Essays on the Economics of Education

Joseph Oliver Regan-Stansfield
MSc and BSc (Hons.) Economics (University of Lancaster)

Supervised by Professor Colin P. Green and Professor Ian Walker

A thesis submitted to the University of Lancaster in partial fulfilment of the requirements of the degree of Doctor of Philosophy in Economics

October 2018
Department of Economics

Essays on the Economics of Education

Joseph Oliver Regan-Stansfield, MSc and BSc (Hons) Economics (University of Lancaster)
Supervised by Professor Colin P. Green and Professor Ian Walker

A thesis submitted to the University of Lancaster in partial fulfilment of the requirements of the degree of Doctor of Philosophy in Economics
October 2018, Department of Economics

Abstract

This thesis consists of three original research articles relating to schooling in England. The first research chapter evaluates a recent English education policy which encourages state primary schools to become academies: state-funded, non-selective, and highly autonomous establishments. The chapter investigates the causal effect of converting to an academy on assessment outcomes, and on entry-year intake composition. Unlike existing evidence focused on academies formed from failing secondary schools, no evidence is found of a converter academy effect on attainment for the average pupil. There is no evidence that becoming a converter academy affects the composition of the entry-year intake.

Standardised tests are a common, yet contentious, feature of many countries' schools. In April 2010, two UK teachers' unions boycotted mandatory age eleven standardised tests. The second research chapter uses a difference-in-differences strategy to estimate the effect of preparing for, but ultimately not completing, standardised tests on subsequent measures of attainment. The chapter finds evidence of a statistically significant adverse effect on age 14 teacher assessed attainment and age 16 secondary school qualification attainment. However, substantial treatment effect heterogeneity exists between sub-groups of pupils. Potential mechanisms are discussed, particularly the role of target setting.

Standardised tests often facilitate school accountability, and pupils usually receive grades (or other feedback) based on their performance. However, providing feedback is not necessary for school accountability. The third research chapter evaluates the effect of receiving integer grades based on a series of low-stakes standardised tests taken by eleven-year-olds in England. The chapter uses raw test marks, typically unobserved by pupils, and grade thresholds to implement a sharp regression discontinuity design. The results indicate that just passing the cut-off to achieve a higher grade in these tests leads to an improvement in secondary school qualification attainment. The estimated effect of just crossing a grade cut-off on secondary school attainment is typically larger for economically disadvantaged pupils. The chapter finds no evidence of an effect on school attendance.

Dedicated to my parents.

Declaration

I declare that this thesis *Essays on the Economics of Education*, and the work presented within it are my own. I confirm that:

1. this thesis has not been submitted for the award of a higher degree elsewhere;
2. this work was completed wholly in the candidature for a higher degree of the University of Lancaster;
3. chapter 2 was presented in the internal PhD seminar series within the Department of Economics, University of Lancaster, and has been published as "Does greater primary school autonomy improve pupil attainment? Evidence from primary school converter academies in England", *Economics of Education Review* (63), 167-179;
4. chapter 3 was presented in the NWDTC PhD Economics conference 2017, the 8th Annual International Workshop on the Applied Economics of Education, the 2nd IZA Workshop: The Economics of Education, and the Bristol Workshop on Assessment and Feedback. It was the subject of a seminar at the Department of Economics, Norwegian University of Science and Technology, and an earlier version was published as "The Impact of Standardised Testing on Later High Stakes Test Outcomes", *Lancaster University Economics Working Paper Series* (2017/009);
5. the research project underpinning chapter 3 became a collaboration with Dr Richard Murphy (Department of Economics, University of Texas at Austin) and Dr Gillian Wyness (UCL Institute of Education). Both collaborators agree that I have completed over seventy-five per cent of the work on the project;
6. chapter 4 was presented in the NWDTC PhD Economics conference 2018;
7. where the work of other authors is referred to, it is accurately referenced;
8. I acknowledge all sources of material help;
9. I grant permission to the institutional repository concerning online access to this thesis;
10. I reserve all rights afforded to myself as the author of this thesis under the Copyright, Designs and Patents Act 1998.

Joseph Oliver Regan-Stansfield

October 2018

Acknowledgements

I owe many thanks to my two supervisors, Colin Green and Ian Walker. They have provided invaluable guidance, support and motivation throughout my candidature. I am grateful for the various opportunities that they have presented to me during my research.

I must also thank Maria Navarro who has made a significant contribution to my development as a researcher since acting as my undergraduate dissertation supervisor in 2014. I am also grateful to Themis Pavlidis, Orestis Troumpounis and Vincent O'Sullivan for their respective contributions to my experiences within the department.

I am grateful to my fellow PhD candidates in the department for providing an enjoyable and stimulating research environment. I particularly thank my friends, Luke Wilson and Jack Higgins, for leaving their impression on my time at Lancaster.

Cartmel College has been my University home since 2011. My employment by the college as an Assistant Dean has been a highly rewarding vocation and a source of personal development. I would like to thank the senior members of the college for tolerating me for so long.

The original analysis presented in this thesis uses numerous data extracts from the Department for Education's National Pupil Database (NPD). I gratefully acknowledge the Department's effort in maintaining the database and thank the Department for providing access to junior researchers such as myself.

The Economic and Social Research Council (ESRC) has funded my PhD studies, via the North West Doctoral Training Centre (NWDTC). I much appreciate the financial support that they have provided.

Finally, I wish to thank my parents, John and Shirley, and probably my brother, Josh, for their love, support and encouragement long before and during my PhD. And Helena, for her love, support and patience with me during the good and bad PhD experiences.

