions of this thesis are justified given what is presented and can consult the methods chapter and appendices for detailed information about the data and analyses used. It is, nevertheless, valuable to briefly highlight and comment on some of the more substantive limitations and make these as explicit as possible before moving to the final conclusion section so that the study conclusions can be viewed in light of these concerns. Three areas of concern are discussed relating to the data used, analytical decisions and limitations of scope.

The results in Chapter 6 are based on analysis of two data sources. Both of these concern the English education system. The advantage of this is that it makes the results particularly relevant to the English system. The disadvantage is that it makes generalisations about value-added more problematic. The major source of difference is likely to be the examination system: the quality of the attainment measures was found to be of paramount importance and so systems or research based on more or less robust attainment (and contextual) measures are likely to give considerably different results. Another key point is that the data set used for performance estimates between key stages (the MGP dataset, see Chapter 5), is based on teacher-assessed data. Similarly, this can be seen as an advantage in terms of making the results more relevant to the (predominantly) teacher-assessed tracking data used in English schools, but raises questions relating to generalising to examination assessed data in England or more generally - although note that teacher-assessed results constitute all or part of the KS1, KS2 and KS3 assessments in England. The most likely effect of using teacher-assessed data is to lower the consistency and stability estimates based on these data. Further problems with the MGP data was that secondary-level results only went up to year 9 (age 14) and there was a reduction in sample size in the final year. Specific limitations of the NPD data used were that the most recent data used was for 2014 and for many of the pupil-level results the most recent available extract obtained related to the 2013 and 2012 data. Given apparent improvements in the data over time and this being a period of significant examination reform, even these recent estimates will differ to some degree from current data.

Many of the analyses involved the straight-forward reporting of specific aspects of the data to explore the properties of the measures. There were some analyses which involved making analytical decisions which are possibly questionable. The most contrived analysis (and also one of the more original) was the simulation of error within Study 1. There is likely to be some disagreement in what constitutes a small, medium and large error rate; indeed, these levels were included to explore rates lower and greater than the author's 'medium' expectations. It is very difficult to know actual error rates and how these are distributed across pupils and schools. Errors are unlikely to be random so the use of random errors is somewhat generous; but introducing systematic errors would defeat the aim of the simulation (to see if pupil-level random error could translate into more systematic school-level error). The simulation attempted to construct data with realistic properties while isolating the effect of error being investigated. There are of course limitations in transferring lessons of a hypothetical situation (of a deterministic relationship in which random error is introduced) to actual value-added measures and data which are wholly naturally occurring (rather than partly constructed). Other analytical decisions in other studies which are noteworthy include the specification of measures where the official measure was not used or replica CVA measures were generated. The intention was always to strike a balance between over-complexity and leaving problematic rates of non-school factor bias. These decisions are unlikely to affect the substantive results. One decision which may be more consequential is the decision to restrict the MGP analysis to mathematics results. The mathematics results were found to be more consistent with exam-assessed performance. The consistency and stability of teacher-assessed reading or writing are likely to be lower than those reported for mathematics.

Limitations of scope were outlined in Chapter 1. This research has had a broad scope and considered many aspects of the overall problem. The notable omissions are that the analyses did not look to break down measures by subjects. As suggested by the review of consistency in Chapter 4, if estimates for specific subject areas were provided rather than average point scores and composite measures, the stability and cohort consistency is likely to be lower. Also, the associations between contextual factors and attainment reported in Study 1 may have differed by subject area. Another major limitation was concentrating on school and cohort value-added measures. The data sources used did not identify individual teachers and so this important aspect of the use of value-added could not be explored. Teacher-level value-added was examined in Chapter 4, the results suggest that problems of instability, inconsistency and measurement error are likely to be considerable greater at this level of analysis than those presented at school or cohort-level. One final limitation of scope is that the evidence base on practical use of value-added in English schools was limited, despite some recent notable studies in the area making great progress to address the gap in knowledge. The research aimed to look at all stages in the process of using value-added described in Chapter 7, Section 7.3.2. But getting detailed information on use and interpretation in the practical context in particular was difficult. As noted in Chapter 1, the decision was made to concentrate on getting the larger picture and this involved accepting limitations of scope concerning highly technical details at one end of the spectrum and specific practical details at the other.

8.2 Conclusions

8.2.1 The Validity of Value-Added Measures

'Are school value-added measures valid measures of school effectiveness?'

The evidence that has been presented and reviewed is that value-added scores are comprised of a) differences in school effectiveness, b) unmeasured individual or cohort-level differences and c) measurement error within the underlying measures of performance. The results suggest that the latter two components are substantial and that assuming otherwise risks seriously misplaced inferences. The evidence suggests that there is a general level of imprecision, error and bias and this means that, at best, value-added scores are only a rough approximation of performance. However, threats to validity (such as omitted variable bias) are likely to affect individual school scores unevenly and so, for a portion of schools, value-added scores will be grossly inaccurate. It is difficult to know to what extent value-added scores do reflect the causal effect of schools rather than bias or error, for particular schools and in general.

At best, one can conclude that value-added scores contain an appreciable amount of error and bias and so need to be interpreted with caution alongside other sources of evidence. At worst, value-added scores may be so dwarfed by error and bias that, in the vast majority of cases, the scores will be highly misleading and any significant action taken as a result of value-added evidence will be ill-advised and potentially damaging. The truth is presumably somewhere between these positions. Value-added certainly cannot be characterised as a robust measure of school performance.

If one takes a sceptical position and seeks justification for why value-added scores should be interpreted as capturing school effects, it is quickly apparent that inferences have very little firm basis. Evidence of consistency or stability does not allay these concerns as, without knowing the underlying causes of the differences, stability could as easily reflect non-school factors as school factors. Even if one was to assume that consistency and stability were evidence of school effects, the evidence in this area gives further cause for concern: value-added has repeatedly found to be inconsistent across outcomes and somewhat unstable over time, particularly so at primary level. This research has also found that consistency is moderate to very low across cohorts within a school at a given point in time, confirming the concerns raised by the small number of early studies in this area. This all leads to the conclusion that value-added is seriously and fundamentally flawed as a measure of school performance. The extent to which the value-added method can isolate school effects from construct irrelevant variance (CIV) is questionable. Moreover, the latent and correlation nature of the evidence means that there is no way to separate systematic non-school factor bias from school effects (Gorard, 2010, p.758), or as Harker and Tymms (2004, p.195) noted when examining compositional effects using value-added evidence, '...the really worrying thing is that the researcher can never be sure about what has been found.'

A key basis for a more optimistic position is that value-added scores have been found here to be highly consistent with gain scores in longitudinal data and therefore do capture real differences in pupil progress, at least to the extent that underlying measures of performance can reliably measure differences in pupil progress. Difficulties of reaching and justifying interpretations based on value-added evidence does not mean that the evidence is meaningless or even that the value-added method has no value. The issue is whether or not one can reliably separate the 'signal' of differences in school effectiveness from the 'noise' of CIV and other confounding signals from non-school factors. This thesis has argued (see Chapter 7, Section 7.3.2) that the value of endeavouring to solve this problem and difficulty of doing so is best considered to be context-dependent.

In conclusion, school value-added measures are seriously and fundamentally flawed, but the design of value-added method as well as the empirical results suggest that value-added evidence could still be put to beneficial use in some contexts. Identifying what these contexts are and developing best practice for the use of value-added evidence within them is a challenging task. The approach which shows most promise for realising this aim is the theory of validation advocated by Kane (2013) and others; where one must ask whether the interpretation-use arguments (IUAs) which draw on value-added evidence are valid rather than ask whether value-added measures are valid independently of their interpretation and use. The problem must be (re)framed as a practical as well as a technical matter (Corcoran and

Goldhaber, 2013). Few of the threats to the validity of value-added are amenable to technical solutions and there is no entirely satisfactory solution to capturing and communicating uncertainty (see Chapter 4). Researchers such as Kelly and Downey (2010) are quite right to position the issue in this way, considering the limitations of the measure, the impact of school value-added measures on different areas and the competing demands of various uses:

> "Whether or not published VA scores are accompanied by confidence intervals, and whether or not they are published as true residuals*, they suggest a degree of precision in the measurement of school performance that is not justified. And, despite their complexity, the measures fail to respond adequately to competing legitimate demands: from the public for *interpretability;* from teachers for *usefulness;* and from policy-makers for *accountability.*"
>
> * (rather than shrunken residuals, see Chapter 2, Section 2.3.6)
>
> (Kelly and Downey, 2010, p.206)

All of these uses of value-added and those in a research context are usefully examined in relation to the validation of IUAs and the theoretical framework outlined in Chapter 7 (Section 7.3.2). In general, claims which are explicitly or implicitly based on the assumption that scores are highly consistent and stable are invalid. Valid IUAs will reflect the weaknesses and uncertainty in the measure, will not overstate how robust the evidence is in terms of causation, will not reduce school performance to a single measure, will not ignore the many other competing explanations for the scores and will not base high-stakes decisions on value-added evidence to any significant degree. These principles alone cast doubt on many current uses and suggest that some must surely be abandoned entirely. Remaining uses must ensure that IUAs reflect what is known about value-added evidence in general and in relation to the particular measure and data. The next section presents several recommendations concerning how to navigate the uncertainty within the measures to reach valid interpretations and states some of the more/less defensible uses of the value-added method.

8.2.2 Implications and Recommendations

General

Valid interpretation and beneficial use of value-added evidence benefits from adopting the following approaches:

1) *Avoiding causal assumptions*: Value-added produces correlation evidence, the difficulties of which are well known (Shadish et al., 2002) and are suitably recognised in many of the research studies reviewed (e.g. Vanlaar et al., 2013). It is only by assumption that value-added scores are attributable to school performance. This assumption is highly problematic. Strictly, value-added measures capture *unexplained* differences in performance. Using value-added to raise questions rather than give answers avoids the mistaken assumption that school-level variance can be safely causally attributed to schools or school factors (Coe and Fitz-Gibbon, 1998); although this means the one has all their work ahead of them to find out what causes any given differences.
2) *Putting Scores in Context*: One major source of differences captured by value-added scores will stem from the 'noise' of – inherently imprecise – educational measurement. The easiest way to be mistaken is to base too much on too few data points or solely use value-added evidence as a basis for inference. Longstanding advice from educational effectiveness researchers stresses the importance of looking across time, outcomes and pupil groups (Reynolds et al., 2012) and the limited value of narrower data:

> On the basis of existing research it is apparent that estimates of schools' effectiveness based on one or two measures of students' educational outcomes in a single year are of limited value. Ideally, data for several years (three being the minimum to identify any trends over time) and several outcomes are needed."
>
> (Teddlie and Reynolds, 2000, p.126)

Maybe the clearest principle for the valid use of value-added is that it should be used in a way that, if it were entirely meaningless, this would be apparent. One should compare value-added across various pupil groups, various subjects and several years. When interpreting the data (see point 2) one should bring value-added evidence together with other sources of information, professional understanding of the school context (and so possible biases) and technical understanding of value-added evidence. This approach is challenging but is the most likely way of reaching valid conclusions. Often value-added evidence is presented without examination of consistency and stability even being possible. Presenting mean values across subjects and cohorts, along with confidence intervals, for example, actively discourages this approach (see below). The primary danger is basing too strong a conclusion on too little evidence, or using additional data to confirm rather than critically examine any initial impressions given by the evidence (see next point).

3) *Exercising due scepticism*: The value-added method does not provide robust evidence. Therefore, as discussed in Chapter 7 (Section 7.3.2), without some ‘slack in the system’ (Kane, 2013, p.3) valid IUAs are impossible. Demanding certainty will lead one to reject all value-added evidence and forgo all potentially beneficial uses. Excess scepticism is not likely to be a major problem, however. As psychologists make clear, people need little encouragement to reach overdrawn conclusions on the basis of scant evidence and often work to confirm rather than scrutinise initial hunches so long as they have superficial plausibility (Kahneman, 2011). Individuals – expert or otherwise – who are actively seeking meaning in data (rather than asking critical questions about its trustworthiness) are prone to reach spuriously coherent judgements about school performance. Moreover, where care is not taken to avoid it, even use of data as described above can become an exercise in confirmation bias, where data are sought to confirm the initial hunches provided by examination of the value-added data. This is an especially serious concern in the context of inspection decisions or those related to teacher performance appraisal (see Kennedy, 2010, for further information on the difficulties of attributing learning to teacher characteristics, for example). The best principle to guard against this is to build attempts at falsification of the conclusions and consideration of alternative explanations into the decision-making process. As noted in Chapter 7, the general answer to the question, ‘could this school value-added score be the result of measurement error or bias?’ will be yes.

4) *Keeping Effects in Perspective*: One final noteworthy principle is that it is important not to lose sight of the scale of any effects. This is another area where the construal of uncertainty as statistical significance and the language associated with it is often counter-productive: there is frequent confusion between statistical significance and substantive significance, the actual size of the effect (in terms of the scale on question or an effect size) is often downplayed or ignored and it sidesteps consideration of measurement error and other threats

to validity (Coe and Fitz-Gibbon, 1998). The size of any effect is the most important piece of information for judging its importance and whether it is of a sufficient size to rule out measurement error (Gorard, 2006b). While this point should be fairly obvious, the current and former (Leckie and Goldstein, 2011) English school performance tables make no attempt to convey what for example a KS2-4 Best 8 value-added score of 981.0 means. It is highly unlikely that many users will access the user guidance and technical information to find out.

Research

Many of the general recommendations above can be adapted to specific areas of application. It is useful nonetheless to make a number of recommendations specifically for particular areas. Research is one application of the value-added which is more likely to result in valid interpretations of the evidence. Many papers reviewed in Chapter 3 and elsewhere recognised the limitations of value-added evidence and researchers tend to seek to find associations with educational factors and residual variance (value-added) rather than interpret the residuals themselves. This is an important distinction although it does not mean that estimates will be unbiased.

5) *Understanding error*: Errors should not be assumed to be random and, as Study 1 showed, even random errors can have systematic effects. There were many examples of papers reviewed which clearly discussed threats to validity (other than ‘sampling error’) and this is to be encouraged.
6) *Aiming for more robust designs (see Chapter 7, Section 7.3.5)*: At best, value-added produces correlational evidence. The difficulties of this are well known and have been clearly examined specifically in relation to school effectiveness research (Coe and Fitz-Gibbon, 1998). While correlational evidence can be valuable, especially in new areas, greater use should be made of more robust designs (Shadish et al., 2002), if causal claims are sought. As concluded in Chapter 7, Section 7.3.5, while it is valuable to continue to generate correlational evidence for this insofar as there are no alternatives and this can lay the groundwork for more robust designs; but it is also surely time to submit some of the effectiveness factors which have been identified to a robust causal test (Gorard, 2007) and see ‘firm and unambiguous evidence’ as the desired end product of the research process. While caveats about causation and threats to validity are appropriate and important, to

continue to make policy recommendations on the basis of such evidence (e.g. Chapman and Muijs, 2013) exposes researchers to accusations of ‘wishing to have their cake and eat it’ too (Goldstein and Woodhouse, 2000, p.354).

7) *Avoiding misinterpretation:* The widespread use of causal language to describe correlation evidence is often highly misleading. Constructing a case that evidence is causal is a difficult task (Hill, 1965) and unthinking use of causal language in the context of highly complex statistics builds inevitable misinterpretation into the methodological framework itself. This raises particular problems in educational effectiveness research where the use of variance partitioning to estimate school effects and the frequent conflation of accounting for variance within a statistical model and ‘explanation’ is widespread. These issues were discussed at length in Chapter 7, Section 7.3.5.
8) *Complexity is not generally the answer:* Technical complexity is another issue addressed at greater length in Chapter 7, Section 7.3.5. No statistical model, however complex, will adequately address the problems of validity discussed in Chapter 4. Advances in methodology (Creemers et al., 2010) should be judged against the robustness of the evidence produced (Gorard, 2013). Greater complexity without greater robustness is a sideways step at best. At worst, this narrows the research field, emphasising technical expertise at the expense of scientific, practical and philosophical expertise.

Policy

Policy issues are discussed at length in Chapter 7, Section 7.3.3. This section briefly reiterates and summarises key points:

9) *Linkage to rewards and sanctions:* Guidance issued by the Department for Education (DfE) (formally the Department for Children, Schools and Families) pointed out that information from value-added measures should be ‘used as a basis for discussion in school improvement and inspections, rather than directly driving any rewards or sanctions’ (Evans, 2008, p.21). This view, however, is not always reflected in guidance documents (e.g. DfE, 2013a) and the research clearly suggests that VA measures *do* play a decisive role in informing high-stakes decisions within the English accountability system (Bradbury, 2011, Acquah, 2013), as a basis for inspection judgements and in public understanding of schools via the school ‘league tables’. High-stakes uses of value-added are likely to have profoundly damaging effects given the likelihood of performance estimates being highly inaccurate for some

schools. Value-added is far better used to raise questions and create a rough boundary on what it is credible to conclude about school performance (see general recommendations).

10) *Taking context into account:* The omission of contextual variables introduces serious biases in the measure and does not seem justifiable if a fair measure of performance is sought (see Section 7.3.3). It should not be the automatic assumption that the staff or indeed any school-factor is the major cause of differences in performance, in fact this is one of the least likely explanations: Only around 5-10% of the differences between pupil performances are associated with school membership (Reynolds, 2008, Wiliam, 2010). This well-established and highly consistent finding is easily lost, in particular in the political rhetoric around standards. Anecdotes about schools beating the odds make poor policy, even if they make good rhetoric.