List of tables

List of figures

List of abbreviations

AP	Advanced Placement
CAT	Cognitive Ability Test
DCSF	Department for Children, Schools and Families – the predecessor to the DfE
DfE	Department for Education – UK government department responsible for schooling since 2010
EBacc	English Baccalaureate
FSM	Free School Meals
GCSE	General Certificate of Secondary Education
HSEE	High School Exit Exam
KS1-4	Key Stage 1, 2, 3, 4
LSYPE	Longitudinal Study of Young People in England – survey dataset of English school pupils
NAHT	National Association of Headteachers
NC	National Curriculum
NCA	National Curriculum Assessments
NCL	National Curriculum Level
NCLB	No Child Left Behind
NPD	National Pupil Database – administrative dataset covering all of England’s state-funded schools and pupils
NUT	National Union of Teachers
SATs	Standardised Assessment Task – a colloquialism for National Curriculum Assessment tests
SEN	Special Educational Needs
STEM	Science, Technology, Engineering and Maths
TGAT	Task Group on Assessment and Teaching

Table of contents

Chapter One:

Introduction

When an individual invests in acquiring additional education, they will on average experience a tranche of positive returns. For example, the economics of education literature is awash with evidence of plausibly causal positive earnings returns to human capital accumulation (Angrist and Keueger, 1991; Devereux and Hart, 2010; Harmon and Walker, 1995). More recently, the literature has turned its attention towards non-pecuniary returns. Researchers have claimed that additional education causes individuals to live healthier lives (Lleras-Muney, 2005; Powdthavee, 2010; Silles, 2009), have greater old-age cognitive ability (Banks and Mazzonna, 2012), and have higher life satisfaction (Oreopoulos, 2007). More education also diminishes the risk that a person engages in criminal activity (Lochner and Moretti, 2004; Machin *et al.*, 2011) and increases the likelihood that they vote (Dee, 2004; Milligan *et al.*, 2004).

In addition to these private returns, there is abundant evidence of external returns to human capital accumulation. That is, other members of society benefit when one member acquires additional education. Parents' education has a positive causal influence on the future educational attainment of their children (Björklund *et al.*, 2006; Holmlund *et al.*, 2011). Meanwhile, economists have studied the positive association between average levels of education in workplaces and cities and an individual's earnings independent of the individual's level of education (Moretti, 2004a, 2004b; Rauch, 1993).

In developed economies, the public sector takes a central role in the provision of education. The existence of external returns to education combined with the non-existence of perfect credit markets means that the private sector would provide a sub-optimal level of education provision. Social norms of fairness and an aspiration for equality of opportunity also justify the public provision of education, particularly in the presence of intergenerational correlations in educational attainment.

In England during January 2018, 21,996 state-funded establishments educated 8 million pupils, equivalent to 93 per cent of the population of pupils in England (Department for Education, 2018). For the 2016-17 financial year, public sector expenditure on education functions in the United Kingdom was £87.2 billion which equates to 11.3 per cent of total public sector spending and 4.5 per cent of UK GDP (HM Treasury, 2017). Education is the third largest category of public sector expenditure after social protection and health. For the same financial year, the Department for Education estimated that the core budget for schools in England is worth £5,439 per pupil per school year (National Audit Office, 2016).

Given the sizable cost to society – taxpayers mainly – of providing education as well as the sizable social returns to educating the population, it is essential that schools, and other education establishments, are both effective and efficient. At best, ineffective schooling is an inefficient use of tax revenues. At worst, ineffective schooling represents a lost opportunity to enhance the future life course of the failed school children and the affected community more broadly.

For these reasons, there is rightly a great deal of interest in the performance of educational institutions and the wisdom of public education policy. Since attainment at school strongly correlates with future education participation (Bradley and Lenton, 2007) and attainment, interest in education policy is generally keenest when pre-schools and schools are concerned. This thesis presents three empirical investigations into the effects of recent schooling policies in England on the educational attainment of pupils. The research provides new evidence to policymakers on the impacts of features of English schooling institutions on the pupils these institutions seek to serve.

Chapter two presents new evidence on the effectiveness of adopting academy school status in the context of primary education (covering pupils between the ages of four and eleven). Academy schools are state-funded, non-selective state schools which are highly autonomous compared to other state schools and mostly independent of local education authorities. The academy school programme was initially an intervention strategy for secondary schools with a history of underperformance. However, the flagship schooling policy of the 2010-2015 government was the extension of the opportunity to become an academy – and thus the chance to enjoy greater autonomy – to all "well-performing" schools. As primary and secondary schools opted to become academies, head teachers and governors instigated the most substantial reorganisation of the English state school sector since the transition to comprehensive schooling.

The chapter uses a difference in differences strategy to estimate the causal effect of becoming an academy school on the attainment of primary school pupils. The research design exploits variation in when schools decide to become academies to use early converters as the treatment group and later converters as a control group. Using administrative data from the National Pupil Database (NPD), the analysis finds that when primary schools elect to become academies, there is neither a positive nor a negative impact on pupil attainment. Pupil attainment is measured by performance on standardised tests completed at the end of primary school and teacher assessments administered midway through primary school. This is in stark contrast to the existing evidence on academies which shows that academy status does improve attainment when applied to underperforming secondary schools. Given the significant costs involved in converting schools into academies, the chapter's findings suggest this is not a wise use of taxpayers' money. Further analysis also finds no evidence that the composition of the entry-year intake changes when a primary school becomes an academy.

While the effectiveness of academies is one of the more recent debates in English education policy, undoubtedly one of the longest-running and most contentious is that of the effectiveness of standardised testing. Since the mid-1990s the attainment of school children in English state schools has been evaluated in various subjects at critical stages of the schooling system. The results of these tests serve a dual purpose, primarily they offer a metric to evaluate school performance, but they are also used to provide feedback on pupil attainment. Proponents of standardised testing argue that test-based accountability incentivises schools to maximise their performance while simultaneously allowing regulators and parents to identify poorly performing schools.

However, critics link standardised tests to numerous adverse consequences. For example, there are concerns that standardised testing incentivises teachers to sacrifice broad curriculums and enjoyable teaching practices in favour of a narrowed focus on testable content and repetitive teaching activities designed to maximise test skills. Chapter three explores whether there is a private attainment return or penalty to participation in standardised tests. Specifically, the chapter investigates whether participation in standardised tests at the age of eleven (the end of primary school) causally impacts contemporaneous and subsequent teacher assessments and secondary school qualification attainment.