11) *Presenting information to the public:* League tables are a highly problematic use of value-added scores: placing large amounts of data online under the heading ‘school performance tables’ makes misinterpretation inevitable. The public presumably have a right to know about school performance however, so options for better presentation of the information or alternative options should be explored. For example, annual data reporting could become a function of the inspectorate, where annual summary data reports are produced and published to supplement periodical inspection reports. Although resource intensive, this would generate expertise in the inspectorate and feed into their core functions. If it is worth reporting school performance data to the public, it should be done in a way which encourages valid interpretation. The other option is to reform the performance tables and their presentation. There are numerous other reports and pieces of research looking in to how best to communicate performance within league tables (e.g. Visscher, 2001 , Allen and Burgess, 2011, Foley and Goldstein, 2013, Bird et al., 2005). Such reports contain excellent advice; although very little of this has currently been acted upon. A recommendation made here is the presentation of proportions rather than means, as discussed in Chapter 7, Section 7.3.3 (also see the general recommendations above). The issue of uncertainty is another key area where change is needed. There has been little to no attempt to communicate any of the issues discussed in this thesis in published performance tables or in guidance documents issued to school leaders such as those cited earlier. The only suggestion that the value-added measure may be less than entirely robust is the inclusion of confidence intervals in the

performance tables but even this is, at best, misleading (see Chapter 4, Section 4.3.3 and Chapter 7, Section 7.3.3).

12) *Accepting the implications of uncertainty:* Best practice for using data has been discussed (in Chapters 3 and 7) in terms of data-informed rather than data-driven decisions. As noted by Kane (see Section 7.3.2), uncertainty 'makes interpretations a bit fuzzy and decisions a bit tentative' (Kane, 2013, p.4). This thesis confirms that value-added evidence is highly uncertain and the majority of generalisations about school performance will be invalid. Yet the review of the policy use of value-added in Chapter 3 raises a serious concern as to whether tentative, multi-faceted and equivocal performance judgements are in keeping with the current English accountability culture. By international accountability standards, England is characterised as having a particularly 'high-pressure system' (Altrichter and Kemethofer, 2015, pp.50, also see Chapter 7, Section 7.3.3). Rating grades used by Ofsted and publication of value-added data in 'School Performance' tables are simplistic and misleading due to their failure to satisfactorily reflect and convey complexity and uncertainty. Complexity and caution is all but lost in the current English accountability climate and this gives rise to decisions which are not conducive to good policy decisions (also see point 10). What is needed is intelligent accountability (O'Neill, 2013) rather than over-reliance on performance measures, quasi-market mechanisms and tough accountability as a driver of school improvement. Within this, value-added scores may be better placed as a monitoring tool (Foley and Goldstein, 2013) for professional use rather than a 'free-standing' evaluative technology for public consumption.

Practice

The use of value-added by educational practitioners is the hardest area to make specific recommendations without further evidence. The main focus of this thesis is school-level value-added; yet for leaders and teachers working in schools, fine-grained monitoring information may be of greater value (Kelly et al., 2010). It is questionable whether school-level value-added scores can provide much information above what is already known; although it may provide a rough indication of the school's performance, preventing highly mistaken judgements about performance. A lot will depend on the specific provision (see Chapter 3) and its ability to promote valid IUAs in the given area (see Chapter 7, Section 7.3). Because of this, no further recommendations for practice are given beyond the discussion in earlier chapters and the general recommendations (above) which all apply to use of value-added data at more fine-grained levels than school-level.

8.2.3 Concluding Remarks: The Value of Value-Added

Value-added measures fall far short of what is demanded of them: rather than being valid, reliable measures of differences in performance which can be causally attributable to schools, value-added measures produce estimates of effectiveness which are approximate and uncertain based on 'messy' data typical of social research.

Given this, we may wonder why it is worth persisting with value-added at all. Yet this thesis, for all its doubts and criticism of value-added, has sought to reach a positive position which outlines principles for beneficial use of value-added evidence. A key part of this positive position is the view that it is best to hold a broad conception of validity which spans from examination of the data quality all the way to the uses of value-added evidence. Simply creating measures and hoping they are interpreted and used correctly is mistaken: Suppose that it was possible to 1) capture school performance within a value-added score and 2) satisfactorily convey the uncertainty of the measure (maybe by using a confidence interval) to users. If this were possible, the value-added score could be seen as a 'free-standing' product which gives an estimate along with everything else the user requires to sensibly interpret the estimate. As a result, producers of value-added evidence could present their estimates along with the measures of uncertainty and have no further role to play. This thesis concludes that a value-added estimate with known validity is a chimera. The producers of value-added estimates cannot reach a definitive position on validity, let alone the users. This all suggests that producers of value-added evidence must carefully consider how to best communicate the problems with the measure and encourage appropriate interpretation and use.

Throughout this thesis the discussion has moved back and forth between doubts about validity and reasons to be more optimistic, trying to grasp the best characterisation of validity. Rather than producing a precise answer, the evidence and analysis has suggested that there are a range of interpretations of validity which could be considered compatible with the available evidence. Because of this, some of the biggest problems relate to interpretation. In the absence of a definitive measure with a clear level of validity, there is scope for a user to 'hold whatever view of this world [s]he finds most agreeable or otherwise to his [or her] taste" (Galbraith, 1998, p.6). Advocates of data-driven accountability will take an optimistic view, opponents a pessimistic view. Poor performing schools will assume that their low value-added stems from unobserved non-school factors and high performing schools will, of course, credit their excellent provision. Encouraging and supporting users to reach valid interpretations may be more of a problem of shifting attitudes, knowledge and culture in relation to data than improving data provision and presentation.

Examining the process of using value-added more generally (see Chapter 7, Section 7.3.2) reveals just how difficult the beneficial use of value-added is likely to be. Every stage of the process is doubtful. Is the test a valid/sufficient measure of educational outcomes? Has sufficient non-school factor data been obtained? Is the value-added model specified correctly, in terms of the non-school factors and the nature of school effect(s)? Are the estimates presented such that users are able to interpret them correctly? Have users appropriately interpreted the estimates? Are the uses which are informed by the value-added evidence educationally valuable? One should not underestimate how difficult it will be to get this process right for any given use.

It is also important to question how essential value-added is to aid decision making. What is the specific role of value-added in any decision-making? By definition, the reason to create a value-added measure is to make comparisons between school performance which are not confounded by differences in intake or circumstance and so justifies a claim that school A has done better than school B (Gray et al., 1986, p.91). What, however, *is* the value in knowing a school's performance compared to other schools and why has such great store been placed in the relative rather than the absolute performance of schools? School value-added ('Progress') measures are about to become the headline measures of performance in the English school system. This places a zero-sum, relative (and problematic), measure at the heart of the system. While the evidence is clear that all schools have a massive impact on children's education (Luyten, 2006), the vast majority of differences in pupil value-added performance are within schools. Differences between schools, while probably an important consideration, are a relatively small part of the overall picture which shows that educational performance is brought about by a vast range of social, cultural, economic and individual factors (school membership accounts for about 5-10% of the variance in outcomes: Reynolds, 2008). Within this bigger picture, value-added has the specific aim of isolating small, *relative* differences in school performance from the myriad and complex combination of non-school factors affecting performance (Teddlie and Reynolds, 2000, Reynolds, 2008), all in the context of performance measures which themselves have a considerable degree of imprecision. It is no surprise that the result is unreliable and uncertain.

The value of data in general and value-added data in particular should not be over-emphasised and nor should the ability of schools to 'compensate for society' (Bernstein, 1970). However, there are some grounds for optimism as well as great challenges in both cases. The evidence strongly suggests that value-added cannot bear the weight of high-stakes accountability; yet, value-added may have some role in informing educational decisions and improving schools (Visscher and Coe, 2002). The value of data in improving schools is not entirely clear and it may be that its potential is yet to be realised. By examining the validity of value-added measures and considering how to work this understanding into practical contexts of use, this thesis is supportive of this endeavour and has thereby aimed to contribute to the realisation of educational aims.

References

Acquah, D. (2013) *School Accountability in England: Past, Present and Future*, Manchester: AQA Centre for Education Research and Policy.

Adnett, N. and Davies, P. (2003) *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*. Online: AERA. Available at: http://edr.sagepub.com/content/early/2015/11/10/0013189X15618385.full.pdf+html (Accessed: 15th January 2016).

Aitkin, M. and Longford, N. (1986) 'Statistical modelling issues in school effectiveness studies', *Journal of the Royal Statistical Society. Series A (General)*, pp. 1-43.

Allen, M. J. and Yen, W. M. (2001) *Introduction to measurement theory*. Waveland Press.

Allen, R. (2012) 'Measuring foundation school effectiveness using English administrative data, survey data and a regression discontinuity design', *Education Economics,* 21(5), pp. 431-446.

Allen, R. (2015a) *National Pupil Database Wiki*. Online: wikispaces.com. Available at: https://nationalpupildatabase.wikispaces.com/ (Accessed: 7th September 2015).

Allen, R. (2015b) *We cannot compare the effectiveness of schools with different types of intakes*. Blog: Education Datalab. Available at: http://www.educationdatalab.org.uk/Blog/May-2015/We-cannot-compare-the-effectiveness-of-schools-wit.aspx#.VaAOkq5Viko (Accessed: 10th July 2015).

Allen, R. and Burgess, S. (2011) 'Can School League Tables Help Parents Choose Schools?', *Fiscal Studies,* 32(2), pp. 245-261.

Allen, R. and Burgess, S. (2013) 'Evaluating the provision of school performance information for school choice', *Economics of Education Review,* 34, pp. 175-190.

Altrichter, H. and Kemethofer, D. (2015) 'Does accountability pressure through school inspections promote school improvement?', *School Effectiveness and School Improvement,* 26(1), pp. 32-56.

Amrein-Beardsley, A. (2014) *Rethinking value-added models in education: Critical perspectives on tests and assessment-based accountability*. Routledge.

Antoniou, P. and Kyriakides, L. (2011) 'The impact of a dynamic approach to professional development on teacher instruction and student learning: results from an experimental study', *School Effectiveness and School Improvement,* 22(3), pp. 291-311.

ASA (2014) *ASA statement on using value-added models for educational assessment*. Available at: https://www.amstat.org/policy/pdfs/ASA_VAM_Statement.pdf (Accessed: 10th January 2016).

Ball, S. J. (2008) *The education debate.* The Policy Press.

Barber, M. (2001) 'The Very Big Picture', *School Effectiveness and School Improvement,* 12(2), pp. 213-228.

Barber, M. (2004) 'The virtue of accountability: System redesign, inspection, and incentives in the era of informed professionalism', *Journal of education,* 185(1), pp. 7-38.

Benton, T. 2014. The relationship between time in education and achievement in PISA in England. *Cambridge Assessment.* Working paper: University of Cambridge.

Berk, R. A. and Freedman, D. A. (2003) *Statistical assumptions as empirical commitments.* Law, Punishment, and Social Control: Essays in Honor of Sheldon Messinger, 2nd edn. Aldine de Gruyter.

Bernstein, B. (1970) 'Education cannot compensate for society', *New society,* 15(387), pp. 344-351.

Bird, S. M., David, C., Farewell, V. T., Harvey, G., Tim, H. and Peter, C. (2005) 'Performance indicators: good, bad, and ugly', *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 168(1), pp. 1-27.

Bloom, H. S. (2009) 'Modern Regression Discontinuity Analysis', *Journal of Research on Educational Effectiveness,* 5(1), pp. 43-82.

Bogotch, I., Mirón, L. and Biesta, G. (2007) '"Effective for what; effective for whom?" Two questions SESI should not ignore', *International handbook of school effectiveness and improvement*: Springer, pp. 93-110.

Boonen, T., Speybroeck, S., Bilde, J., Lamote, C., Damme, J. and Onghena, P. (2013a) '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), pp. 441-466.

Boonen, T., Van Damme, J. and Onghena, P. (2013b) 'Teacher effects on student achievement in first grade: which aspects matter most?', *School Effectiveness and School Improvement,* 25(1), pp. 126-152.

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), pp. 741-751.

Bosker, R. J. and Scheerens, J. (1994) 'Alternative models of school effectiveness put to the test', *International Journal of Educational Research,* 21(2), pp. 159-180.

Bradbury, A. (2011) 'Equity, Ethnicity and the Hidden Dangers of "Contextual" Measures of School Performance', *Race, Ethnicity and Education,* 14(3), pp. 277-291.

Brennan, R. L. (2013) 'Commentary on "Validating the Interpretations and Uses of Test Scores"', *Journal of Educational Measurement,* 50(1), pp. 74-83.

Brewer, D. J. and McEwan, P. J. (2010) *Economics of education.* Elsevier.

Bridges, D. and McLaughlin, T. H. (1994) *Education and the market place.* Psychology Press.

Burgess, S. 2014. Understanding the success of London's schools. Working Paper 14: Centre for Market and Public Organisation (CMPO).

Burgess, S. and Thomson, D. (2013a) *Key Stage 4 Accountability: Progress Measure and Intervention Trigger*, http://www.bristol.ac.uk/cubec/portal/: BUBeC, University of Bristol.

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*, http://www.bristol.ac.uk/cubec/portal/: BUBeC, University of Bristol.

Cahan, S. and Cohen, N. (1989) 'Age versus schooling effects on intelligence development', *Child development,* 60(5), pp. 1239-1249.

Cahan, S. and Davis, D. (1987) 'A Between-Grade-Levels Approach to the Investigation of the Absolute Effects of Schooling on Achievement', *American Educational Research Journal,* 24(1), pp. 1-12.

Cahan, S. and Elbaz, J. G. (2000) 'The Measurement of School Effectiveness', *Studies in Educational Evaluation,* 26(2), pp. 127-42.

Camilli, G. (1996) 'Standard errors in educational assessment', *Education Policy Analysis Archives,* 4, pp. 4.

Carlson, D., Borman, G. D. and Robinson, M. (2011) 'A multistate district-level cluster randomized trial of the impact of data-driven reform on reading and mathematics achievement', *Educational Evaluation and Policy Analysis,* 33(3), pp. 378-398.

CEM (2015) *Publication Of Schools' Performance Data.* Online: CEM. Available at: http://www.cem.org/publication-of-schools-performance-data (Accessed: 4th September 2015).

Chang, H. (2004) *Inventing temperature: Measurement and scientific progress.* Oxford University Press.

Chapman, C. and Muijs, D. (2013) 'Does school-to-school collaboration promote school improvement? A study of the impact of school federations on student outcomes', *School Effectiveness and School Improvement,* 25(3), pp. 351-393.

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.

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.

Chetty, R., Friedman, J. and Rockoff, J. (2014a) 'Discussion of the American Statistical Association's Statement (2014) on Using Value-Added Models for Educational Assessment', *Statistics and Public Policy,* 1(1), pp. 111-113.

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. (2014b) 'Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates', *American Economic Review,* 104(9), pp. 2593-2632.

Chetty, R., Friedman, J. N. and Rockoff, J. E. (2014c) 'Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood', *American Economic Review,* 104(9), pp. 2633-79.

Chitty, C. (2009) *Education policy in Britain.* 2nd edn. London: Palgrave Macmillan.

Cliffordson, C. (2010) 'Methodological Issues in Investigations of the Relative Effects of Schooling and Age on School Performance: The Between-Grade Regression Discontinuity Design Applied to Swedish TIMSS 1995 Data', *Educational Research and Evaluation,* 16(1).

Coe, R. (1998) *The Significance of Significance*. Available at: http://community.dur.ac.uk/r.j.coe/teaching/critsig.htm (Accessed: 12th January 2016).

Coe, R. (2009) 'Unobserved but not unimportant: the effects of unmeasured variables on causal attributions', *Effective Education,* 1(2), pp. 101-122.

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

Coe, R., Aloisi, C., Higgins, S. and Major, L. E. (2014) *What makes great teaching? Review of the underpinning research*, London: Sutton Trust.

Coe, R. and Fitz-Gibbon, C. T. (1998) 'School effectiveness research: Criticisms and recommendations', *Oxford Review of Education,* 24(4), pp. 421-438.

Coe, R., Karen Jones, Jeff Searle, Dimitra Kokotsaki, Azlina Mohd Kosnin and Skinner, P. (2008) 'Evidence on the effects of selective educational systems', *A report for the Sutton Trust. Durham: CEM Centre, University of Durham, for the Sutton Trust.*

Coleman, J. S. (1968) 'Equality of educational opportunity', *Integrated Education,* 6(5), pp. 19-28.

Condie, S., Lefgren, L. and Sims, D. (2014) 'Teacher heterogeneity, value-added and education policy', *Economics of Education Review,* 40, pp. 76-92.

Corcoran, S. and Goldhaber, D. (2013) 'Value Added and its Uses: Where You Stand Depends on Where You Sit', *Education,* 8(3), pp. 418-434.

Crawford, C., Dearden, L. and Greaves, E. (2013) 'The drivers of month of birth differences in children's cognitive and non-cognitive skills: a regression discontinuity analysis'', *Institute for Fiscal Studies (IFS), Working Paper W13/08.*

Crawford, C., Dearden, L. and Meghir, C. (2007) *When you are born matters: The impact of date of birth on child cognitive outcomes in England.* Centre for the Economics of Education, London School of Economics and Political Science.

Crawford, C., Dearden, L. and Meghir, C. (2010) 'When you are born matters: the impact of date of birth on educational outcomes in England', *Institute for Fiscal Studies (IFS), Working Paper W10/06.*

Creemers, B. P., Kyriakides, L. and Sammons, P. (2010) *Methodological advances in educational effectiveness research.* Routledge.

Davies, P., Coates, G., Hammersley-Fletcher, L. and Mangan, J. (2005) 'When 'becoming a 50% school'is success enough: a principal–agent analysis of subject leaders' target setting', *School Leadership and Management,* 25(5), pp. 493-511.

de Bilde, J., Van Damme, J., Lamote, C. and De Fraine, B. (2013) 'Can alternative education increase children's early school engagement? A longitudinal study from kindergarten to third grade', *School Effectiveness and School Improvement,* 24(2), pp. 212-233.

Dearden, L., Micklewright, J. and Vignoles, A. (2011a) 'The Effectiveness of English Secondary Schools for Pupils of Different Ability Levels*', *Fiscal Studies,* 32(2), pp. 225-244.

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.

Demie, F. (2013) *Using Data to Raise Achievement: Good Practice in Schools*, London: Lambeth Council.

DfE (2010) *The Importance of Teaching - White Paper.* London: Her Majesty's Stationery Office.

DfE (2011) *How do pupils progress during Key Stages 2 and 3? Research Report DFE-RR096*, London: DfE. Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/182413/DFE-RR096.pdf.

DfE (2012) *New data reveals the truth about school performance*: Department for Education. Available at: http://www.education.gov.uk/inthenews/inthenews/a00202531/secperftables12 (Accessed: 14th December 2012).

DfE (2013a) *A Guide to Value Added Key Stage 2 to 4 in 2013 School Performance Tables & RAISEonline*: DfE. Available at: http://www.education.gov.uk/schools/performance/2013/secondary_13/KS2-4_Performance_Tables_General_VA_Guide_2013_FINAL.pdf.

DfE (2013b) *New advice to help schools set performance-related pay*. Online: DfE. Available at: https://www.gov.uk/government/news/new-advice-to-help-schools-set-performance-related-pay (Accessed: 5th September 2015).

DfE (2013c) *Reforming the accountability system for secondary schools*. Government response to the February to May 2013 consultation on secondary school accountability: Department for Education. Available at: https://www.gov.uk/government/consultations/secondary-school-accountability-consultation.

DfE (2014a) *Progress 8 measure in 2016*: Department for Education. Available at: https://www.gov.uk/government/organisations/department-for-education.

DfE (2014b) *Progress 8 school performance measure*: Department for Education. Available at: https://www.gov.uk/government/organisations/department-for-education.

DfE (2015) *School Performance Tables*. Available at: http://www.education.gov.uk/schools/performance/ (Accessed: 26th March 2015 2015).

DfE (2016) *Primary school accountability*: Department for Education. Available at: https://www.gov.uk/government/publications/primary-school-accountability.

Downey, C. and Kelly, A. (2013) 'Professional attitudes to the use of data in England', *Data-based decision making in education*: Springer, pp. 69-89.

Dumay, X., Coe, R. and Anumendem, D. N. (2013) 'Stability over time of different methods of estimating school performance', *School Effectiveness and School Improvement,* 25(1), pp. 64-82.

Easen, P. and Bolden, D. (2005) 'Location, Location, Location: What Do League Tables Really Tell Us about Primary Schools?', *Education 3-13,* 33(3), pp. 49-55.

Evans, H. (2008) *Value-Added in English Schools*. London: Dept for Children, Schools, Families. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.178.9363&rep=rep1&type=pdf (Accessed: 3rd February 2014).

Ferrão, M. E. (2012) 'On the stability of value added indicators', *Quality & Quantity,* 46(2), pp. 627-637.

Ferrão, M. E. and Couto, A. P. (2013) 'The use of a school value-added model for educational improvement: a case study from the Portuguese primary education system', *School Effectiveness and School Improvement,* 25(1), pp. 174-190.

Ferrão, M. E. and Goldstein, H. (2009) 'Adjusting for measurement error in the value added model: evidence from Portugal', *Quality & Quantity,* 43(6), pp. 951-963.

Fitz-Gibbon, C. (1996) *Monitoring School Effectiveness: Simplicity and Complexivity. Merging Traditions: The Future of Research on School Effectiveness and School Improvement.* London.

Fitz-Gibbon, C. T. (1997) *The Value Added National Project: Final Report: Feasibility Studies for a National System of Value-added Indicators.* SCAA.

Foley, B. and Goldstein, H. 2013. Measuring success: League tables in the public sector. British Academy.

Freddie, W. (2015) *Automatic registration for free school meals gets cross-party backing | Schools Week*: Schools Week. Available at: http://schoolsweek.co.uk/automatic-registration-for-free-school-meals-gets-cross-party-backing/ (Accessed: January 26th 2016).

Galbraith, J. K. (1998) *The Affluent Society*. Houghton Mifflin Harcourt.

Gates Foundation (2013) *Annual Letter 2013.* Online. Available at: http://www.gatesfoundation.org/Who-We-Are/Resources-and-Media/Annual-Letters-List/Annual-Letter-2013 (Accessed: 15th April 2015).

Gershenson, S. and Langbein, L. (2015) 'The Effect of Primary School Size on Academic Achievement', *Educational Evaluation and Policy Analysis,* 37(1 suppl), pp. 135S-155S.

Gibbs, B. G., Shafer, K. and Miles, A. (2015) 'Inferential statistics and the use of administrative data in US educational research', *International Journal of Research & Method in Education*, pp. 1-7.

Gillard, D. (2011) *Education in England: a brief history*. Available at: www.educationengland.org.uk/history (Accessed: 17th March 2015).

Glass, G. V. (2014) 'A response to Gorard', *The Psychology of Education Review,* 38(1), pp. 12-13.

Goldhaber, D. (2015) 'Exploring the Potential of Value-Added Performance Measures to Affect the Quality of the Teacher Workforce', *Educational Researcher,* 44(2), pp. 87-95.

Goldhaber, D. D., Goldschmidt, P. and Tseng, F. (2013) 'Teacher Value-Added at the High-School Level: Different Models, Different Answers?', *Educational Evaluation and Policy Analysis,* 35(2), pp. 220-236.

Goldstein, H. (1997) 'Methods in School Effectiveness Research', *School Effectiveness and School Improvement,* 8(4), pp. 369-395.

Goldstein, H. (2008) 'Evidence and education policy–some reflections and allegations', *Cambridge Journal of Education,* 38(3), pp. 393-400.

Goldstein, H., Burgess, S. and McConnell, B. (2007) 'Modelling the effect of pupil mobility on school differences in educational achievement', *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 170(4), pp. 941-954.

Goldstein, H., Huiqi, P., Rath, T. and Hill, N. (2000) *The use of value-added information in judging school performance.* Institute of Education, University of London London.

Goldstein, H., Kounali, D. and Robinson, A. (2008) 'Modelling measurement errors and category misclassifications in multilevel models', *Statistical Modelling,* 8(3), pp. 243-261.

Goldstein, H. and Leckie, G. (2008) 'School league tables: what can they really tell us?', *Significance,* 5(2), pp. 67-69.

Goldstein, H. and Noden, P. (2004) 'A response to Gorard on social segregation', *Oxford Review of Education,* 30(3), pp. 441-442.

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)*, pp. 385-443.

Goldstein, H. and Woodhouse, G. (2000) 'School effectiveness research and educational policy', *Oxford Review of Education,* 26(3-4), pp. 353-363.

Gorard, S. (2006a) 'Is There a School Mix Effect?', *Educational Review,* 58(1), pp. 87-94.

Gorard, S. (2006b) 'Towards a Judgement-Based Statistical Analysis', *British Journal of Sociology of Education,* 27(1), pp. 67-80.

Gorard, S. (2006c) 'Value-added is of little value', *J. Educ. Policy,* 21(2), pp. 235-243.

Gorard, S. (2007) 'The dubious benefits of multi-level modeling', *International Journal of Research & Method in Education,* 30(2), pp. 221-236.

Gorard, S. (2010) 'Serious doubts about school effectiveness', *British Educational Research Journal,* 36(5), pp. 745-766.

Gorard, S. (2011a) *Comments on 'The value of educational effectiveness research'.* BERA Conference: BERA. Available at: http://rab.bham.ac.uk/pubs.asp?id=049c1113-9b7a-4978-b370-98cdc55d3e3a.

Gorard, S. (2011b) 'Doubts about school effectiveness exacerbated–By attempted justification', *Research Intelligence,* 114(26), pp. Q8.

Gorard, S. (2011c) 'Now You See It, Now You Don't: School Effectiveness as Conjuring?', *Research in Education,* 86(1), pp. 39-45.

Gorard, S. (2012a) 'The Increasing Availability of Official Datasets: methods, limitations and opportunities for studies of education', *Br. J. Educ. Stud.,* 60(1), pp. 77-92.

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), pp. 1003-1017.

Gorard, S. (2013) *Research design: Robust approaches for the social sciences.* Sage, London, UK.

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), pp. 3-10.

Gorard, S. (2015) 'Rethinking 'quantitative' methods and the development of new researchers', *Review of Education,* 3(1), pp. 72-96.

Gorard, S., Hordosy, R. and Siddiqui, N. (2012) 'How Unstable are 'School Effects' Assessed by a Value-added Technique?', *International Education Studies,* 6(1), pp. p1.

Gray, J. (2004) 'School effectiveness and the 'other outcomes' of secondary schooling: A reassessment of three decades of British research', *Improving Schools,* 7(2), pp. 185-198.

Gray, J., Goldstein, H. and Thomas, S. (2001) 'Predicting the future: the role of past performance in determining trends in institutional effectiveness at A level', *British Educational Research Journal,* 27(4), pp. 391-405.

Gray, J., Jesson, D. and Jones, B. (1986) 'The search for a fairer way of comparing schools' examination results', *Research papers in education,* 1(2), pp. 91-122.

Grisay, A. (1997) *Evolution des acquis cognitifs et socio-affectifs des élèves au cours des années de collège.* Ministère de l'éducation nationale, de la recherche et de la technologie, Direction de l'évaluation et de la prospective.

Guardian (2014) *Education chief fights back in battle with Michael Gove over schools*: theguardian.com. Available at: http://www.theguardian.com/politics/2014/dec/06/nicky-morgan-battle-michael-gove-schools (Accessed: 19th August 2015).

Guarino, C., Dieterle, S. G., Bargagliotti, A. E. and Mason, W. M. (2013) 'What Can We Learn About Effective Early Mathematics Teaching? A Framework for Estimating Causal Effects Using Longitudinal Survey Data', *Journal of Research on Educational Effectiveness,* 6(2), pp. 164-198.

Guldemond, H. and Bosker, R. (2009) 'School effects on students' progress - a dynamic perspective', *School Effectiveness and School Improvement,* 20(2), pp. 255-268.

Gunter, H. M. and Fitzgerald, T. (2015) 'Educational administration and neoliberalism: historical and contemporary perspectives', *Journal of Educational Administration and History,* 47(2), pp. 101-104.

Gustafsson, J.-E. (2013) 'Causal inference in educational effectiveness research: a comparison of three methods to investigate effects of homework on student achievement', *School Effectiveness and School Improvement,* 24(3), pp. 275-295.

Gutman, L. M. and Vorhaus, J. (2012) *The impact of pupil behaviour and wellbeing on educational outcomes.* Online. Available at: http://dera.ioe.ac.uk/16093/1/DFE-RR253.pdf.

Harker, R. and Tymms, P. (2004) 'The effects of student composition on school outcomes', *School effectiveness and school improvement,* 15(2), pp. 177-199.

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), pp. 245-270.

Harris, A. and Goodall, J. (2008) 'Do parents know they matter? Engaging all parents in learning', *Educational Research,* 50(3), pp. 277-289.

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), pp. 319-350.

Harris, D. N., Ingle, W. K. and Rutledge, S. A. (2014) 'How Teacher Evaluation Methods Matter for Accountability: A Comparative Analysis of Teacher Effectiveness Ratings by Principals and Teacher Value-Added Measures', *American Educational Research Journal,* 51(1), pp. 73-112.

Haworth, C. M. A., Asbury, K., Dale, P. S. and Plomin, R. (2011) 'Added Value Measures in Education Show Genetic as Well as Environmental Influence', *PLoS ONE,* 6(2), pp. e16006.

He, Q., Hayes, M. and Wiliam, D. (2013) 'Classification accuracy in Key Stage 2 National Curriculum tests in England', *Research Papers in Education,* 28(1), pp. 22-42.

He, Q. and Tymms, P. (2014) 'The principal axis approach to value-added calculation', *Educational Research and Evaluation,* 20(1), pp. 25-43.

Heck, R. H. and Mahoe, R. (2010) 'Student Course Taking and Teacher Quality: Their Effects on Achievement and Growth', *International Journal of Educational Management,* 24(1), pp. 56-72.

Hill, A. B. (1965) 'The environment and disease: association or causation?', *Proceedings of the Royal Society of Medicine,* 58(5), pp. 295.

Hill, P. and Rowe, K. (1996) 'Multilevel Modelling in School Effectiveness Research', *School Effectiveness and School Improvement,* 7(1), pp. 1-34.

Howe, C. (2014) 'A response to Gorard', *The Psychology of Education Review,* 38(1).

Isaacs, T., Zara, C., Smith, C., Herbert, G. and Coombs, S. J. (2013) *Key Concepts in Educational Assessment.* Sage.

Isac, M. M., Maslowski, R., Creemers, B. and van der Werf, G. (2013) 'The contribution of schooling to secondary-school students' citizenship outcomes across countries', *School Effectiveness and School Improvement,* 25(1), pp. 29-63.

Isenberg, E., Teh, B.-r. and Walsh, E. (2015) 'Elementary School Data Issues for Value-Added Models: Implications for Research', *Journal of Research on Educational Effectiveness,* 8(1), pp. 120-129.

Jencks, C. (1972) *Inequality: A reassessment of the effect of family and schooling in America.* New York: Basic Books, Inc.

Johnson, M. T., Lipscomb, S. and Gill, B. (2014) 'Sensitivity of Teacher Value-Added Estimates to Student and Peer Control Variables', *Journal of Research on Educational Effectiveness,* 8(1), pp. 60-83.

Johnson, S. (2013) 'On the reliability of high-stakes teacher assessment', *Research Papers in Education,* 28(1), pp. 91-105.

Kahneman, D. (2011) *Thinking, Fast and Slow.* New York: Farrar, Straus and Giroux, p. 499.

Kane, M. T. (2013) 'Validating the Interpretations and Uses of Test Scores', *Journal of Educational Measurement,* 50(1), pp. 1-73.

Kane, T. J. and Staiger, D. O. (2008) *Estimating teacher impacts on student achievement: An experimental evaluation*: National Bureau of Economic Research.

Kay, B. (1976) 'The Assessment of Performance Unit: its task and rationale', *Education 3-13,* 4(2), pp. 108-112.

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), pp. 181-198.

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

Kelly, A. and Downey, C. (2011b) *Using effectiveness data for school improvement: Developing and utilising metrics.* Abbingdon Oxon: Routledge.

Kelly, A., Downey, C. and Rietdijk, W. (2010) *Data dictatorship and data democracy: understanding professional attitudes to the use of pupil performance data in schools*, Online: CfBT. Available at: http://eprints.soton.ac.uk/147597/4/SUMMARY_REPORT_DataDictatorship_web.pdf.

Kennedy, M. M. (2010) 'Attribution error and the quest for teacher quality', *Educational Researcher,* 39(8), pp. 591-598.

Kirkup, C., Sizmur, J., Sturman, L. and Lewis, K. (2005) *Schools' Use of Data in Teaching and Learning. Research Report RR671.* ERIC.

Knuver, A. and Brandsma, H. (1993) 'Cognitive and Affective Outcomes in School Effectiveness Research', *School Effectiveness and School Improvement,* 4(3), pp. 189-204.

Konstantopoulos, S. and Sun, M. (2013) 'Are teacher effects larger in small classes?', *School Effectiveness and School Improvement*, pp. 1-17.

Koretz, D. M. (2008) *Measuring up.* Harvard University Press.

Kvanvig, J. 2008. Coherentist theories of epistemic justification. *Stanford Encyclopedia of Philosophy.*

Kyriakides, L. and Luyten, H. (2009) 'The contribution of schooling to the cognitive development of secondary education students in Cyprus: An application of regression discontinuity with multiple cut-off points', *School Effectiveness and School Improvement,* 20(2), pp. 167-186.

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), pp. 1-17.

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), pp. 367-414.

Le Grand, J. (1991) 'Quasi-markets and social policy', *The Economic Journal,* 101(408), pp. 1256-1267.

Leckie, G. (2013) 'England's Multilevel Model Based Value-Added School League Tables: Measuring and communicating statistical uncertainty to parents', 68, pp. 1 - 6.

Leckie, G. and Goldstein, H. (2009) 'The limitations of using school league tables to inform school choice', *Journal of the Royal Statistical Society,* 172, pp. 835-851.

Leckie, G. and Goldstein, H. (2011) 'Understanding Uncertainty in School League Tables*', *Fiscal Studies,* 32(2), pp. 207-224.

Lee, J. and Fitz, J. (1997) 'HMI and OFSTED: evolution or revolution in school inspection', *British Journal of Educational Studies,* 45(1), pp. 39-52.

Lenkeit, J. (2013) 'Effectiveness measures for cross-sectional studies: a comparison of value-added models and contextualised attainment models', *School Effectiveness and School Improvement,* 24(1), pp. 39-63.

Little, T. D. (2013) *The Oxford handbook of quantitative methods in psychology*. Oxford University Press.

Liu, H., Van Damme, J., Gielen, S. and Van Den Noortgate, W. (2015) 'School processes mediate school compositional effects: model specification and estimation', *British Educational Research Journal,* 41(3), pp. 423-447.

Lumby, J. and Muijs, D. (2014) 'Corrupt language, corrupt thought: the White Paper The importance of teaching', *British Educational Research Journal,* 40(3), pp. 523-538.