This is possible due to a widespread head teacher-led boycott of standardised tests that occurred in 2010. The first stage of the analysis involves using propensity score matching to identify schools that did not participate in the boycott but are similar to boycotting schools in other observable dimensions. The analysis' second stage involves estimating difference in differences models based on a panel of pupil level NPD data.

The models reveal evidence that participation in the boycott had a small adverse effect on the average pupil's attainment. However, the boycott participation effect varies significantly between different subgroups of pupils. The chapter speculates as to the mechanisms, focusing mainly on the role of attainment targets imposed by the government. The chapter also reports evidence that the boycott caused pupils to: change their subject choices at the end of secondary school, and be absent for slightly fewer secondary school lessons.

Chapter four evaluates the effect of providing pupils with feedback on their performance on the same series of standardised tests. The primary function of these tests is to produce a measure of school performance for school accountability. Neither pupils, parents nor teachers need to know how well each pupil has performed on the tests to achieve this central function. Feedback is provided based on the assumption that it is harmless and useful. Feedback takes the form of discrete numerical grades ranging from 2 to 5. These grades are inextricably mapped to national standards of expected attainment: level 4 denotes meeting the expected standard, level 5 denotes surpassing the expected standard and levels 2 and 3 denotes attainment below the expected standard. The testing authority awards grades based on pupils' raw test marks. Because pupils and parents generally only know the level that has been awarded and not the raw test mark achieved, pupils who perform near identically on the standardised tests can receive markedly different feedback on their attainment.

For example, consider two identical pupils. One scores 70 marks on the standardised maths test, the other scores one mark less. This could be due to a variety of trivial reasons. Perhaps the second pupil misread one question or made a minor arithmetic error. Maybe the pupils' tests were marked by different examiners who slightly differed in their method of awarding marks. If the threshold to achieve a level 5 grade is 70 marks, then the first pupil is told that their performance surpasses the national standard, whereas the second pupil is told that their performance meets the national standard. Neither pupil knows how “close” they were to receiving the other pupil’s grade.

As data is available on the raw test marks achieved by pupils as well as the grade thresholds, a regression discontinuity design can be used to uncover the effect of narrowly crossing the raw mark threshold for specific grades on the standardised tests. The analysis finds evidence that ‘just’ passing the raw mark threshold for the level 4 and level 5 grades has a slight positive impact on attainment in secondary school qualifications; the effect estimates are more substantial for pupils with a history of free school meal eligibility. The effect of English test grades on FSM ineligible pupils is precisely zero. Whereas for FSM eligible pupils, the effect estimates are statistically different from zero at the one per cent significance level. The chapter suggests that it is an undesirable feature of a schooling system that feedback on supposedly low-stakes standardised tests effects subsequent high-stakes secondary school qualification attainment. The chapter further contends that it is not desirable that the polarising effect of the test feedback be more pronounced for the most economically disadvantaged group of pupils. Further analysis finds no evidence of an effect on pupil effort as measured by school attendance during the school year after the standardised test.

Chapter Two:

Does Greater Primary School Autonomy Improve Pupil Attainment? Evidence from Primary School Converter Academies in England

A recent English education policy has been to encourage state primary schools to become academies: state-funded, non-selective, and highly autonomous establishments. Primary schools have been able to opt-in to academy status since 2010 and academies now account for twenty-one per cent of the primary sector. This chapter investigates the causal effect of becoming a converter academy on primary school assessment outcomes, and on entry-year intake composition. Unlike existing evidence focused on earlier academies formed from failing secondary schools, no evidence is found of a converter academy effect on attainment for the average pupil. Although, there is evidence of a slight positive effect on age eleven attainment for pupils eligible for free school meals. There is no evidence that becoming a converter academy affects the composition of the entry-year intake.

2.1 Introduction

The relentless growth in the number of academies represents arguably the most significant transformation of the English state school sector since the introduction of comprehensive schools in the mid-1960s. First introduced in the early 2000s, academies are state-funded, non-selective, yet highly autonomous schools operating mainly without local authority interference. Since the change in the UK government in 2010, the Department for Education (DfE) has overseen a process of "mass academisation" whereby all state schools have been encouraged to become academies. 65 per cent of secondary and twenty-one per cent of primary schools are now academies.

Some studies suggest that the high priority attached to the mass academisation programme is justified. The conversion of existing secondary schools between 1988 and 1997 into foundation schools, which enjoyed greater autonomy than their predecessors, was estimated to increase the proportion of pupils passing five GCSEs or more by five percentage points on average (Clark, 2009). A second intervention, the sponsored academies programme, established 200 sponsored academies between 2002 and 2010 to replace historically underperforming schools. Research suggests that the replacement of these schools with academies led to an improvement in pupils' GCSE attainment (Eyles *et al.*, 2018). Pupils attending these academies were also more likely to complete a degree following their schooling (Eyles *et al.*, 2016b). These two interventions were different and affected schools with dissimilar performance records, but both increased schools' autonomy and the affected pupils' attainment.

The existing body of research into academies focuses overwhelmingly on secondary sponsored academies established before 2010. Sponsored academies are far less prevalent than converter academies, the latter of which are formed by schools that voluntarily elect to become academies. These schools tend to be already well-performing and educate relatively advantaged pupils. Researchers have only recently turned their attention towards converter academies. For example, Eyles *et al.* (2017) and Worth (2015) both show that attainment in primary converter academies does not improve following academy conversion.

This chapter uses a difference-in-differences strategy to exploit the availability of data before and after conversions to identify the effect of becoming a converter academy on pupil attainment in primary schools. This chapter also considers whether voluntary academy conversion alters the composition of the primary schools' entry-year intake.