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), pp. 197-216.

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), pp. 31-51.

Luyten, H. (2006) 'An empirical assessment of the absolute effect of schooling: regression-discontinuity applied to TIMSS-95', *Oxford Review of Education,* 32(3), pp. 397-429.

Luyten, H. and de Wolf, I. (2011) 'Changes in student populations and average test scores of Dutch primary schools', *School Effectiveness and School Improvement,* 22(4), pp. 439-460.

Luyten, H., Peschar, J. and Coe, R. (2008) 'Effects of Schooling on Reading Performance, Reading Engagement, and Reading Activities of 15-Year-Olds in England', *American Educational Research Journal,* 45(2), pp. 319-342.

Luyten, H. and Sammons, P. (2010) 'Multilevel Modelling', in Creemers, B., Kyriakides, L. & Sammons, P. (eds.) *Methodological Advances in Educational Effectiveness Research*. London: Routledge, pp. 246-276.

Luyten, H., Tymms, P. and Jones, P. (2009) 'Assessing school effects without controlling for prior achievement?', *School Effectiveness and School Improvement,* 20(2), pp. 145-165.

Luyten, H. and Veldkamp, B. (2011) 'Assessing Effects of Schooling With Cross-Sectional Data: Between-Grades Differences Addressed as a Selection-Bias Problem', *Journal of Research on Educational Effectiveness,* 4(3), pp. 264-288.

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), pp. 249-279.

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), pp. 203-216.

Manzi, J., San Martín, E. and Van Bellegem, S. (2014) 'School system evaluation by value added analysis under endogeneity', *Psychometrika,* 79(1), pp. 130-153.

Marks, G. N. (2014) 'The size, stability, and consistency of school effects: evidence from Victoria', *School Effectiveness and School Improvement,* (ahead-of-print), pp. 1-18.

Marks, G. N. (2015) 'Are school-SES effects statistical artefacts? Evidence from longitudinal population data', *Oxford Review of Education,* 41(1), pp. 122-144.

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), pp. 279-295.

McCaffrey, D. F., Lockwood, J., Koretz, D., Louis, T. A. and Hamilton, L. (2004) 'Models for value-added modeling of teacher effects', *Journal of educational and behavioral statistics,* 29(1), pp. 67-101.

McNaughton, S., Lai, M. K. and Hsiao, S. (2012) 'Testing the effectiveness of an intervention model based on data use: a replication series across clusters of schools', *School Effectiveness and School Improvement,* 23(2), pp. 203-228.

Melhuish, E., Quinn, L., Sylva, K., Sammons, P., Siraj-Blatchford, I. and Taggart, B. (2013) 'Preschool affects longer term literacy and numeracy: results from a general population longitudinal study in Northern Ireland', *School Effectiveness and School Improvement,* 24(2), pp. 234-250.

Messick, S. (1987) 'VALIDITY', *ETS Research Report Series,* 1987(2), pp. i-208.

Meyer, R. H. (1997) 'Value-added indicators of school performance: A primer', *Economics of Education Review,* 16(3), pp. 283-301.

Morgan, N. (2015) *Nicky Morgan: why knowledge matters*: DfE. Available at: https://www.gov.uk/government/speeches/nicky-morgan-why-knowledge-matters (Accessed: 15th April 2015).

Morganstein, D. and Wasserstein, R. (2014) 'ASA Statement on Value-Added Models', *Statistics and Public Policy,* 1(1), pp. 108-110.

Morris, R. (2015) 'Free Schools and disadvantaged intakes', *British Educational Research Journal,* 41(4), pp. 535-552.

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,* 1(3), pp. 315-332.

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, pp. 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), pp. 231-256.

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), pp. 273-303.

Muñoz-Chereau, B. and Thomas, S. M. (2015) 'Educational effectiveness in Chilean secondary education: comparing different ‘value added’ approaches to evaluate schools', *Assessment in Education: Principles, Policy & Practice*, pp. 1-27.

Murray, J. (2013) 'Critical issues facing school leaders concerning data-informed decision-making', *School Leadership & Management*, pp. 1-9.

Neale, D. (2015) 'Defending the logic of significance testing: a response to Gorard', *Oxford Review of Education*, pp. 1-12.

Newton, P. and Shaw, S. (2014) *Validity in educational and psychological assessment.* Sage.

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), pp. 1-21.

Nor, M. Y. M. (2014) 'Potentials of Contextual Value-Added Measures in Assisting Schools Become More Effective', *International Education Studies,* 7(13), pp. p75.

Noyes, A. (2013) 'The effective mathematics department: adding value and increasing participation?', *School Effectiveness and School Improvement,* 24(1), pp. 1-17.

Nye, B., Konstantopoulos, S. and Hedges, L. V. (2004) 'How Large Are Teacher Effects?', *Educational Evaluation and Policy Analysis,* 26(3), pp. 237-257.

O'Neill, O. (2013) 'Intelligent accountability in education', *Oxford Review of Education,* 39(1), pp. 1-13.

OECD (2008) *Measuring improvements in learning outcomes: Best practices to assess the value-added of schools.* OECD Publishing.

Ofsted (2013) *Data Dashboard.* Online. Available at: http://dashboard.ofsted.gov.uk/ (Accessed: 25th June 2013).

Ofsted (2015) *School inspection handbook.* Online: Ofsted. Available at: https://www.gov.uk/government/publications/school-inspection-handbook-from-september-2015.

Plewis, I. and Fielding, A. (2003) 'What is multi-level modelling for? A critical response to Gorard (2003)', *British Journal of Educational Studies,* 51(4), pp. 408-419.

Pokropek, A. (2014) 'Phantom Effects in Multilevel Compositional Analysis: Problems and Solutions', *Sociological Methods & Research*, pp. 1-29.

Popper, K. (2005) *The logic of scientific discovery.* Routledge.

Pring, R. (2012) 'Putting persons back into education', *Oxford Review of Education,* 38(6), pp. 747-760.

Putwain, D. (2014) 'A response to Gorard', *The Psychology of Education Review,* 38(1), pp. 17-19.

Rasbash, J., Leckie, G., Pillinger, R. and 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), pp. 657-682.

Raudenbush, S. W. (2015) 'Value Added: A Case Study in the Mismatch Between Education Research and Policy', *Educational Researcher,* 44(2), pp. 138-141.

Raudenbush, S. W. and Willms, J. (1995) 'The Estimation of School Effects', *Journal of Educational and Behavioral Statistics,* 20(4), pp. 307-335.

Ray, A. (2006) 'School Value Added Measures in England', *DfES, A paper for the OECD Project on the Development of Value-Added Models in Education Systems*.

Ready, D. D. (2013) 'Associations Between Student Achievement and Student Learning Implications for Value-Added School Accountability Models', *Educational Policy,* 27(1), pp. 92-120.

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), pp. 109-127.

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), pp. 197-230.

Reynolds, D., Sammons, P., Stoll, L., Barber, M. and Hillman, J. (1996) 'School effectiveness and school improvement in the United Kingdom', *School Effectiveness and School Improvement,* 7(2), pp. 133-158.

Rutter, M. (1983) 'School effects on pupil progress: Research findings and policy implications', *Child development*, pp. 1-29.

Rutter, M. and Maughan, B. (2002) 'School effectiveness findings 1979–2002', *Journal of school psychology,* 40(6), pp. 451-475.

Rutter, M., Maughan, B., Mortimore, P. and Ouston, J. (1979) *Fifteen thousand hours: Secondary schools and their effects on children.* Somerset, England: Open Books Publishing Ltd.

Sammons, P. (1995) *Key characteristics of effective schools.* University of London.

Sammons, P. (1996) 'Complexities in the Judgement of School Effectiveness', *Educational Research and Evaluation,* 2(2), pp. 113-149.

Sammons, P. (2014) *Influences on students' GCSE attainment and progress at age 16: Effective Pre-School, Primary & Secondary Education Project (EPPSE): September 2014*, London: Institute of Education, University of London.

Sammons, P., Hall, J., Sylva, K., Melhuish, E., Siraj-Blatchford, I. and Taggart, B. (2012) 'Protecting the development of 5–11-year-olds from the impacts of early disadvantage: the role of primary school academic effectiveness', *School Effectiveness and School Improvement*, pp. 1-18.

Sammons, P. and Luyten, H. (2009) 'Editorial article for special issue on alternative methods for assessing school effects and schooling effects', *School Effectiveness and School Improvement,* 20(2), pp. 133-143.

Sammons, P., Mortimore, P. and Thomas, S. (1996) 'Do schools perform consistently across outcomes and areas', *Merging traditions: The future of research on school effectiveness and school improvement*, pp. 3-29.

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), pp. 381-405.

Sammons, P., Sylva, K., Melhuish, E., Siraj-Blatchford, I., Taggart, B., Grabbe, Y. and Barreau, S. (2007) *Effective Pre-School and Primary Education 3-11 Project (EPPE 3-11): Summary Report, Influences on Children's Attainment and Progress in Key Stage 2: Cognitive Outcomes in Year 5.* Institute of Education, University of London/Department for Education and Skills.

Sammons, P., Thomas, S., Mortimore, P., Owen, C. and Pennell, H. (1994) *Assessing school effectiveness: Developing measures to put school performance in context.* Institute of Education, International School Effectiveness & Improvement Centre.

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), pp. 233-256.

Saunders, L. (2000) 'Understanding schools' use of 'value added' data: the psychology and sociology of numbers', *Research Papers in Education,* 15(3), pp. 241-258.

Saunders, L. and Rudd, P. (1999) 'Schools' use of "value added" data: a science in the service of an art?'', *British Educational Research Association Conference*, Brighton, University of Sussex

SCAA (1994) *Value Added Performance Indicators for Schools.* London: School Curriculum and Assessment Authority.

Scheerens, J. (1993a) 'Basic School Effectiveness Research: Items for a Research Agenda', *School Effectiveness and School Improvement,* 4(1), pp. 17-36.

Scheerens, J. (1993b) 'Effective schooling: Research, theory and practice', *School Effectiveness and School Improvement,* 4(3), pp. 230-235.

Schildkamp, K., Poortman, C. L. and Handelzalts, A. (2015) 'Data teams for school improvement', *School Effectiveness and School Improvement*, pp. 1-27.

Scimago Lab (2016) *The SCImago Journal & Country Rank.* Available at: http://www.scimagojr.com/index.php (Accessed: January 11th 2016).

Shadish, W. R., Cook, T. D. and Campbell, D. T. (2002) *Experimental and quasi-experimental designs for generalized causal inference.* Wadsworth Cengage learning.

Sireci, S. G. (2015) 'On the validity of useless tests', *Assessment in Education: Principles, Policy & Practice*, pp. 1-10.

Slee, R. and Weiner, G. (2001) 'Education Reform and Reconstruction as a Challenge to Research Genres: Reconsidering School Effectiveness Research and Inclusive Schooling', *School Effectiveness and School Improvement,* 12(1), pp. 83-98.

Smith, D., Tomlinson, S., Bonnerjea, L., Hogarth, T. and Thomes, H. (1989) *The School Effect: A Study of Multi Racial Comprehensives.* Policy Studies Institute.

Smith, G. (2000) 'Research and Inspection: HMI and OFSTED, 1981-1996-a commentary', *Oxford Review of Education,* 26(3-4), pp. 333-352.

Snijders, T. and Bosker, R. (2011) *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* 2nd edn. London: Sage.

Spinath, B., Eckert, C. and Steinmayr, R. (2014) 'Gender differences in school success: What are the roles of students' intelligence, personality and motivation?', *Educational Research*, pp. 1-14.

Stankov, L. and Lee, J. (2014) 'Quest for the best non-cognitive predictor of academic achievement', *Educational Psychology,* 34(1), pp. 1-8.

Stobart, G. (2008) *Testing times*. Routledge.

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), pp. 209-225.

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), pp. 289-314.

Strand, S. (2014a) 'Ethnicity, gender, social class and achievement gaps at age 16: intersectionality and 'getting it' for the white working class', *Research Papers in Education,* 29(2), pp. 131-171.

Strand, S. (2014b) 'School effects and ethnic, gender and socio-economic gaps in educational achievement at age 11', *Oxford Review of Education*, pp. 1-23.

Stringfield, S. and Herman, R. (1996) 'Assessment of the State of School Effectiveness Research in the United States of America', *School Effectiveness and School Improvement,* 7(2), pp. 159-180.

Styles, B. (2014) 'A response to Gorard', *The Psychology of Education Review,* 38(1), pp. 20-21.

Syverson, P. (2008) 'An ecological view of literacy learning', *Literacy,* 42(2), pp. 109-117.

Taylor, J. and Nguyen, A. N. (2006) 'An Analysis of the Value Added by Secondary Schools in England: Is the Value Added Indicator of Any Value?*', *Oxford Bulletin of Economics and Statistics,* 68(2), pp. 203-224.

Teddlie, C., Lang, M. H. and Oescher, J. (1995) 'The Masking of the Delivery of Educational Services to Lower-Achieving Students', *Urban Education,* 30(2), pp. 125-149.

Teddlie, C. and Reynolds, D. (2000) *The international handbook of school effectiveness research.* Routledge.

Televantou, I., Marsh, H. W., Kyriakides, L., Nagengast, B., Fletcher, J. and Malmberg, L.-E. (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), pp. 75-101.

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), pp. 45-55.

TGAT (1988) *National curriculum: Task Group on Assessment and Testing : three supplementary reports / Welsh Office Great Britain, Department of Education and Science Great Britain, Task Group on Assessment and Testing.* London: London : HMSO.

Thomas, S. (1998) 'Value-added measures of school effectiveness in the United Kingdom', *Prospects,* 28(1), pp. 91-108.

Thomas, S. (2001) 'Dimensions of secondary school effectiveness: Comparative analyses across regions', *School Effectiveness and School Improvement,* 12(3), pp. 285-322.

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), pp. 261-295.

Thomas, S., Sammons, P., Mortimore, P. and Smees, R. (1997) 'Stability and Consistency in Secondary Schools' Effects on Students' GCSE Outcomes over Three Years', *School Effectiveness and School Improvement,* 8(2), pp. 169-197.

Thomas, S., Smees, R., MacBeath, J., Robertson, P. and Boyd, B. (2000) 'Valuing Pupils Views in Scottish Schools', *Educational Research and Evaluation,* 6(4), pp. 281-316.

Timmermans, A. C., Bosker, R. J., de Wolf, I. F., Doolaard, S. and van der Werf, M. P. C. (2014) 'Value added based on educational positions in Dutch secondary education', *British Educational Research Journal,* 40(6), pp. 1057-1082.

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), pp. 393-413.

Trafimow, D. and Rice, S. (2009) 'A test of the null hypothesis significance testing procedure correlation argument', *The Journal of general psychology,* 136(3), pp. 261-270.

Trochim, W. M. (1984) *Research design for program evaluation: The regression-discontinuity approach.* Sage Newbury Park, CA.

Trust, T. S. (2014) *Extra-curricular Inequalities.* Online: The Sutton Trust. Available at: http://www.suttontrust.com/researcharchive/enrichment-brief/.

Tymms, P. (1996) 'Theories, models and simulations: school effectiveness at an impasse', *Merging Traditions: The Future of Research on School Effectiveness and School Improvement. J. Gray, D. Reynolds, C. Fitz-Gibbon and D. Jesson. London & New York, Cassell*, pp. 121-135.

Tymms, P. (1999) 'Baseline assessment, value-added and the prediction of reading', *Journal of Research in Reading,* 22(1), pp. 27-36.

Tymms, P. and Albone, S. (2002) 'Performance indicators in primary schools', *School improvement through performance feedback*, pp. 191-218.

Tymms, P. and Dean, C. (2004) 'Value-added in the primary school league tables: a report for the National Association of Head Teachers', *Durham: CEM Centre.*

Tymms, P., Merrell, C. and Jones, P. (2004) 'Using baseline assessment data to make international comparisons', *British educational research journal,* 30(5).

van der Werf, G., Opdenakker, M. C. and 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), pp. 447-462.

Vanlaar, G., Denies, K., Vandecandelaere, M., Van Damme, J., Verhaeghe, J. P., Pinxten, M. and De Fraine, B. (2013) 'How to improve reading comprehension in high-risk students: effects of class practices in Grade 5', *School Effectiveness and School Improvement,* 25(3), pp. 408-432.

Vardardottir, A. (2013) 'Peer effects and academic achievement: a regression discontinuity approach', *Economics of Education Review,* 36, pp. 108-121.

Verachtert, P., Van Damme, J., Onghena, P., Ghesquiere, P., Verachtert, P., Van Damme, J., Onghena, P. and Ghesquiere, P. (2009) 'A seasonal perspective on school effectiveness: evidence from a Flemish longitudinal study in kindergarten and first grade', *School Effectiveness and School Improvement,* 20(2), pp. 215-233.

Verhaeghe, G., Schildkamp, K., Luyten, H. and Valcke, M. (2015) 'Diversity in school performance feedback systems', *School Effectiveness and School Improvement*, pp. 1-27.

Verhaeghe, G., Vanhoof, J., Valcke, M. and Van Petegem, P. (2010) 'Using school performance feedback: perceptions of primary school principals', *School Effectiveness and School Improvement,* 21(2), pp. 167-188.

Visscher, A. J. (2001) 'Public school performance indicators: Problems and recommendations', *Studies in Educational Evaluation,* 27(3), pp. 199-214.

Visscher, A. J. and Coe, R. (2002) *School improvement through performance feedback.* Routledge.