This chapter finds no evidence that the average pupil performs any better in reading or maths tests at the end of primary school because their school became a converter academy. However, evidence is uncovered that one sub-group of pupils, those eligible for free school meals (FSM), perform slightly better in age eleven maths and reading tests. There is also a small positive effect on age eleven reading attainment for schools that had the least autonomy before becoming a converter academy, but no effect is found for schools that were already relatively autonomous before conversion. No evidence is found that average pupil attainment at age seven is affected by converter academy status. Lastly, the composition of the entry-year intake does not appear to change with respect to several observable pupil characteristics following conversion.

This chapter informs a lively public debate over the merits of academies, which are opposed by most teacher unions, some local authorities and major opposition political parties. The debate was galvanised by the 2016 government white paper *Education Excellence Everywhere* which declared the DfE's aspiration for every English state school to become an academy (or be in the process of doing so) by 2020 (Department for Education, 2016).[1] While full academisation is no longer a policy priority, schools continue to become academies at a vast rate. The scale and speed of the reform are unprecedented. If the trend continues then, state-funded schools will be mostly independent of local government, and the English state schooling system will secure its position as the world's most decentralised.

Academies are relatively less prevalent in the primary sector than the secondary sector. Furthermore, the government has already ensured that many of the worst performing primary schools have become sponsored academies. As such, the most significant consequence of further academisation will be a substantial increase in the number of primary converter academies – the specific type of academy studied in this chapter.

The conversion process is known to place significant administrative and financial burdens on the DfE, local authorities and schools themselves. For example, the DfE incurred additional costs of £1bn due to the academies programme between April 2010 and April 2012 (National Audit Office, 2012). This includes one-off costs such as the £25,000 grant paid to schools to facilitate the conversion process, as well as the additional recurrent cost per open academy. In 2012/13, this was estimated at £260,000 per annum on average. At a time when the English state school sector is facing resource pressures, such as teacher shortages, and expecting other radical reforms, such as the introduction of a national school funding formula, this timely analysis is unable to provide evidence of any benefit from academy conversion to the average primary school pupil.

[1] The white paper stated that schools would be forced to become academies by 2022 even if this was against schools' wishes. A hostile backlash led to a policy revision whereby state schools would be encouraged but not compelled to become academies by 2022.

2.2 Institutional background

There are two broad types of state school in England: maintained schools and academies. Maintained schools receive funding and some professional and pupil-facing services from local education authorities (LEAs), to whom the government has historically delegated schooling provision. These authorities also set, or constrain, the policies and processes of their maintained schools; although the degree of control LEAs have over schools varies between different types of maintained school. The types of maintained school are, from least to most autonomous: community, voluntary-controlled, voluntary-aided and foundation schools. Academies, on the other hand, are funded directly by the DfE and are mostly independent of LEAs.

Academies recruit and contract their staff, unlike community and voluntary-controlled schools whose staff are employed by their LEAs. Academies may impose their own employment terms and can disregard nationally negotiated teacher pay and conditions. They also have considerable freedom in devising their curriculum which must be "broad and balanced" and include English, maths, science and religious studies (Department for Education, 2010). However, they do not have to follow the national curriculum in these subjects, unlike maintained schools who are bound to the full national curriculum. Academies set their admission policy unlike community and voluntary-controlled schools which are subject to an LEA admission policy.[2]

Maintained schools are run by a board of between 9 and 20 governors. In community schools, one-fifth of the governors are appointed by the LEA. In foundation, voluntary-aided and voluntary-controlled schools, a separate charitable (often faith-based) foundation appoints between one-quarter and a majority of the governors, reducing the

[2] However, admission policies must comply with the national School Admissions Code which forbids selection by ability.

LEA's control. Academies are governed by private charitable trusts independent of the LEA. These trusts set their own budget and policies, including the length of the school day and year. Academies are effectively the UK equivalent of charter schools in the USA.

Officially academies should not be funded advantageously relative to maintained schools. However, a 2012 National Audit Office survey of converter academy head teachers found that 77 per cent of academies converted to obtain more funding for front-line education (National Audit Office, 2012). Academies and maintained schools receive comparable Dedicated Schools Grant (DSG) funding which covers mainstream education provision and is the primary source of funding for schools. However, there has been a historical disparity between academies and maintained schools in respect of funding for auxiliary functions. LEAs centrally provide some services to maintained schools that academies need to procure independently. Academies formerly received an additional grant to provide these functions.[3] It boosted some academies' budgets by more than 10 per cent and was widely considered to overcompensate academies. This grant has now been replaced with the Educational Services Grant (ESG), paid on a common per-pupil rate. Since the 2015/16 school year academies and maintained schools are financed on a comparable basis (Department for Education, 2014).

An understanding of the academy sector's expansion is vital as academies can be grouped into two very different subcategories. By 2000 it was apparent to the then Labour government that there was a pervasive problem of under-performance, poor behaviour and low aspirations in inner-city secondary schools. The government's solution was to inject innovative management and private sector best practices into

[3] The grant was known as the Local Authority Central Spend Equivalent Grant.

these failing schools. The government set about matching selected schools to sponsors – an individual, business or charitable organisation – who would influence the management, ethos, and curriculum of the school as it re-opened as an academy. These original academies would often occupy new or extensively refurbished facilities co-financed by the sponsor.[4] Between 2002 and 2010, 203 such academies were established; all were secondary schools, and most were former maintained schools.[5] Academies founded due to the DfE imposing academy status on failing schools are now referred to as sponsored academies.

The composition of the academy sector changed dramatically following the formation of the Conservative-Liberal Democrat coalition government in May 2010. The new Secretary of State for Education was keen to offer academy freedoms to schools that were not failing or located within inner-city or deprived neighbourhoods. In July, the Academies Act 2010 became one of the fastest pieces of education legislation to be adopted by the UK parliament. It gave all schools the option to voluntarily become academies from the 2010/11 school year, ultimately leading to the first primary academies. Academies formed from schools which voluntarily chose to become academies are known as converter academies.