Visscher, A. J. and Coe, R. (2003) 'School performance feedback systems: Conceptualisation, analysis, and reflection', *School Effectiveness and School Improvement,* 14(3), pp. 321-349.

von Hippel, P. T. (2009) 'Achievement, Learning, and Seasonal Impact as Measures of School Effectiveness: It's Better to Be Valid than Reliable', *School Effectiveness and School Improvement,* 20(2), pp. 187-213.

Wainer, H. and Robinson, D. H. (2003) 'Shaping up the practice of null hypothesis significance testing', *Educational Researcher,* 32(7), pp. 22-30.

Waldegrave, H. and Simons, J. (2014) 'Watching the Watchmen: The future of school inspections in England', *London: Policy Exchange*.

Walsh, E. and Isenberg, E. (2015) 'How Does Value Added Compare to Student Growth Percentiles?', *Statistics and Public Policy,* 2(1), pp. 1-13.

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

West, A. and Pennell, H. (2000) 'Publishing school examination results in England: incentives and consequences', *Educational Studies,* 26(4), pp. 423-436.

Whetton, C. (2009) 'A brief history of a testing time: national curriculum assessment in England 1989-2008', *Educ. Res.,* 51(2), pp. 137-159.

White, P. (2014) 'A response to Gorard', *The Psychology of Education Review,* 38(1), pp. 24-28.

Wiliam, D. (2010) 'Standardized testing and school accountability', *Educational Psychologist,* 45(2), pp. 107-122.

Willms, J. D. (2003) *Monitoring school performance: A guide for educators.* Routledge.

Wilson, D., Piebalga, A., Wilson, D. and Piebalga, A. (2008) 'Performance measures, ranking and parental choice: An analysis of the English school league tables', *Int. Public Manag. J.,* 11(3), pp. 344-366.

Woodhouse, G. and Goldstein, H. (1988) 'Educational performance indicators and LEA league tables', *Oxford Review of Education,* 14(3), pp. 301-320.

Appendices

Appendix A

Materials related to Chapter 2, Section 2.3 on Value-Added designs

A1 Technical Details: School-Level Value-Added Models

Basic Model Specification (Figure 2.4.2a)

A simple value-added model for k schools (j) based on aggregated school-level data can be specified as follows:

(1a) $\bar{Y}_j = \alpha + \beta_1 \bar{X}_{1j} + r_j + u_j \qquad j = 1, 2, \ldots, k$

Where $\bar{Y}_j$ is the school mean of the pupil performance scores;

α is a constant intercept term;

$\bar{X}_{1j}$ is a measure of prior attainment, again aggregated to school-level;

β_1 is the prior attainment coefficient, estimating a linear relationship between $\bar{Y}_j$ and $\bar{X}_{1j}$;

r_j is the school effect, this is a latent variable estimated by the model residual; and

u_j is a random error term, assumed to have an expected value of zero (see the multi-level modelling section for further details on the variance).

The latent school effect, r_j, is not entered as a variable in the model but is estimated from the model residual (see below). Equation 1a estimates the linear relationship between cohort average performances at two different time points. This linear relationship is shown in Equation 1b, where $\hat{Y}_j$ is the expected score for school j.

(1b) $\hat{Y}_j = \alpha + \beta_1 \bar{X}_{1j} \qquad j = 1, 2, \ldots, k$

Value-added captures this difference using the model residual r_j which is the difference between the actual value of $\bar{Y}_j$, as in Equation 1a, and the expected value, given by $\hat{Y}_j$ in Equation 1b. Algebraically, this is as follows:

(1a) $\bar{Y}_j = \alpha + \beta_1 \bar{X}_{1j} + r_j + u_j \qquad j = 1, 2, \ldots, k$

(1b) $\hat{Y}_j = \alpha + \beta_1 \bar{X}_{1j}$

Substituting the predicted value, $\hat{Y}_j$ in Equation 1b for the observed linear relationship in 1a:

(2a) $\bar{Y}_j = \hat{Y}_j + r_j + u_j$

Subtracting the predicted value, $\hat{Y}_j$, to find the difference between the school performance and the expected performance:

(2b) $Y_j - \hat{Y}_j = r_{\ j} + u_j$

What this process has ostensibly achieved is to remove the predictable effect of prior attainment (as captured in $\hat{Y}_j$) from the variation between schools, leaving a measure of relative school performance.

Extending the Model: Alternative Functional Forms (Figure 2.4.2b)

As described in Chapter 2, Section 2.3.2, the model can be adapted to take alternative functional forms into account, such as a quadratic term to capture a non-linear relationship. This change is show in shown in Equation 3, below:

(3) $\bar{Y}_j = \alpha + \beta_1 \bar{X}_{1j} + \beta_2 \bar{X}^2_{2j} + r_j + u_j \qquad j = 1, 2, \ldots, k$

Note on the Problems Associated with the use of School-Level Data

Early methodological debates addressed the issue of whether one ought to estimate school value-added using school-level scores or if estimates calculated and then derived from pupil-level data were more appropriate, concluding that pupil-level data were required to meaningfully compare school performances (Aitkin and Longford, 1986, Woodhouse and Goldstein, 1988, Raudenbush and Willms, 1995). This research highlighted the potential difficulties with using aggregated school data to create value-added estimates (Snijders and Bosker, 2011), as follows:

1. *Less efficient estimators will be produced* (Raudenbush and Willms, 1995). Identical school mean scores can arise from different underlying pupil distributions so the ability to produce estimators at pupil-level allows more precise fitting of relationships between performance and the factors included in the model.
2. *School-level data ignores within-school differences:* Using aggregate school data prevents analysis of within school differences and so risks an ‘ecological fallacy’ in the interpretation of the results in which characteristics of the school overall are assumed to apply equally to all pupils in the school (Aitkin and Longford, 1986, p.11).

3. *Relationships at different levels potentially conflict:* It is possible for relationships between factors and performance to act in a different or even opposite direction at pupil- and school-level (Dettmers et al., 2009, Snijders and Bosker, 2011). In a seminal paper leading to a widespread adoption of multi-level models, Aitkin and Longford (1986: p42), despite not finding this to be a problem in their own data, issued the following warning: "… There is no reason in general to expect [aggregated school-level and pupil-level] models to give results that are at all comparable. In these circumstances, reliance on [aggregated school-level models] is dangerous at best, and disastrous at worst."

Aitkin, M. and Longford, N. (1986) 'Statistical modelling issues in school effectiveness studies', *Journal of the Royal Statistical Society. Series A (General)*, pp. 1-43.

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), pp. 375-405.

Raudenbush, S. W. and Willms, J. (1995) 'The Estimation of School Effects', *Journal of Educational and Behavioral Statistics,* 20(4), pp. 307-335.

Snijders, T. and Bosker, R. (2011) *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* 2nd edn. London: Sage.

Woodhouse, G. and Goldstein, H. (1988) 'Educational performance indicators and LEA league tables', *Oxford Review of Education,* 14(3), pp. 301-320.

A2 Technical Details: Pupil-Level Value-Added Models

Model Specification 1- OLS

One specification of a pupil level model uses ordinary least squares (OLS) multiple regression. This applies the school-level approach described above, to pupil-level data, as follows:

$$(4) \qquad y_{ij} = \alpha + \beta_1 x_{1ij} + r_{ij} + u_{ij} \qquad i = 1, 2, \ldots, n \quad j = 1, 2, \ldots, k$$

Pupils (i) are nested in schools (j) and the relationship between performance (y) and prior performance (x) is estimated at pupil-level. As in the school-level example, the regression equation (as given by the right-hand-side and disregarding the latent pupil-level value-added (r_{ij}) and the unobserved error, (u_{ij}) can be used as an estimate of y, ($\hat{y}_{ij}$). The difference between this estimate and the actual values of y can be used to generate pupil-level residuals r_{ij} (i.e. pupil-level value-added scores), as was the case at school-level. To get school-level value-added score, the school mean of these pupil-level residuals can be calculated.

Extending the Model: Contextual Variables

Equation 4, above, can be extended to account for a greater range of non-school factors. This is shown in equation 9 below which gives a simple example of a model which has been extended to adjust for 3 non-school factors (x_1, x_2 and x_3):

$$(9) \qquad y_{ij} = \alpha + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3ij} + u_{ij} \qquad i = 1, 2, \ldots, n \quad j = 1, 2, \ldots, k$$

Any number of non-school factors can be included. Also, alternative functional forms (see above) and interactions between non-school factors can be considered if required.

Technical Note on the Specification of Progress 8

As described in Chapter 2, Section 2.3.3, the Progress 8 measure calculates the mean pupil KS4 point score for every possible pupil KS2 score for all pupils in the national cohort. The key tension in this approach is that the prior attainment score (in this case at KS2) must be a) sufficiently fine-grained to allow an efficient estimator of KS4 performance (too crude a measure would leave a considerable amount of variation within each gradation) while b) not allowing the mean for each gradation to be based on too few pupils. In the middle of the distribution, there are large numbers of pupils in the national cohort even when using very fine-grained KS2 data. However, at the edges of the performance distribution the number of pupils for each KS2 score falls very low and the estimation line gives a 'saw-toothed' pattern (Burgess and Thomson, 2013: 16). This difficulty is addressed in the Progress 8 measure in two ways: First, the KS2 score is expressed to 1 decimal place. Second, scores are truncated at the upper and lower extremes of the distribution such that several gradations in the KS2 score are grouped and a single mean KS4 score is produced for the range of scores (Burgess and Thomson, 2013).

Burgess, S. and Thomson, D. (2013) *Key Stage 4 Accountability: Progress Measure and Intervention Trigger, Technical Annex: Techniques for producing an unbiased national pupil progress line*, http://www.bristol.ac.uk/cubec/portal/: BUBeC, University of Bristol.

A3 Technical Details: Multi-level Value-Added Models

Model Specification

In terms of specification, a key difference between a multi-level model and the OLS pupil-level model described in equation 4 (see last section) is the partitioning of the residual into a school-level term and a pupil-level term. Drawing on Goldstein (1997) and remaining consistent to the above notation, a simple multi-level model can be specified as follows:

$$(5) \qquad y_{ij} = \alpha + \beta_1 x_{1ij} + \overline{r_j} + r_{ij} + u_{ij} \qquad i = 1, 2, \dots, n \quad j = 1, 2, \dots, k$$

The performance variables (y and x_1), intercept (α) and random error term (u_{ij}) are unchanged from equation 4. However, the model now divides the residual into two 'random effects': the school-level 'value-added' ($\overline{r_j}$) and the pupil-level deviation from this school average (r_{ij}). Within the multi-level modelling framework, statistical programmes do not treat these deviations as being independent at pupil-level (see technical note, below).

Extending the Model: Allowing school-level variation in estimates

The other advantage of multi-level models is the ability to allow relationships to vary by school. This can be demonstrated using the following model, again following Goldstein (1997). In this example, the coefficient on the prior attainment variable, x_{1ij}, is allowed to vary by school, denoted by the new subscript on β1:

$$(6) \qquad y_{ij} = \alpha + \beta_{1j} x_{1ij} + \overline{r_j} + r_{ij} + u_{ij} \qquad i = 1, 2, \dots, n \quad j = 1, 2, \dots, k$$

Where β_{1j} comprises of a general prior attainment coefficient, β_1, and a school specific 'differential' school effectiveness coefficient d_j, as follows:

$$(7) \qquad \beta_{1j} = \beta_1 + d_j$$

By substitution and expanding the brackets the model can be shown to have two terms for the effect of prior performance (x_1): a 'fixed effect' $\beta_1 x_{1ij}$ which is constant for all schools and a 'random effect' $d_j x_{1ij}$ which gives each school's deviation in the slope coefficient and therefore the extent to which value-added varies according to pupil prior ability.

$$(8) \qquad y_{ij} = \alpha + \beta_1 x_{1ij} + d_j x_{1ij} + \overline{r_j} + r_{ij} + u_{ij}$$

Goldstein, H. (1997) 'Methods in School Effectiveness Research', *School Effectiveness and School Improvement,* 8(4), pp. 369-395.

Technical Note on the Problems Associated with the use of Pupil-Level Models (such as described in Appendix A2, above)

This inability of pupil-level models to capture the hierarchical structure has two key negative consequences (Goldstein, 1997):

1. *Limiting the analytical possibilities of the model:* Use of pupil-level models prevents relationships within the data to vary by school. It would be possible to estimate these manually from the pupil-level residuals but this would prove difficult and/or time-consuming for a large number of schools or in more complex analyses. Also, as noted in the technical note in Appendix 1A, relating to the problems associated with school-level models, another problem is that factor relationships may be different at different levels of analysis; in such circumstances, a multi-level approach is needed to detect any differences (Snijders and Bosker, 2011). Again, while this problem can be addressed in pupil-level models to some degree using school-level variables and interaction terms, this is generally unfeasible for larger samples or more complex models.
2. *Violation of independence assumptions:* The second limitation of using the pupil-level models relates to non-independence of observations violating the assumptions underpinning statistical tests within the model (Aitkin and Longford, 1986). Pupil-level models assume that two pupils from the same school are not expected to be any more similar than two pupils from different schools. This has implications for statistical tests because the model treats 100 schools with 100 pupils per school as being a sample of 10,000 independent pupils and so assumes more statistical information than is available. When one allows for correlation between pupil-level errors within schools, larger standard errors are produced. Without doing this, statistical tests are ‘biased and typically over-optimistic’ (Goldstein, 1997: 377). The size of the standard errors is the main difference between equivalent estimates produced in multi-level and a single-level models; where standard errors and the stringency of statistical tests tend to be higher in a multi-level framework (Snijders and Bosker, 2011).

Aitkin, M. and Longford, N. (1986) 'Statistical modelling issues in school effectiveness studies', *Journal of the Royal Statistical Society. Series A (General)*, pp. 1-43.

Goldstein, H. (1997) 'Methods in School Effectiveness Research', *School Effectiveness and School Improvement,* 8(4), pp. 369-395.

Snijders, T. and Bosker, R. (2011) *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* 2nd edn. London: Sage.

Appendix B

Materials related to the methodological survey in Section 3.2.1

B1 Methodological Survey Papers and their Categorisation

Table B.1a – Raw Data for Categorisation of Methods in Educational Effectiveness Research Papers Summarised in Chapter 3, Table 3.2.1b

	Study Design and Journal			Main Findings relate to…							
				Fixed Effects				Random Effects			
	Journal	Longitudinal	Cross-sectional	School/Leader Practices or Characteristics	Teacher Practices or Characteristics	Pupil Characteristics	System/other Practices or Characteristics	School Effects	Teacher Effects	Methodology	Interventions/ Systemic
(Al Otaiba et al., 2014)	JREE	✓									✓
(Altrichter and Kemethofer, 2015)	SESI		✓				✓				
(Anders et al., 2012)	SESI	✓			✓	✓	✓				
(Arshavsky et al., 2013)	SESI	✓		✓							✓
(Askell-Williams et al., 2012)	SESI	✓									✓
(Boonen et al., 2013)	SESI	✓			✓				✓		
(Chapman and Muijs, 2013)	SESI	✓		✓							
(de Bilde et al., 2012)	SESI	✓		✓							
(de Haan et al., 2012)	SESI	✓			✓						✓
(de Lange et al., 2013)	SESI		✓	✓			✓				
(Demanet and Van Houtte, 2012)	SESI		✓	✓		✓					
(Devos et al., 2012)	SESI		✓	✓							
(Dumay et al., 2013)	SESI	✓						✓		✓	
(Ebert et al., 2012)	SESI	✓				✓	✓				
(Ferrão and Couto, 2013)	SESI	✓						✓		✓	
(Gaertner et al., 2013)	SESI	✓									✓
(González and Jackson, 2012)	SESI	✓		✓			✓				
(Gottfried et al., 2013)	JREE	✓				✓					
(Gottfried, 2012)	SESI	✓				✓					
(Guarino et al., 2013)	JREE	✓			✓					✓	
(Gustafsson, 2013)	SESI	✓	✓		✓					✓	
(Hall et al., 2012)	SESI	✓		✓	✓	✓					
(Houtveen et al., 2013)	SESI		✓		✓						✓
(Isac et al., 2013)	SESI		✓	✓				✓			
(Isenberg et al., 2015)	JREE	✓							✓	✓	

(Table continued overleaf)

(Table B.1a continued)

	Study Design and Journal			Main Findings relate to…							
				Fixed Effects				Random Effects			
	Journal	Longitudinal	Cross-sectional	School/Leader Practices or Characteristics	Teacher/Class Practices or Characteristics	Pupil Characteristics	Wider/other Practices or Characteristics	School Effects	Teacher Effects	Methodology	Interventions/ Systemic
(Johansson et al., 2013)	SESI		✓		✓						
(Johnson et al., 2014)	JREE		✓						✓	✓	
(Kieffer, 2013)	JREE	✓		✓							
(Konstantopoulos and Sun, 2013)	SESI		✓		✓				✓		
(Lenkeit, 2012)	SESI	✓						✓		✓	
(Leucht et al., 2013)	SESI		✓	✓							
(Ma et al., 2013)	SESI		✓				✓				
(Melhuish et al., 2012)	SESI	✓		✓	✓						
(Miller et al., 2014)	JREE	✓				✓					
(Noyes, 2012)	SESI	✓		✓				✓			
(Othman and Muijs, 2012)	SESI		✓	✓			✓				
(Paterson et al., 2013)	SESI	✓				✓					✓
(Sammons et al., 2012)	SESI	✓		✓	✓	✓		✓			
(Tan, 2013)	SESI		✓	✓							
(Televantou et al., 2015)	SESI		✓					✓		✓	
(Tuytens and Devos, 2013)	SESI		✓	✓							
(Vanlaar et al., 2013)	SESI	✓			✓	✓					
(Walker et al., 2014)	SESI		✓	✓							
(Weinstein and Muñoz, 2013)	SESI		✓	✓							
(You, 2012)	SESI	✓		✓		✓					

B2 Secondary Reference List for Appendix B, Table B.1a

Al Otaiba, S., Kim, Y.-S., Wanzek, J., Petscher, Y. and Wagner, R. K. (2014) 'Long-Term Effects of First-Grade Multitier Intervention', *Journal of Research on Educational Effectiveness,* 7(3), pp. 250-267.

Altrichter, H. and Kemethofer, D. (2015) 'Does accountability pressure through school inspections promote school improvement?', *School Effectiveness and School Improvement,* 26(1), pp. 32-56.

Anders, Y., Grosse, C., Rossbach, H.-G., Ebert, S. and Weinert, S. (2012) 'Preschool and primary school influences on the development of children's early numeracy skills between the ages of 3 and 7 years in Germany', *School Effectiveness and School Improvement,* 24(2), pp. 195-211.

Arshavsky, N., Edmunds, J. A., Miller, L. C. and Corritore, M. (2013) 'Success in the college preparatory mathematics pipeline: the role of policies and practices employed by three high school reform models', *School Effectiveness and School Improvement,* 25(4), pp. 531-554.

Askell-Williams, H., Dix, K. L., Lawson, M. J. and Slee, P. T. (2012) 'Quality of implementation of a school mental health initiative and changes over time in students' social and emotional competencies', *School Effectiveness and School Improvement,* 24(3), pp. 357-381.

Boonen, T., Van Damme, J. and Onghena, P. (2013) 'Teacher effects on student achievement in first grade: which aspects matter most?', *School Effectiveness and School Improvement,* 25(1), pp. 126-152.

Chapman, C. and Muijs, D. (2013) 'Does school-to-school collaboration promote school improvement? A study of the impact of school federations on student outcomes', *School Effectiveness and School Improvement,* 25(3), pp. 351-393.

de Bilde, J., Van Damme, J., Lamote, C. and De Fraine, B. (2012) 'Can alternative education increase children's early school engagement? A longitudinal study from kindergarten to third grade', *School Effectiveness and School Improvement,* 24(2), pp. 212-233.

de Haan, A., Elbers, E., Hoofs, H. and Leseman, P. (2012) 'Targeted versus mixed preschools and kindergartens: effects of class composition and teacher-managed activities on disadvantaged children's emergent academic skills', *School Effectiveness and School Improvement,* 24(2), pp. 177-194.

de Lange, M., Dronkers, J. and Wolbers, M. H. J. (2013) 'Single-parent family forms and children's educational performance in a comparative perspective: effects of school's share of single-parent families', *School Effectiveness and School Improvement,* 25(3), pp. 329-350.

Demanet, J. and Van Houtte, M. (2012) 'Grade retention and its association with school misconduct in adolescence: a multilevel approach', *School Effectiveness and School Improvement,* 24(4), pp. 417-434.

Devos, G., Hulpia, H., Tuytens, M. and Sinnaeve, I. (2012) 'Self-other agreement as an alternative perspective of school leadership analysis: an exploratory study', *School Effectiveness and School Improvement,* 24(3), pp. 296-315.

Dumay, X., Coe, R. and Anumendem, D. N. (2013) 'Stability over time of different methods of estimating school performance', *School Effectiveness and School Improvement,* 25(1), pp. 64-82.

Ebert, S., Lockl, K., Weinert, S., Anders, Y., Kluczniok, K. and Rossbach, H.-G. (2012) 'Internal and external influences on vocabulary development in preschool children', *School Effectiveness and School Improvement,* 24(2), pp. 138-154.

Ferrão, M. E. and Couto, A. P. (2013) 'The use of a school value-added model for educational improvement: a case study from the Portuguese primary education system', *School Effectiveness and School Improvement,* 25(1), pp. 174-190.

Gaertner, H., Wurster, S. and Pant, H. A. (2013) 'The effect of school inspections on school improvement', *School Effectiveness and School Improvement,* 25(4), pp. 489-508.

González, R. L. and Jackson, C. L. (2012) 'Engaging with parents: the relationship between school engagement efforts, social class, and learning', *School Effectiveness and School Improvement,* 24(3), pp. 316-335.

Gottfried, A. E., Marcoulides, G. A., Gottfried, A. W. and Oliver, P. H. (2013) 'Longitudinal Pathways From Math Intrinsic Motivation and Achievement to Math Course Accomplishments and Educational Attainment', *Journal of Research on Educational Effectiveness,* 6(1), pp. 68-92.

Gottfried, M. A. (2012) 'The achievement effects of tardy classmates: evidence in urban elementary schools', *School Effectiveness and School Improvement,* 25(1), pp. 3-28.

Guarino, C., Dieterle, S. G., Bargagliotti, A. E. and Mason, W. M. (2013) 'What Can We Learn About Effective Early Mathematics Teaching? A Framework for Estimating Causal Effects Using Longitudinal Survey Data', *Journal of Research on Educational Effectiveness,* 6(2), pp. 164-198.

Gustafsson, J.-E. (2013) 'Causal inference in educational effectiveness research: a comparison of three methods to investigate effects of homework on student achievement', *School Effectiveness and School Improvement,* 24(3), pp. 275-295.

Hall, J., Sylva, K., Sammons, P., Melhuish, E., Siraj-Blatchford, I. and Taggart, B. (2013) 'Can preschool protect young childrens cognitive and social development? Variation by center quality and duration of attendance', *School Effectiveness and School Improvement,* 24(2), pp. 155-176.

Houtveen, A. A. M., van de Grift, W. J. C. M. and Brokamp, S. K. (2013) 'Fluent reading in special primary education', *School Effectiveness and School Improvement,* 25(4), pp. 555-569.

Isac, M. M., Maslowski, R., Creemers, B. and van der Werf, G. (2013) 'The contribution of schooling to secondary-school students' citizenship outcomes across countries', *School Effectiveness and School Improvement,* 25(1), pp. 29-63.

Isenberg, E., Teh, B.-r. and Walsh, E. (2015) 'Elementary School Data Issues for Value-Added Models: Implications for Research', *Journal of Research on Educational Effectiveness,* 8(1), pp. 120-129.

Johansson, S., Strietholt, R., Rosén, M. and Myrberg, E. (2013) 'Valid inferences of teachers' judgements of pupils' reading literacy: does formal teacher competence matter?', *School Effectiveness and School Improvement,* 25(3), pp. 394-407.

Johnson, M. T., Lipscomb, S. and Gill, B. (2014) 'Sensitivity of Teacher Value-Added Estimates to Student and Peer Control Variables', *Journal of Research on Educational Effectiveness,* 8(1), pp. 60-83.

Kieffer, M. J. (2013) 'Development of Reading and Mathematics Skills in Early Adolescence: Do K-8 Public Schools Make a Difference?', *Journal of Research on Educational Effectiveness,* 6(4), pp. 361-379.

Konstantopoulos, S. and Sun, M. (2013) 'Are teacher effects larger in small classes?', *School Effectiveness and School Improvement,* 25(3), pp. 312-328.

Lenkeit, J. (2013) 'Effectiveness measures for cross-sectional studies: a comparison of value-added models and contextualised attainment models', *School Effectiveness and School Improvement,* 24(1), pp. 39-63.

Leucht, M., Prieß-Buchheit, J., Pant, H. A. and Köller, O. (2013) 'Sometimes less is more: educational effectiveness of English as a foreign language instruction in German classrooms', *School Effectiveness and School Improvement,* 24(4), pp. 435-451.

Ma, X., Shen, J. and Krenn, H. Y. (2013) 'The relationship between parental involvement and adequate yearly progress among urban, suburban, and rural schools', *School Effectiveness and School Improvement,* 25(4), pp. 629-650.

Melhuish, E., Quinn, L., Sylva, K., Sammons, P., Siraj-Blatchford, I. and Taggart, B. (2013) 'Preschool affects longer term literacy and numeracy: results from a general population longitudinal study in Northern Ireland', *School Effectiveness and School Improvement,* 24(2), pp. 234-250.

Miller, A. C., Fuchs, D., Fuchs, L. S., Compton, D., Kearns, D., Zhang, W., Yen, L., Patton, S. and Kirchner, D. P. (2014) 'Behavioral Attention: A Longitudinal Study of Whether and How It Influences the Development of Word Reading and Reading Comprehension Among At-Risk Readers', *Journal of Research on Educational Effectiveness,* 7(3), pp. 232-249.

Noyes, A. (2013) 'The effective mathematics department: adding value and increasing participation?', *School Effectiveness and School Improvement,* 24(1), pp. 1-17.

Othman, M. and Muijs, D. (2012) 'Educational quality differences in a middle-income country: the urban-rural gap in Malaysian primary schools', *School Effectiveness and School Improvement,* 24(1), pp. 1-18.

Paterson, L., Gow, A. J. and Deary, I. J. (2013) 'School reform and opportunity throughout the lifecourse: the Lothian Birth Cohort 1936', *School Effectiveness and School Improvement,* 25(1), pp. 105-125.

Sammons, P., Hall, J., Sylva, K., Melhuish, E., Siraj-Blatchford, I. and Taggart, B. (2012) 'Protecting the development of 5–11-year-olds from the impacts of early disadvantage: the role of primary school academic effectiveness', *School Effectiveness and School Improvement*, pp. 1-18.

Tan, C. Y. (2013) 'Influence of contextual challenges and constraints on learning-centered leadership', *School Effectiveness and School Improvement,* 25(3), pp. 451-468.

Televantou, I., Marsh, H. W., Kyriakides, L., Nagengast, B., Fletcher, J. and Malmberg, L.-E. (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), pp. 75-101.

Tuytens, M. and Devos, G. (2013) 'How to activate teachers through teacher evaluation?', *School Effectiveness and School Improvement,* 25(4), pp. 509-530.

Vanlaar, G., Denies, K., Vandecandelaere, M., Van Damme, J., Verhaeghe, J. P., Pinxten, M. and De Fraine, B. (2013) 'How to improve reading comprehension in high-risk students: effects of class practices in Grade 5', *School Effectiveness and School Improvement,* 25(3), pp. 408-432.

Walker, A. D., Lee, M. and Bryant, D. A. (2014) 'How much of a difference do principals make? An analysis of between-schools variation in academic achievement in Hong Kong public secondary schools', *School Effectiveness and School Improvement,* 25(4), pp. 602-628.

Weinstein, J. and Muñoz, G. (2013) 'When duties are not enough: principal leadership and public or private school management in Chile1', *School Effectiveness and School Improvement,* 25(4), pp. 651-670.

You, S. (2012) 'Gender and ethnic differences in precollege mathematics coursework related to science, technology, engineering, and mathematics (STEM) pathways', *School Effectiveness and School Improvement,* 24(1), pp. 64-86.

Appendix C

Materials related to Study 1, Section 6.1

C1 Details of National Pupil Database variables used during analysis

Table C.1a – School-level National Pupil Database Variables Used during RQ 1.1.1

Year	Value-added Measure	KS2 Average Point score	Capped GCSE and equivalents point score.
2004	(VA2NEWE)	(KS2EVAIN)	(PTSCNEWE)
2005-2006	(CVA_KS2)	(CVA2APS)	
2007-2010	(CVA24SCO*SHRINK24)		(TTAPSCP)
2011-2014	(B8VAMEA)	(KS2APS)	

Table C.1b – School-level National Pupil Database Variables Used during RQ 1.1.2

Year	Value-added Measure	Average Point Score at KS1	Average Point Score at KS2
2011-2012	(EMVAMEAS)	(TKS1APS)	(TAPS)
2013-2014	(OVAMEAS)	(TKS1APS)	(TAPS)

C2 Model Specifications and Raw Output RQ 1.1.3

Multiple regression model for results presented in Table 6.1.1c:

Model 1.1.3a)

$$Best8VAMeasure_j = \beta_0 + \beta_1 SEN_j + \beta_2 EAL_j + \beta_3 FSMLA_j + \beta_4 Eligiblepupils_j + \beta_5 KS2APS_j + \beta_6 Coverage_j + \beta_7 PercentGirls_j + \varepsilon_j$$

Where the subscript, j, denotes schools;

β_0 is a constant intercept term;

ε_j is the model residual;

All other variables are described in Table 6.1.1c and the main text.

Note on the use of school-level data:

Model 1.1.3a is estimated using school-level data. It is worth making two points about this: first, there is some loss of statistical efficiency (i.e. the optimality of the estimator) from using school-level results. Second, it is widely held that standard errors (and so derivative measures of statistical significance) are more appropriately estimated within a multi-level framework (e.g. Snijders and Bosker, 2011). These limitations are noted but are thought unlikely to have an impact on the substantive interpretation of the results of this particular analysis. The advantages of a school-level analysis is that it uses publically available data, making it readily replicable; it avoids all the complexities of multi-level analysis, making it highly transparent; and it delivers all results at school-level, and so gives results most relevance to the immediate practical context of school accountability.

Snijders, T. and Bosker, R. (2011) *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* 2nd edn. London: Sage.

C3 Model Specifications and Raw Output RQ 1.1.3

Multilevel model to replicate the official 2013 KS2-4 VA measure - Specification:

This model uses the specification given in DfE (2013). Schools which were not mainstream or state maintained were dropped from the analysis. Also, there were 8 schools out of 3020 who had official VA scores of lower than -100 (in the context of a distribution spanning from approximately -50 to 50). These 8 schools were dropped from the analysis, it is likely that these schools have particular circumstances unknown to this researcher. The NPD data contains a variable indicating whether pupils were included in the VA calculations; this was used to remove pupils not included in the official measure from the analysis.

Model 1.1.3b) $Best8ScorePlusBonus_{ij} = \beta_0 + \beta_1 KS2APS_{ij} + \beta_2 KS2APS^2_{ij} + \beta_3\ KS2APS^3_{ij} + \beta_4 EngDev_{ij} + \beta_5 MatDev_{ij} + \varepsilon_{ij}$

Where the subscript j denotes schools;

The subscript i denotes pupils;

β_0 is a constant intercept term;

ε_j is the model residual;

$KS2APS_{ij}$ is the Key Stage 2 average point score;

$EngDev_{ij}$ and $MatDev_{ij}$ are the deviations from the Key Stage 2 average point score;

This was estimated within a multilevel framework in which pupil and school level random effects could be obtained from the model residuals, as follows:

Model 1.1.3c) $\varepsilon_{ij} = u_j + e_{ij}$

The R^2 of this model was calculated in an equivalent standard regression model, giving a value of 0.43. The residuals of this model were saved as replica VA scores. These scores had a pupil-level correlation of 0.999 (3dp) with the DfE official VA measure so were concluded to be, for all intents and purposes, identical.

Multilevel model to replicate a 2013 KS2-4 CVA measure - Specification:

Model 1.1.3d) $Best8ScorePlusBonus_{ij} = \beta_0 + \beta_1 KS2APS_{ij} + \beta_2 KS2APS^2{}_{ij} + \beta_3\ KS2APS^3{}_{ij} + \beta_4 EngDev_{ij} + \beta_5 MatDev_{ij} + \beta_5 SENgroup_{ij} + \beta_5 EALgroup_{ij} + \beta_5 FSMstatus_{ij} + \beta_5 Gender_{ij} + \beta_5 IDACI_{ij} + \beta_5 IDACI^2{}_{ij} + \varepsilon_{ij}$

Where, in addition to the variables in 1.1.3b:

$SENgroup_{ij}$ records one of three special educational needs categories. This was entered as a series of dummy variables using Stata's factor entry option.

$EALgroup_{ij}$ records one of three English as an Additional Language categories, again entered as a series of dummy variables.

$FSMstatus_{ij}$ is a binary variable recording pupils' free school meals status.

$Gender_{ij}$ is a binary variable recording pupils' gender

$IDACI_{ij}$ is the Income Deprivation Affecting Children Index. This is a measure of deprivation based on pupils' neighbourhood and is a continuous variable.

As before, a standard OLS regression model equivalent was calculated to estimate the model R squared value, giving an R^2 value of 0.48. The addition of the contextual variables, therefore, explains a further 5% of the variance in pupil performance.

DfE (2013) *A Guide to Value Added Key Stage 2 to 4 in 2013 School Performance Tables & RAISEonline*: DfE. Available at: http://www.education.gov.uk/schools/performance/2013/secondary_13/KS2-4_Performance_Tables_General_VA_Guide_2013_FINAL.pdf.