Schools rated "outstanding" by OFSTED, the national school inspections body, originally had their applications pre-approved meaning they could become academies from September 2010. From April 2011, all applications from "well-performing" schools received priority from the DfE.[6] The application process is relatively swift, with

[4] This requirement was subsequently dropped to encourage more sponsors.

[5] Some academies were new establishments with no predecessor school, some were previously private schools.

[6] According to National Audit Office (2012), "well performing" is based on the last three years' test/exam results; prior OFSTED inspections, particularly OFSTED judgements on leadership and the capacity to improve; financial management, and any other evidence deemed significant.

eight months elapsing on average between an initial expression of interest and the actual re-opening of a school as an academy. The approval rate for applications to become a primary converter academy is 90 per cent, which should allay fears that schools are "cherry picked" to become academies.[7] It is not uncommon for conversions to take place mid-school year, although many conversions occur over the summer school break.

The DfE continues to identify under-performing schools, match them with sponsors and impose academy status. Weak schools that apply to become converter academies can have their application withdrawn and face a sponsor-led academy takeover thrust upon them.

Table 2.1 shows the number of each type of state primary school open at the start of every school year since 2008. Five years after their introduction, converter academies account for 11.1 per cent of the primary school sector. 5.4 per cent of primary schools are now sponsored academies. Table 2.2 depicts the number of primary conversions during each school year by predecessor school type. Around 120 primary schools converted during the 2010/11 school year. Since then between 350 and 450 conversions have taken place each school year. Although a slightly disproportionate number of early converters were community schools, it appears that the overall predecessor school type distribution corresponds to the prevalence of each type in the pre-academy period.

In England, pupils start primary school at the age of four or five and complete seven school years at primary level before joining a secondary school at age ten or eleven. Primary school is split into three stages: reception which lasts a single school year; key stage 1 (KS1) which covers the second and third years of primary school (known as

[7] This statistic is calculated from the author's own analysis of the DfE's *Open Academies and Applications Dec '15* dataset, and refers to the number of all applications received by the end of December 2015 to be approved.

year 1 and 2), and key stage 2 (KS2) which encompasses the final four years of primary schooling.

At the end of both key stages, schools assess the attainment of their pupils in English, maths and science. Schools have good reasons to encourage their pupils to perform well in the KS2 tests. KS2 assessment performance is an integral component of school league tables and the broader school accountability system. KS2 performance can also affect pupils' secondary school experience if their secondary school tracks students by ability since KS2 performance may be used by secondary schools to gauge the ability of pupils joining from primary schools.

2.3 Literature review

2.3.1 US evidence: charter schools

Other nations have introduced new, more autonomous school types to improve attainment. A well-established literature exists on charter schools, which were introduced to the US in 1992. Like academies, charter schools are highly autonomous, fee-free and non-selective. Unlike academies, charter schools tend to be new establishments with no predecessor state school.

The causal effect of charter school attendance is often identified using charter admission lotteries to instrument the number of years spent in a charter school. Identification depends on the lotteries being fair and, by implication, lottery winners and losers not being systematically different. Angrist *et al.* (2010) find that lottery winners test scores are 0.35σ and 0.12σ higher per year of charter attendance in maths and English Language Arts (ELA) tests, respectively. σ denotes the standard deviation of the test score distribution for a given subject, grade and year. Based on different samples, Abdulkadiroğlu *et al.* (2011) and Dobbie and Fryer Jr (2011) report quantitatively similar effects for maths test scores, but find ELA test score effects in limited circumstances only.

There are good reasons to interpret these results cautiously. Admission lotteries are held when schools are oversubscribed which is a consequence of good performance. Therefore, the studies pre-condition on school quality. These studies also condition on schools retaining lottery records which might be associated with the efficiency or competence of the school (Dobbie and Fryer, 2011a). The interaction of these factors means that the samples of the studies above are small. The sample of eight schools in Abdulkadiroğlu *et al.* (2011) is the largest of the three. Hoxby and Murarka (2009) use a larger sample of 42 charter schools located across New York City. They report a much smaller per year of charter school attendance effect of 0.09σ on maths test scores and a statistically insignificant reading test score effect.

Other lottery based (Gleason *et al.*, 2010) and matching evidence (CREDO, 2013) suggests some charter schools are ineffective. Urban charter schools seem to be effective whereas non-urban charters appear to be ineffective or harmful. Angrist *et al.* (2013) argue that student demographic differences explain a small portion of the urban/non-urban distinction; whereas variation in the policies and practices of urban and non-urban charter schools have more explanatory power. The *No Excuses* philosophy, incorporating strict discipline, academic rigour and high expectations, may be driving the urban charter school effect (Angrist *et al.*, 2011). 45 per cent of the variation in charter school effectiveness is associated with policies aligned with the *No Excuses* model (Dobbie and Fryer, 2011b).

Evidence on the medium-term effect of charter school attendance is similarly mixed. Teen pregnancy and incarceration are less likely among charter attendees (Dobbie and Fryer, 2015), yet charter attendance does not appear to affect the likelihood of high school graduation or college enrollment (Angrist *et al.*, 2016).

State to charter school conversions, which are more comparable to England’s experience with academy schools, have also been studied. However, charter school takeovers are considerably less common than start-up charter schools. Abdulkadiroğlu *et al.* (2016) focuses on nine charter takeovers of failing New Orleans, LA public schools, and another in Boston, MA. To accommodate selection into and out of takeover schools, the authors use enrolment in the schools pre-takeover to instrument enrolment post-takeover. Takeovers are shown to have significant positive effects on maths and reading test scores. A similar study by Fryer Jr (2014) imposes the freedom and practices associated with effective charter schools on eight randomly selected failing elementary schools in Houston, TX. After two years of exposure, maths test scores in the treated schools improve by 0.15σ on average relative to their closest matched school from the control group.