Multilevel model to replicate a 2013 KS2-4 CVA measure – Full Output:

```
Mixed-effects ML regression
Group variable: URN_SPR13                Number of obs         =     534686
                                         Number of groups      =     3020
Wald chi2(14)  =      468838             Obs per group: min    =     1
Log likelihood =      -2956838           avg                   =     177.0
Prob > chi2    =      0.00               max                   =     585
```

KS4 Best 8 Score Plus Bonus	Coef.	Std.Err.	z	P>z	[95% Conf. Interval]	
KS2 APS	27.04	0.82	32.86	0.00	25.42	28.65
KS2 APS Squared	0.02	0.00	46.79	0.00	0.02	0.02
KS2 APS Cubed	-1.17	0.03	-33.87	0.00	-1.24	-1.11
KS2 Maths Deviation	1.20	0.04	28.04	0.00	1.12	1.29
KS2 English Deviation	0.90	0.04	22.03	0.00	0.82	0.98
SEN Code						
None	19.18	0.29	65.44	0.00	18.60	19.75
School Action Plus	-35.35	0.43	-82.93	0.00	-36.18	-34.51
Statement of Special Educational Needs	-16.36	0.66	-24.96	0.00	-17.64	-15.08
EAL Group						
2 (EAL status)	30.84	0.33	93.65	0.00	30.20	31.49
3 (Unclassified)	-0.86	2.42	-0.36	0.72	-5.61	3.89
Free School Meals Eligible	-18.78	0.26	-72.81	0.00	-19.29	-18.27
Gender is Male	-18.72	0.18	-101.66	0.00	-19.08	-18.35
IDACI Score Squared	71.06	2.73	26.07	0.00	65.72	76.41
IDACI Score	-85.15	1.78	-47.97	0.00	-88.63	-81.67
Constant	124.83	6.43	19.41	0.00	112.23	137.44

Random-effects Parameters	Estimate	Std. Err.	[95% Conf. Interval]	
URN_SPR13: Identity				
var(constant)	441.18	12.05	418.19	465.44
var(Residual)	3659.23	7.1	3645.34	3673.16
LR test vs. linear regression: chibar2(01) 46581.75 Prob >= chibar2 = 0.0000				

C4 Model Specifications and Raw Output RQ 1.1.4

Multilevel model to replicate a 2012 KS1-2 CVA measure - Specification:

The replica KS1-2 CVA model used an identical model specification as the replica KS2-4 CVA measure, given in Appendix C3, above. The only difference is that prior attainment corresponds to KS1 rather than KS2. The analysis also used the same analytical steps above; although there are a couple of specific details particular to the KS1-2 level analysis which are noted here:

First, All pupils without an official value-added score were dropped to ensure the measure is as close as possible to the actual measure.

Second, schools with fewer than 5 pupils in their entry were removed, this meant the removal of 902 pupils (of 507,693) from the analysis. A small number of outlier schools were also removed (with scores lower than –9). This involved the removal of a further 75 pupils. These were all some way from the overall distribution and predominately from schools with very small cohort numbers. It is likely these schools are not representative of overall mainstream, maintained schools for reasons unknown to this researcher.

From this point, a replica VA score was created based on the actual DfE measure (DfE, 2011). In an OLS equivalent, the replica model had a R^2 of 0.64. The pupil-level correlation between the replica VA scores and the official scores was 0.998 (3DP) so the replica is almost identical to the actual measure. Then, the same contextual variables as at KS4 were entered into the replica model to produce the replica CVA model. The OLS equivalent model had a R^2 of 0.67, suggesting that a further 3% of the pupil attainment variance was explained with the contextual variables.

DfE (2011) *A Guide to Value Added Key Stage 1 to 2 in 2011 School and College Performance Tables*. Available at: http://www.education.gov.uk/schools/performance/2011/primary_11/2011_KS1-2_VA_Guide_FINAL.pdf.

Multilevel model to replicate a 2012 KS1-2 CVA measure – Full Output:

```
Mixed-effects ML regression
Group variable: URN_SPR12              Number of obs           =     503951
                                       Number of groups        =     14321
Wald chi2(14) =1160000                 Obs per group: min      =     4
Log likelihood = -1150700              avg                     =     35.2
Prob > chi2=0.00                       max                     =     211
```

KS2 APS Fine Graded	Coef.	Std. Err.	z	P>z	[95% Conf.	Interval]
KS1APS	0.81	0.02	48.93	0.00	0.78	0.84
KS1 APS Squared	0.00	0.00	-2.98	0.00	-0.01	0.00
KS1 APS Cubed	0.00	0.00	7.94	0.00	0.00	0.00
KS1 Maths Deviation	-0.32	0.00	-120.14	0.00	-0.33	-0.32
KS1 Reading Deviation	-0.04	0.00	-13.10	0.00	-0.05	-0.03
SEN Code						
None	1.24	0.01	107.67	0.00	1.22	1.26
School Action Plus	-0.84	0.02	-55.69	0.00	-0.87	-0.81
Statement of Special Educational Needs	-2.56	0.03	-98.07	0.00	-2.61	-2.51
EAL Group						
2 (EAL status)	0.71	0.01	58.69	0.00	0.69	0.73
3 (Unclassified)	-0.11	0.13	-0.84	0.40	-0.37	0.15
Free School Meals Eligible	-0.33	0.01	-35.48	0.00	-0.35	-0.32
Gender is Male	0.27	0.01	38.13	0.00	0.25	0.28
IDACI Score Squared	2.01	0.11	17.53	0.00	1.79	2.24
IDACI Score	-1.91	0.08	-25.19	0.00	-2.06	-1.76
Constant	14.91	0.07	217.97	0.00	14.77	15.04

Random-effects Parameters	Estimate	Std. Err.	[95% Conf. Interval]	
URN_SPR12: Identity				
var(Constant)	1.19	0.02	1.16	1.22
var(Residual)	5.32	0.01	5.30	5.34

LR test vs. linear regression: chibar2(01) = 71108.90 Prob >= chibar2 = 0.0000

C5 Model Specification and Output for RQ 1.3.1

Model used to create a deterministic relationship between KS2 and KS4 scores:

The following model is based on the replica of the official 2013 VA measure, specified above in Appendix C3 (Model 1.1.3b). To simplify the analysis, the maths deviation and English deviation scores were dropped from the value-added model. The effect of omitting the two deviation scores is minimal: the correlation between the predicted scores with and without is .995 (3dp). The advantage of dropping these is that, without these deviations, the model predicts a one-to-one correspondence between KS2 and KS4 scores. This model used to create KS4 predictions is specified as follows:

Model 1.3.1a) $Best8ScorePlusBonus_{ij} = \beta_0 + \beta_1 KS2APS_{ij} + \beta_2 KS2APS^2_{ij} + \beta_3\ KS2APS^3_{ij} + \varepsilon_{ij}$

Where the subscript j denotes schools;

The subscript i denotes pupils;

$Best8ScorePlusBonus_{ij}$ is the official KS4 attainment measure (in 2013);

β_0 is a constant intercept term;

ε_j is the model residual;

$KS2APS_{ij}$ is the Key Stage 2 average point score.

This was estimated within a multilevel framework in which pupil and school level random effects could be obtained from the model residuals, as follows:

Model 1.3.1b) $\varepsilon_{ij} = u_j + e_{ij}$

The model prediction was saved and subsequently treated as the actual KS4 score, giving a deterministic relationship between KS2 and KS4 scores. After introducing error, model 1.3.1a was used to estimate school value-added, this time using the deterministic scores with an added measurement error. Without error, the relationship between KS2 and KS4 was deterministic, so gave a model R^2 of 1. As these errors were introduced, the model R^2 dropped to 0.91 (small errors), to 0.75 (medium errors) and finally to 0.58 (large errors). These R^2 values were created in an OLS model equivalent of the multi-level model. Reduced model outputs are below:

Application of Model 1.3.1a in the presence of small error rates – reduced output:

VARIABLES	Best 8 KS4 with small error
KS2 APS with a small error added	0.709
	(0.209)
KS2 APS with a small error added squared	0.193
	(0.00871)
KS2 APS with a small error added cubed	0.00132
	(0.000118)
Constant	224.9
	(1.639)
Observations	535,890
Number of groups	3,028

Standard errors in parentheses

Application of Model 1.3.1a in the presence of medium error rates – reduced output:

VARIABLES	Best 8 KS4 with medium error
KS2 APS with a medium error added	-8.677
	(0.279)
KS2 APS with a medium error added squared	0.767
	(0.0114)
KS2 APS with a medium error added cubed	-0.00960
	(0.000152)
Constant	281.9
	(2.232)
Observations	535,890
Number of groups	3,028

Standard errors in parentheses

Application of Model 1.3.1a in the presence of large error rates – reduced output:

VARIABLES	Best 8 KS4 with large error
KS2 APS with a large error added	-1.532
	(0.281)
KS2 APS with a large error added squared	0.503
	(0.0112)
KS2 APS with a large error added cubed	-0.00753
	(0.000146)
Constant	245.6
	(2.303)
Observations	535,890
Number of groups	3,028

Standard errors in parentheses

C6 Model Specification and Output for RQ 1.3.2

Application of Model 1.3.1a using KS1-2 APS with small error rates – reduced output:

VARIABLES	KS2 APS
KS1 APS with a large error added	0.883
	(0.00774)
KS1 APS with a large error added squared	0.0105
	(0.000616)
KS1 APS with a large error added cubed	-0.000465
	(1.53e-05)
Constant	14.34
	(0.0307)
Observations	507,461
Number of groups	14,762

Standard errors in parentheses, Model R^2 in OLS equivalent = 0.87

Application of Model 1.3.1a using KS1-2 APS with medium error rates – reduced output:

VARIABLES	KS2 APS
KS1 APS with a large error added	0.888
	(0.0108)
KS1 APS with a large error added squared	0.00454
	(0.000834)
KS1 APS with a large error added cubed	-0.000434
	(2.02e-05)
Constant	15.65
	(0.0446)
Observations	507,461
Number of groups	14,762

Standard errors in parentheses, Model R^2 in OLS equivalent = 0.60

Application of Model 1.3.1a using KS1-2 APS with large error rates – reduced output:

VARIABLES	KS2 APS
KS1 APS with a large error added	0.872
	(0.0106)
KS1 APS with a large error added squared	-0.00717
	(0.000805)
KS1 APS with a large error added cubed	-0.000143
	(1.90e-05)
Constant	17.64
	(0.0460)
Observations	507,461
Number of groups	14,762

Standard errors in parentheses, Model R^2 in OLS equivalent = 0.36

Appendix D

Materials related to Study 2, Section 6.2

In all specifications within Appendix D1 the following notation is used: models examine the performance (P) of pupils (*i*), within cohorts (*j*), within schools (*k*) across the three years (*t*) for which data are available (see methods chapter, Section 5.5.2). Performance (P) refers to the teacher-assessed mathematics point scores at time periods 1, 2 and 3, where the time period is given in subscript.

Note that the models were run one cohort at a time. This means that subscripts referring to schools (k), below, actually refer to a cohort with the school. Similarly, in the case the RD design, which looks over two years at a time, school (k) refers to the consecutive cohorts for which the school effect estimate is desired pooled with the consecutive cohort below it.

D1 Model Specification and Selected Output: CVA Measure, Study 2

Model Specification – Contextualised Value-added (CVA) Model

Value-added scores were estimated using multi-level models where the residual variance was partitioned between school-level (u) and pupil-level (e) and the school-level residual recorded as the value-added score of the school for time period t. The school level residual can be considered a value-added score for the school as it gives the mean difference of the schools' pupils' actual scores from their predicted scored based on the model. The model is formally specified as follows:

Model 6.2CVAa) $P_{ik(t)} = \beta_0 + \beta_1 P_{ik(t-1)} + \beta_2 KS1_{ik} + \beta_3 FSM_{ik}$

$\beta_4 GENDER_{ik} + \varepsilon_{ik}$

School- and pupil-level residuals are calculated in a multilevel model such that:

Model 6.2CVAb) $\varepsilon_{ik} = u_k + e_{ik}$

Full Output from a selected CVA Model (CVA_i, Year4-T2)::

As an example of the model output from the CVA models, the first is given in full below. The following output concerns measure CVA_i, the estimate for Year 4 in time period 2.

```
Mixed-effects ML regression                        Number of obs          =    12225
Group variable: Z_ESTAB_Y2                         Number of groups        =      271


                                                   Obs per
Wald chi2(4)       =  43083.70                     group:          min  =       1
Log likelihood = -23622.253                                        avg  =     45.1
Prob > chi2        =    0.0000                                     max  =     237
```

T2 Maths Score	Coef.	Std. Err.	z	P>z	[95% Conf. Interval]	
T1 Maths Score	0.757	0.008	98.41	0	0.742	0.772
KS1 APS	0.276	0.007	37.34	0	0.261	0.290
Free School Meals Eligible	-0.181	0.041	-4.43	0	-0.261	-0.101
Gender is male	-0.205	0.031	-6.68	0	-0.265	-0.145
Constant	3.460	0.115	30.12	0	3.235	3.685

Random-effects Parameters	Estimate	Std. Err.	[95% Conf. Interval]	
Z_ESTAB_Y2: Identity				
var(Constant)	0.564	0.058	0.461	0.689
var(Residual)	2.663	0.034	2.596	2.731

LR test vs. linear regression: chibar2(01) = 1385.79 Prob >= chibar2 = 0.0000

D2 Model Specification and Selected Output: RD1 Measure, Study 2

Model Specification – Cross-Sectional Regression Discontinuity Model

The basic RD model (RD1) is formally stated as follows:

Model 6.2RD1$_a$) $P_{ik(t)} = \beta_{0k} + \beta_1 Age_{ik} + \beta_{2k} Year_{ik} + \varepsilon_{ik}$

Where β_{0k} is the intercept for school *k*;

Age_{ik} is the number of months between the August cut-off and month of birth of pupil *i* in school *k* (with July scored as 1, June as 2 and so forth).

$Year_{ik}$ takes the value of 1 for pupils in the upper year of the two consecutive cohorts and, therefore, estimates the added-year effect (see note below).

ε_{ik} is the model residual.

Note that β_{0k} and β_{2k} are school-specific. To estimate a school effect, the coefficient on the added-year effect (β_{2k}) can be separated into an overall mean effect of an added-year of schooling (β_{20}) and the school-specific deviation (S_{2k}) from this for school *k*, as follows:

Model 6.2RD1$_b$) $\beta_{2k} = \beta_{20} + S_{2k}$

This was calculated for each cohort (*j*) at a time. Each school-specific deviation, above, therefore corresponds to a given cohort in given school. This RD model is specified based on that in Luyten et al. (2009: 147), whose results this study replicates in relation to the two RD applications (cross-sectional and longitudinal). One technical difference to note is that the age-within year variable is used rather than an age relative to the cut-off variable. This saves computing a new variable for when pupils are used as the lower year as a baseline for performance estimates of the year above and as the upper year for estimates of their own performance. The difference in interpretation of the results this has is that the coefficient on the added-year effect gives the gross added-year effect. To estimate the added-year effect net of the age (maturity) effect, one can multiply the age effect coefficient by 12 and subtract this from the gross added-year effect (as is done in the results section RQ 2.1.1). This minor technical difference also applies to the other RD models, below.

Luyten, H., Tymms, P. and Jones, P. (2009) 'Assessing school effects without controlling for prior achievement?'. *School Effectiveness and School Improvement,* 20(2), pp. 145-165.

Full Model Output for a Selected Cross-Sectional Regression Discontinuity Model (RD1_i, Year4-T2):

```
Mixed-effects ML regression                    Number of obs     =    26253
Group variable: Z_ESTAB_Y2                     Number of groups   =      271

Wald chi2(2)       =   3234.28                 Obs per group:         min =        2
Prob > chi2        =    0.0000                                        avg =     96.9
Log likelihood = -68782.248                                           max =      506
```

T2 Maths Score	Coef.	Std. Err.	z	P>z	[95% Conf. Interval]	
Age within Year	0.170	0.006	28.87	0	0.158	0.181
Upper Year Flag	2.800	0.057	48.86	0	2.687	2.912
Constant	17.177	0.079	218.54	0	17.023	17.331

Random-effects Parameters	Estimate	Std. Err.	[95% Conf. Interval]	
Z_ESTAB_Y2: Unstructured				
var(Upper Year)	0.345	0.074	0.226	0.525
var(Constant)	1.081	0.122	0.867	1.349
cov(Upper Year, Constant)	-0.023	0.070	-0.160	0.115
var(Residual)	10.739	0.095	10.555	10.926

```
LR test vs. linear regression:       chi2(3) =  1658.08   Prob > chi2 = 0.0000
```

D3 Model Specification and Selected Output: RD2 Measure, Study 2

Model Specification – Regression Discontinuity Model with an Interaction Term:

RD2 extends RD1 to consider the effect of contextual variables on the size of the overall school effect. Several versions of the model were tested (see results chapter RQ 2.1.2, for further details). A model estimating the interaction effect between free school meals status and the added year effect as well as controlling for other contextual factors is specified below (Model 6.2RD2).

Model 6.2RD2$_a$) $P_{ik(t)} = \beta_{0k} + \beta_1 Age_{ik} + \beta_{2k} Year_{ik} + \beta_3 Gender_{ik} + \beta_4 FSM_{ik} + \beta_5 Gender_{ik} Year_{ik} + \beta_6 FSM_{ik} Year_{ik} + \varepsilon_{ik}$

Where, in addition to variables specified in *Model 6.2RD1$_a$:*

$Gender_{ijk}$ is a binary variable recording the pupil *i*'s gender

FSM_{ijk} is a binary variable recording free school meals status (a measure of poverty)

$FSM_{ijk}Year_{ijk}$ is the interaction effect between FSM and the added-year effect.

$\beta_5 Gender_{ik}Year_{ik}$ is the interaction effect between Gender and the added-year effect.

β_{0k} and β_{2k} are school-specific. To estimate a school effect, the coefficient on the added-year effect (β_{2k}) can be separated into an overall mean effect of an added-year of schooling (β_{20}) and the school/cohort-specific deviation (S_{2k}) in school *k*, as follows:

Model 6.2RD2$_b$) $\beta_{2k} = \beta_{20} + S_{2k}$

These school-specific deviations were then regressed on the cohort mean KS1 scores to ensure that the RD2 school-effects were not bias according to pupil prior attainment. The residual (ε_k) of this regression was saved as the RD2 measure.

Model 6.2RD2$_c$) $S_{2k} = \beta_0 + \beta_3 CohortMeanKS1_k + \varepsilon_k$

As before, as these were calculated for consecutive cohorts at a time (and the estimate saved only for the upper cohort), each score relates to a single cohort in a single school.