Another difference between academies and charter schools is that charter schools are not generally part of a centralised admissions system. They instead require parents to make a separate application to them whereas, academy admissions are handled through the same centralised process as applications to maintained schools. Abdulkadiroğlu *et al.* (2015) investigates charter school effectiveness in the Denver, CO school district which has a rare unified, centralised admission system incorporating charter schools. The authors find positive attainment effects from charter school attendance similar to Abdulkadiroğlu *et al.* (2011).

2.3.2 English literature: grant-maintained and academy schools

The academies programme is not the first initiative to increase the autonomy of England's schools. Between 1988 and 1997, if maintained schools won a majority vote of current parents, they could partially opt out of LEA control by becoming a grant-maintained (GM) school.[8] One-third of secondary schools held such a vote. Clark (2009) uses a fuzzy regression discontinuity design to estimate the GM conversion effect. GM conversion meant greater autonomy, including control over staffing and admission policies, and more generous capital and current expenditure funding (according to estimates). Clark reports that the percentage of pupils in converters passing five GCSEs or more increased by 4 to 6 percentage points (from a base of 60 per cent). The prior attainment of the entry year intake increased for converters, and they experienced higher teacher turnover and a net rise in teacher numbers. No evidence is found that schools neighbouring a GM converter were affected by their neighbour's conversion.

The majority of research into academies is based on the first generation of sponsored academies. An early, government commissioned, evaluation of the academies programme reported that improvements in the GCSE attainment of the first 27 academies exceeded the national average improvement (PriceWaterhouseCoopers, 2008). However, this finding may merely reflect mean reversion. These academies replaced some of England's most poorly performing schools and had greater scope for improvement than the average school. A more rigorous early analysis is provided by Machin and Wilson (2009) who compare each academy to a closest matched non-academy twin and also to other secondary schools in the same local authority. They

[8] GM schools are the predecessors to today's foundation schools.

report positive academy effects on GCSE performance. However, their estimates are not statistically significant at standard levels.

A series of papers estimate difference-in-differences models using a treatment group of approximately 100 sponsored academies which opened between 2001/02 and 2008/09. The control group consists of a further 100 sponsored academies which re-opened in later school years. Using school-level data, Machin and Vernoit (2011) find that average GCSE attainment and prior (KS2) attainment of the entry-year intake both increase following an academy takeover. However, these effects take time to materialise. The authors also present evidence that the KS2 attainment of nearby schools' entry-year intake decreases, although schools neighbouring the best performing sponsored academies also experience an improvement in their average GCSE performance.

The estimated GCSE attainment effect for sponsored academies could be biased from pupils non-randomly switching into or away from academies in response to sponsored academy takeovers. Indeed, the increase in the prior attainment of the entry-year intake suggests this is a valid concern. Using the same sample of schools, but with pupil level data, Eyles and Machin (2018) account for this potential source of bias by instrumenting attendance at an academy with attendance at the academy's predecessor school before the takeover.[9] The authors report that the GCSE point score of pupils who attend an academy for one school year is 0.04σ higher on average; while for those attending an academy for four school years the average effect is 0.24σ.[10] Only seven per cent of pupils in the sample attend university. However, each school year spent in a sponsored academy increases the likelihood of attendance by 0.7 percentage points.

[9] For similar analysis see (Eyles *et al.*, 2016a).

[10] Eyles and Machin (2018) suggest that the improvement in GCSE performance is only experienced by sponsored academies which takeover former community schools.

The authors provide a brief insight into the potential mechanisms behind these attainment effects. Sponsored academies are much more likely to undergo a leadership change than control group schools. Academies also add extra pupils and teachers, including unqualified teachers (one of their new freedoms). The teacher-pupil ratio slightly increases.

There are also improvements in the average prior KS2 attainment of the entry-year intake for newer secondary sponsored academies (takeovers after the Academies Act 2010); the magnitude of the effect is comparable to that for older academies (Eyles *et al.*, 2015). The same paper finds no evidence of a change in the prior attainment of the entry-year intake of secondary converter academies.

A National Audit Office (2010) evaluation suggests that sponsored academies improve other student outcomes. Sponsored academies are more effective at reducing the percentage of school days lost to absence than comparable maintained schools. Additionally, they are more effective than similar non-academies at reducing the number of their pupils not in education employment or training (NEET) after age 16.

A fundamental challenge with evaluating sponsored academies is disentangling the effects of increased school autonomy, changes in school leadership and heavily refurbished or newly built school buildings. It is not clear how these factors interact to produce a "sponsored academy effect". By comparison converter academies generally experience an increase in the first of these factors, but no change in the latter two.

To date, there are two evaluations of converter primary academies. Worth (2015) uses propensity score matching to compare KS2 performance in the 2014/15 school year between primary converter academies and matched non-academies. The analysis does not uncover any statistically significant academy status effect on KS2 performance for the average pupil or several sub-groups of pupils. Since this study is cross-sectional, the author is unable to control for any time-invariant differences between academies and non-academies.

Eyles *et al.* (2017) applies the methodology of Eyles and Machin (2018) to an analysis of primary converters. The authors find no effect of voluntary academy conversion on KS2 attainment. Primary schools that converted between 2010/11 and 2014/15 form the treatment group, while schools that converted in 2015/16 and 2016/17 are the control group. However, the approval criteria for academy conversion applications weakened significantly in April 2011. In the methodology section (Table 2.4), I show that in the pre-treatment period of the present study, primary schools converting between 2010 and 2012 had a better attainment record and educated more advantaged pupils than primary schools that became converter academies after 2012. If these observable differences are accompanied by unobservable differences between primary schools established either side of the approval criteria change, then enrolment in a predecessor school is not a validly excluded instrument for enrolment in a converter academy. To address this, Eyles *et al.* stratify their sample according to schools' most recent OFSTED rating. This ensures there are no differences in the means of baseline characteristics between their control and treatment schools.

An aspect of the academy programme yet to be thoroughly analysed is academy chains. Half of all academies are a constituent of one of nearly 300 chains: academies linked together through a common sponsor and/or as a single legal entity (typically, a multi-academy trust). The development of chains has been encouraged to mitigate the risks associated with increased autonomy and to facilitate the sharing of best practice. Focusing on long-established chains, (Hutchings *et al.*, 2014) offers a descriptive analysis of the effectiveness of chains in the secondary sector.[11] The report reveals persistent variation between and within chains in their ability to improve disadvantaged pupils' attainment. Other evidence indicates that sponsored academies in chains perform marginally better than standalone sponsored academies.

2.4 Data

This chapter is based on extracts from the DfE's National Pupil Database (NPD), a collection of linked administrative datasets providing detailed information on England's state schools and their pupils. The School Census links pupils to the school they attend at a given point in time. It contains rich demographic information such as gender, ethnicity, first language, as well as month and year of birth. Proxy variables including FSM eligibility history capture socioeconomic circumstances. School Census records can be directly matched to pupils' KS1 and KS2 attainment records. I also use data from the School Level Database (SLD) to facilitate between school comparisons of aggregate pupil demographics and attainment.

State primary schools are statutorily required to assess their pupils' attainment using national curriculum (NC) assessments. This includes externally set and marked tests and externally moderated teacher-based assessments. Primary schools must register their pupils for these assessments at the end of key stages 1 and 2 (years (i.e. grades) 2 and 6).

The KS2 assessments feature mathematics and reading tests, as well as a combined spelling, punctuation and grammar test (since 2012/13). Separately, year six pupils undergo teacher assessments in English, mathematics and science. Since 2005, pupils

[11] See also Hutchings *et al.* (2015)

receive a teacher assessment in reading, writing, speaking and listening, mathematics and science at the end of KS1.[12]

Primary NC assessments were graded using a five-point grade scale (levels 1 to 5) until 2012 when, with the intention of challenging high performing pupils, the government introduced level 6. A pupil achieves level 6 at KS2 in a subject if they pass an additional test. Consequentially, the grading and difficulty of the level 1 to 5 KS2 tests did not systematically change in 2012. Pupils are expected to be working at level 2 at the end of KS1. Pupils should make two levels worth of progress throughout KS2. Therefore, year six pupils are expected to attain level 4.

The analysis separately assesses the effects of academy conversion on pupil attainment in reading and maths, as academies may on average place greater emphasis on either subject than non-academies following the national curriculum. KS2 attainment is measured using test marks standardised to zero mean, unit standard deviation. However, it is important to note that pupils who are deemed by their schools to be working below the level assessed by the KS2 tests gain exemption from the tests. As such, pupils at the bottom of the attainment distribution are excluded from the analysis. Including these pupils necessitates using NC level point score as the attainment outcome which is a far coarser variable.[13] This chapter uses this outcome in a robustness exercise. As teacher assessment is the exclusive measure of KS1 attainment, the outcome variable for KS1 analysis is the NC level point score.

[12] Pupils previously also sat KS2 writing and science tests, discontinued in 2012 and 2009 respectively. Before 2005, KS1 attainment was assessed using formal testing.

[13] The NC level point score is a simple numerical transformation of the NC level. For example, level 1 is coded as 9 points, level 2 as 15 points. The NC level point score does not convey any more detail than the NC level.

This chapter uses a data extract covering school years 2007/08 to 2014/15. 2014/15 is the last school year before NC assessments undergo significant reform. The analysis uses data on every year 2 and year 6 pupil in each of these school years to determine how academy status may affect pupil attainment. Separately, the analysis uses data on every reception pupil (the entry-year) to explore whether academy status affects the composition of the entry-year intake. Primary schools that do not cover reception and key stages 1 and 2 in their entirety or schools that cater to special educational or behavioural needs are excluded from the analysis. [14]

2.5 Methodology

The causal effects of a primary school opting to become a converter academy are estimated using difference-in-differences (DiD) models. The baseline estimating equation is

$$y_{ist} = \alpha_s + \alpha_t + \beta_1 Academy_{st} + \gamma' x_{ist} + \varepsilon_{ist} \tag{2.1}$$

where i, s and t are pupil, school and school year (i.e. cohort) identifiers respectively. y_{ist} refers generically to an attainment measure. α_s is a school fixed effect and α_t is a school year (time) effect. Binary variable $Academy_{st}$ equals 1 if school s is a primary converter academy in school year t and 0 otherwise. Conversion is an “absorbing” state since no academies revert to maintained school status. The parameter of interest is β_1 representing the estimated average causal effect of treatment on the treated (ATT). This is the estimated average change in attainment in converter academies caused by conversion to academy status. x_{ist} is a vector of time-varying pupil-level control variables. Under the parallel trends assumption, the error term, ε_{ist}, is orthogonal to

[14] In other words, lower and middles schools are excluded from the analysis.

$Academy_{st}$. I assume this term has a school/school year specific component that is likely to exhibit serial correlation over time. Therefore, I estimate robust standard errors clustered at the school level, as advocated by Bertrand *et al.* (2004). There are approximately 1,300 clusters which exceeds the standard minimum number of clusters required to estimate robust clustered standard errors (Cameron and Miller, 2015).

When outcome y_{ist} is a measure of KS2 attainment, a value-added model can be estimated using prior KS1 attainment. This model is motivated by the lack of observed historical school and parental inputs. These important unobserved inputs are proxied using prior attainment. The model incorporates prior KS1 attainment in vector x_{ist}, which assumes the effects of historical inputs experience a common rate of geometric decay. The alternative case where y_{ist} is equal to the difference of current and prior attainment assumes prior inputs are as relevant as current inputs.

The value-added model does not account for contemporaneous changes in parental inputs. Parents may interpret a school's decision to become a converter academy as a positive or negative signal of the school's quality and may adjust their parental inputs accordingly. Therefore, the estimated treatment effects are net of the average parental response to their child's school becoming an academy. Value-added models are thoroughly critiqued in Todd and Wolpin (2003), which also discusses the unavoidable restrictions that such models place on the underlying education production function.