Full Model Output for a Selected Cross-Sectional Regression Discontinuity Model with Interaction and Term (RD2_i, Year4-T2):

```
Mixed-effects ML regression                     Number of obs      =     26253
Group variable: Z_ESTAB_Y2                      Number of groups   =       271

Wald chi2(6)       =   4376.63                  Obs per group:         min =         2
Prob > chi2        =    0.0000                                         avg =      96.9
Log likelihood = -68299.453                                            max =       506
```

T2 Maths Score	Coef.	Std. Err.	z	P>z	[95% Conf. Interval]	
Age within Year	.170	.006	29.53	0.000	.159	.182
Upper Year Flag	2.88	.072	40.26	0.000	2.74	3.020
Gender (is male)	-.300	.0570	-5.26	0.000	-.412	-.188
Free School Meals (FSM) Eligible	-1.447	.0758	-19.09	0.000	-1.596	-1.298
Upper Year*Gender	.063	.080	0.78	0.434	-.094	.220
Upper Year*FSM	-.350	.103	-3.39	0.001	-.552	-.148
Constant	17.617	.0777	226.65	0.000	17.465	17.769

Random-effects Parameters	Estimate	Std. Err.	[95% Conf. Interval]	
Z_ESTAB_Y2: Unstructured				
var(Upper Year)	0.322	0.070	0.210	0.493
var(Constant)	0.803	0.097	0.635	1.017
cov(Upper Year, Constant)	-0.091	0.062	-0.213	0.031
var(Residual)	10.382	0.092	10.204	10.563

LR test vs. linear regression: chi2(3) = 1141.13 Prob > chi2 = 0.0000

Estimates of Contextual and Interaction Effects from RD2 Models (Step 1, Model 6.2RD2a, above):

	Year 3-4	Year 4-5	Year 5-6	Year 7-8	Year 8-9	Year 3-4	Year 4-5	Year 5-6	Year 7-8	Year 8-9
	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)	(ix)	(x)
	T2 Maths Score					T3 Maths Score				
Age Within Year in	0.170	0.177	0.167	0.155	0.158	0.166	0.185	0.177	0.159	0.161
Months	(0.0058)	(0.0067)	(0.0074)	(0.0098)	(0.011)	(0.0065)	(0.0075)	(0.0083)	(0.0118)	(0.0131)
Gross Upper Year Effect	2.880	2.982	3.814	2.769	3.289	3.024	3.031	3.795	2.761	2.999
	(0.0715)	(0.0918)	(0.100)	(0.153)	(0.174)	(0.0844)	(0.0915)	(0.109)	(0.181)	(0.234)
Gender (is male)	-0.300	-0.244	-0.441	-0.316	-0.169	-0.129	-0.377	-0.257	-0.219	-0.194
	(0.0571)	(0.0668)	(0.0729)	(0.0985)	(0.113)	(0.0642)	(0.0740)	(0.0827)	(0.120)	(0.130)
Free School Meals (FSM)	-1.447	-1.782	-2.121	-2.443	-2.733	-1.484	-1.759	-2.231	-2.279	-2.674
	(0.0758)	(0.0854)	(0.0951)	(0.121)	(0.141)	(0.0853)	(0.0958)	(0.107)	(0.141)	(0.154)
Upper Year*Gender	0.0627	-0.200	-0.0907	0.172	-0.147	-0.240	0.120	-0.0847	0.0426	0.00720
	(0.0801)	(0.0934)	(0.103)	(0.139)	(0.159)	(0.0899)	(0.104)	(0.117)	(0.167)	(0.187)
Upper Year*FSM	-0.350	-0.297	-0.0495	-0.296	-0.350	-0.280	-0.437	0.168	-0.442	-0.186
	(0.103)	(0.119)	(0.133)	(0.172)	(0.201)	(0.117)	(0.132)	(0.148)	(0.196)	(0.223)
Constant	17.62	20.46	23.48	29.89	32.66	17.72	20.65	23.73	30.21	32.97
	(0.0777)	(0.0855)	(0.110)	(0.203)	(0.268)	(0.0890)	(0.102)	(0.114)	(0.223)	(0.299)
Observations	26,253	27,319	27,831	28,842	28,414	19,897	20,628	21,013	20,499	20,168
Number of groups	271	271	276	69	69	226	226	230	52	52

Standard errors in parentheses

Appendix E

Materials related to Study 3 and 4, Section 6.3 and 6.4

E1 Model Specification and Selected Output: CVA Measure, Study 3-4

Model Specification – Contextualised Value-added (CVA) Model

The final question within study 3 and the first question in study 4 required the creation of a value-added measure of performance. The following simple contextualised value-added measure was produced for use in both studies:

Model 6.3CVAa) $P_{ik} = \beta_0 + \beta_1 Prior_{ik} + \beta_2\ Prior^2_{ik} + \beta_4 GENDER_{ik} + \beta_3 FSM_{ik} + \varepsilon_{ik}$

Where the subscript k denotes schools;

The subscript i denotes pupils;

P_{ik} refers to teacher-assessed mathematics point scores;

β_0 is a constant intercept term;

$Prior_{ik}$ is the exam-assessed prior attainment score at the previous key stage (KS1 for primary school pupils and KS2 for the secondary school pupils);

$GENDER_{ik}$ is a binary variable recording pupil gender;

FSM_{ik} is a binary variable recording pupil free school meal eligibility;

ε_{jk} is the model residual.

Note that the models were run one cohort at a time. This means that subscripts referring to schools (k), below, actually refer to a cohort with the school. School- and pupil-level residuals are calculated in a multilevel model such that:

Model 6.3CVAb) $\varepsilon_{ik} = u_k + e_{ik}$

Example output from the model is given below:

Full Model Output for a Selected Contextualised Value-Added Model Used in Study 3 RQ 3.3 and Study 4 RQ 4.1:

```
                                     Number of obs      =    12981
Mixed-effects ML regression
Group variable: Z_ESTAB_Y1           Number of groups   =      271


Wald chi2(4)     =  20166.98         Obs per group:      min =       1

Prob > chi2      =    0.0000                             avg =    47.9

Log likelihood = -27191.304                              max =     247
```

Teacher Assessed Maths Score in T1	Coef.	Std. Err.	z	P>z	[95% Conf. Interval]	
KS1 Prior Attainment	0.627	0.030	20.69	0	0.567	0.686
KS1 Prior Attainment Squared	0.004	0.001	3.73	0	0.002	0.006
Gender (is male)	-0.780	0.034	-22.75	0	-0.847	-0.713
Free School Meals Eligibility	-0.105	0.046	-2.28	0.023	-0.196	-0.015
Constant	8.818	0.233	37.91	0	8.362	9.274

Random-effects Parameters	Estimate	Std. Err.	[95% Conf. Interval]	
Z_ESTAB_Y1: Identity				
var(Constant)	0.659	0.068	0.539	0.807
var(Residual)	3.701	0.046	3.611	3.793

LR test vs. linear regression: chibar2(01) = 1193.76 Prob >= chibar2 = 0.0000

Appendix F

Summary of Key Findings

F1 Specific Results Forming Headline Findings

Chapter 8, Section 8.1.3 presents ‘headline’ empirical findings. These summary statements are underpinned by numerous specific results. These more specific results are not presented in Chapter 8 to avoid repetition. Nevertheless, it is valuable to provide details of specific results which underpin the headline findings given. Note that even the following list is a summary of the numerous more fine-grained findings discussed in Chapter 6 and 7.

Key Findings of Study 1 relating to Bias and Error

1. *English School VA scores, despite the method controlling for prior attainment at pupil-level, are not wholly independent of prior attainment at school-level.*
   - There is a small *positive* correlation (r=0.29) between mean KS2 score of the intake and KS2-4 VA in the most recent (2014) data. In line with previous studies (Leckie and Goldstein, 2009), a grammar (selective) school effect was found which was found to be a marked example of a more general bias (see below for further details).
   - There is a small *negative* correlation (r=-0.18 between mean KS2 score of the intake and KS1-2 VA in the most recent (2014) data. Differences between measures with a baseline at the mid-point of the general school age range (e.g. mid primary to the end of primary, KS1-2) seem to differ from those which typically have the baseline at a separate school (e.g. end of primary to end of secondary, KS2-4). Differences relating to the baseline have also been found between KS3-4 and KS2-4 in previous research (Benton, 2004).

2. *Failure to include contextual variables in the more recent English VA measures has resulted in a number of substantial and systematic biases related to intake characteristics. These are consistent with factors found in school effectiveness literature.*
   - At KS2-4, roughly 35% of the variance in school-level value-added scores is accounted for by the following variables: special educational needs (SEN) (%), English as an addition language (%), disadvantage as measured by free school meals (FSM) and child looked after status (%), number of pupils in cohort, average cohort

prior attainment (at KS1/KS2), percentage of eligible pupils who are female and coverage (inclusion in the measure) as a percentage of eligible pupils. The largest factors within these were the rates of EAL and FSM. Rates of EAL were strongly positively associated with VA performance; FSM rates were strongly negatively associated with performance. A selective intake is associated with an increase in value-added of around 23 points (4 GCSE grades per pupil on the Best 8 KS4 measure). There is also a single-sex school effect over and above the grammar school effect (and considerable overlap in these two characteristics).

- If a contextual value-added model was produced to take into account key intake characteristics (above), typical schools could expect KS2-4 scores to change (upwards or downwards) by up to about 12 points, the equivalent of 2 GCSE grades per pupil at KS4. Schools with particularly (un)favourable intakes would see scores be adjusted by 12 to 30 points, or 2 to 5 GCSE grades per pupil at KS4.
- At KS1-2, key contextual variables (same as at KS2-4, above) account for about 10% of the (school-level) variance in the KS1-2 school VA measure. The most important contextual factors influencing the VA scores were found to be proportion of pupils with English as and additional language (EAL), rates of deprivation (measured by FSM rates) and mean KS1 cohort attainment.
- If a contextual value-added model was produced to take into account key intake characteristics (see point 3), typical schools could expect KS1-2 scores to change (upwards or downwards) by up to about 1 NC point, the equivalent of 4 months' progress *per pupil* at KS2. Over half of schools would change by 1/4 NC point (1 months' progress per pupil), and over 10% would change by 1/2 NC (2 months' progress per pupil).

3. *There are several specific problems within the National Pupil Database that pose serious threats to the validity of school value-added measures. These relate to both measures of attainment and other contextual variables*
    - The free school meals (FSM) variable was found to have a markedly different association with performance according to the percentage of pupils at a school who were eligible. Where the proportion of FSM pupils in a school is low, there is a strong negative association between FSM and performance. The relationship dramatically reduces for school with higher proportions and, for the schools with the highest rates

is even has a slight positive association with performance. This suggests that FSM is a poor proxy for characteristics of disadvantaged pupils which affect educational performance.

- Rates of missing data for contextual variables within the NPD are generally low (such as those in point 2, above). 0.2% and 0.5% of pupils are missing data for the FSM and IDACI at KS2, respectively. 1% and 1.3% are missing these at KS4. These rates may still prove problematic where missingness is concentrated by school, as previous research suggests it is (Gorard, 2012). Also, these figures do not reveal the extent to which pupils are misclassified or values have been replaced with a default value in the absence of other information.
- The most problematic case of missing data was the KS1 attainment figures where approximately 5% of pupils were missing a KS1 score. The KS1 data which were available were found to differ considerably from the normal distribution which would be expected from a robust measure of attainment.
- There are clear ceiling and floor effects in all of the measures of attainment examined. Earlier KS2 distributions show 5% of pupils with a score at the floor. This has improved with the more recent data (2012). The data seem to be improving over time but appreciable concerns remain and these will take a number of years to pass through the system.

4. *A simulation of pupil-level random measurement error suggests that even random measurement error will translate to substantial school-level errors. This is especially the case for KS2-4 value-added and is likely to apply to VA measures more generally.*
    - When a moderate amount of pupil-level random error was introduced into KS2 and KS4 scores, KS2-4 VA changed by up to about 2 GCSE grades per pupil. Typical rates in the school-level distribution were about half of this.
    - The introduction of random measurement error into KS1 and KS2 scores produces a more modest impact in the school-level KS1-2 VA results than at secondary level. Moderate error rates produce biases of up to 0.5 NC points. This equates to about 2 months’ progress per pupil. These figures are the edges of the distribution and rates for more typical schools will be lower than this.
    - Errors were found to produce systematic error patterns within value-added calculations, leading to a general systematic bias relating to intake average prior

attainment and spurious grammar school effects comparable with earlier estimates (see point 2, above). The evidence suggested that the tendency to create systematic error rates are a general feature of the value-added method which varies in severity according to characteristics of the specific data used.

Key Findings of Study 2 relating to Absolute School Effects and the Regression Discontinuity Design

5. *Progress is heavily patterned by year group in the English system. Year 6 results (as used in KS1-2 VA and KS2-4), for example, are considerably higher than other years.*
    - Analysis of teacher-assessed national curriculum (NC) levels for NC years 3 to 9 showed that score gains were heavily patterned by NC year. Year 6 results were considerably higher than gain scores for all other NC years.
6. *Value-added is successfully capturing differences in progress between pupils. At least to the extent to which the underlying measure of performance can accurately capture these.*
    - VA scores in these teacher-assessed data were found to be highly consistent with gain scores by pupil age within year. This suggests that VA is capturing genuine differences in relative performance to the extent that the underlying measure of attainment is a valid and reliable measure of performance.
7. *The regression discontinuity design is not suitable for comparing the effectiveness of different schools on an individual basis.*
    - The regression discontinuity design was found to be unreliable when estimating the performance of single cohorts/schools. This was because it required the lower consecutive year group to be a suitable control group for the year group in question but the volatility in cohort performances (see Study 4), prevented this from consistently holding.
8. *The regression discontinuity design shows considerable promise as a way of monitoring or comparing systems or large groups of schools.*
    - The regression discontinuity design was found, however, to be a viable design for estimating system-level absolute school effect. This design shows considerable promise for the monitoring of standards between key stages and for other outcomes, subject to the collection of suitably large sample sizes.

- The regression discontinuity design can be used to examine how progress varies according to pupil characteristics. RD estimates of the school effect were found to be associated with similar contextual factors as in the VA design (notably, rates of FSM).

Key Findings of Study 3 relating to Stability over Time

9. *Secondary school value-added scores have moderate stability over time. Rates of stability have been inflated by failure to account for contextual variables.*
   - Secondary level raw scores are very stable over a period of three years, although they have been affected by reforms to qualifications relating to GCSE equivalents. Raw attainment scores at KS2 are only moderately correlated over time.
   - The stability of KS2-4 value-added is considerably lower than that of raw scores. Estimates suggest correlations of approximately 0.8, 0.7 and around 0.5 for VA scores 1,2 and 3 years apart, respectively. The latter of these is likely to have been reduced to the qualifications reforms and so might expected to increase in future data. These scores are on the high side of estimates pertaining to the former CVA measure (Gorard et al., 2012), suggesting that the removal of contextual variables (see point 2, above) has reintroduced stabilising intake biases into the measure.
10. *Primary school value-added scores have moderate to very low stability over time. Correlations drop very quickly when at performance looking 1, 2, and 3 years apart.*
    - The stability of primary-level school VA scores ranges from moderate to very low. VA scores 1 year apart are moderately correlated but scores 2 and 3 years apart have low to very low correlations. Only 12% of the variance is common to both years when looking at scores 3 years apart.
11. *Instability is not strongly linked with initial poor performance. It is a general characteristic of the value-added scores across the performance range.*
    - At both secondary and primary level, there is no suggestion that rates of instability are due to disproportionate rates of change within lower performing schools. The instability appears to be across the board and related to measurement unreliability.
12. *Cohort performance over time has moderate levels of stability.*
    - The consistency of performance for the same cohorts within the same school across time was found to be moderate (0.43 to 0.73). The estimates are based on teacher-assessed data, which is likely to contribute to this instability. Nonetheless, these results

suggest that instability for schools is a more general problem of measurement unreliability rather than a result of cohorts with different characteristics passing through the examination year.

Key Findings of Study 4 relating to Consistency within and across Cohorts

13. *Consistency between the performances of different KS3 year groups in the same school at a point in time is moderate.*
    - The consistency of performance for different (year 7 to 9) cohorts within a secondary school (KS3) at a given time is estimated to be 0.6 and 0.45 for cohorts 1 and 2 years apart, respectively. As per the previous point, these estimates are based on teacher-assessed data for 48-71 schools but are considered a reasonable estimate of the consistency of performance between year 7 to 9 cohorts.
14. *Consistency in the performance of different KS2 cohorts in the same school at a point in time is moderate for adjacent cohorts and low to very low for cohorts 1 and 2 years apart.*
    - The consistency of performance for different (year 3 to 6) cohorts within a primary school at a given time is estimated to be 0.5, 0.3 and 0.2 for cohorts 1, 2 and 3 years apart, respectively. These are very low: performance of cohorts only 1 year apart is moderate; the performance of cohorts 2 or 3 years apart is hardly related. These estimates are based on teacher-assessed data in a large sample of primary schools.
15. *School value-added scores mask very large differences in pupil performance for schools across the performance range.*
    - Average score value-added scores were found to mask a wide range of pupil value-added scores at secondary and primary level and for schools at all performance levels. Even the best/worst performing schools tended to contain pupils who received high/low value-added scores.

Benton, T. (2004) 'Study of the performance of maintained secondary schools in England'. Available at: http://www.leeds.ac.uk/educol/documents/00003494.htm.

Gorard, S. (2012) 'Who Is Eligible for Free School Meals? Characterising Free School Meals as a Measure of Disadvantage in England', *British Educational Research Journal,* 38(6), pp. 1003-1017.

Leckie, G. and Goldstein, H. (2009) 'The limitations of using school league tables to inform school choice', *Journal of the Royal Statistical Society,* 172, pp. 835-851.