The analysis extends equation (2.1) in several ways to accommodate different forms of treatment effect heterogeneity. Equation (2.1) imposes a constant average treatment effect for every school year following academy conversion. It is unlikely that academies fully realise and exploit the implications of their enhanced independence straight after conversion. Instead, there may be an adjustment period during which academies gradually implement changes that would not have been possible as a maintained school. It is appropriate to adopt a specification that allows the treatment effect to vary according to the length of time elapsed since conversion occurred. A more flexible variant of equation (2.1) is

$$y_{ist} = \alpha_s + \alpha_t + \sum_{\tau=-4}^{\tau=2} \beta_\tau Academy\,Yr\,\tau_{ts} + \gamma' x_{ist} + \varepsilon_{ist} \qquad (2.2)$$

where $AcademyYrX_{ts}$ equals one if the difference between school year t and the school year that school s becomes an academy is X school years, and zero otherwise. This is sometimes referred to as the leads and lags DiD estimator and attributed to Autor (2003). If the control and treated groups have differential trends in the absence of treatment, then the pre-treatment beta estimates $(\hat{\beta}_{-4}, \dots, \hat{\beta}_{-1})$ will be significantly different from zero. Estimates that are not significantly different from zero lend support in favour of the identifying assumption.

26 per cent of primary schools participated in a boycott of KS2 assessment tests in May 2010. Since participation in the boycott was non-random and widespread, the 2009/10 school year is dropped from the panel for all KS2 attainment analysis. This means the pre-treatment period spans four schools-years (two either side of the dropped year). As such, I correct the pre-treatment indicators in equation (2.2) such that, for example, if a school becomes an academy in 2012/13, then 2008/09 is coded as the third school year before that school’s conversion, and not the fourth school year prior.

Also, equation (2.1) does not allow the treatment effect to vary between academies with different predecessor school types, despite academies experiencing varying degrees of autonomy before conversion. As schools experience differential increases in autonomy following conversion to academy status, there is an element of treatment intensity which could be captured. I interact a binary variable equal to 1 if an academy was previously a community or voluntary-controlled school (CVC_s) and 0 otherwise, with $Academy_{st}$. In equation (2.3), $\beta_1 + \beta_2$ is the ATT for academies which were previously community or voluntary-controlled schools, whereas β_1 is the ATT for academies whose predecessor school was another maintained school type.

$$y_{ist} = \alpha_s + \alpha_t + \beta_1 Academy_{st} + \beta_2(Academy_{st} \times CVC_s) + \gamma' x_{ist} + \varepsilon_{ist} \quad (2.3)$$

Specific sub-groups of the pupil population may be affected differently by academy conversion than the average pupil. The autonomy accompanying academy status may allow academies to redirect their attention and resources towards or away from certain pupil groups. An important sub-group is pupils from disadvantaged backgrounds. I use FSM eligibility to indicate disadvantage. I further estimate equation (2.4).

$$y_{ist} = \alpha_s + \alpha_t + \beta_1 Academy_{st} + \beta_2(Academy_{st} \times FSM_i) + \gamma' x_{ist} + \varepsilon_{ist} \quad (2.4)$$

FSM eligibility is recorded in vector x_{ist}. In equation (2.4), β_1 is the ATT for pupils who are ineligible for FSM, while $\beta_1 + \beta_2$ is the ATT for FSM pupils. Equations (2.1) to (2.4) are estimated using pupil level data.

For the entry-year intake analysis, the baseline estimating equation is

$$y_{st} = \alpha_s + \alpha_t + \beta_1 Academy_{st} + \varepsilon_{st} \quad (2.5)$$

where y_{st} refers to the entry-year cohort average of a certain pupil characteristic for school s in school year t. The interpretation of the equation’s remaining components is the same as in the preceding equations. β_1 is the ATT estimate which is the estimated average change in the cohort average of a certain attribute of the entry-year intake experienced by schools when they become academies.

The β estimates in equations (2.1) to (2.5) provide unbiased treatment effect estimates if the parallel trend assumption holds conditional on the control variable vector x_{ist}. The school fixed effect controls for differences in time invariant characteristics between treatment and control schools. It remains a possibility that schools become academies based on unobserved trends. I depend on the parallel trends assumption to dismiss this remaining identification threat.

The treatment and control groups should be as similar as possible in observed and unobserved dimensions; this maximises the likelihood that outcomes for the groups share a common time trend in the absence of treatment. While the application procedure and criteria for approval for academy conversion changed during the 2010/11 school year, it has not significantly changed since. As such, schools that later become academies should be similar to already opened academies.

The treatment group is defined as all schools that become converter academies in the school years 2012/13 to 2014/15. The control group is schools that become converter academies during the 2015/16 school year. The treatment group includes schools that experience one to three school years of academy status. The implication of this research design for the primary outcome of interest, value-added at KS2, is that I observe cohorts who spend between one and three years of KS2 (which spans four years) at an academy; the treatment schools experience academy status for 21 months on average. The minimum observed pre-treatment period is four school years.

Panel A of Table 2.3 compares several measures of attainment and pupil demographics, for the year six cohort averaged at school-level, for the last pre-treatment school year, between the control and treatment groups. Column (3) tests the equality of means between the two groups. The means are not significantly different at conventional levels of significance, providing good evidence that the groups are alike regarding observable factors the school year before the first treatment schools become converter academies.

Panel B of Table 2.3 shows the change in the same school level attainment and demographic measures of the year six cohorts between 2007/08 and 2011/12 for the control and treatment groups. Column (3) tests whether the difference in the mean change is equal between the two groups. There are no statistical differences between the groups at typical significance levels. This suggests that the overall trend in these measures in the pre-treatment period do not vary between the groups.

Table 2.4 compares the means of the same variables averaged over the pre-treatment period for a more extensive selection of schools. Column (1) shows means for schools that become converter academies in 2010/11 and 2011/12. Column (2) and (3) contains means for the schools that are considered in this analysis. Comparison of Column (1) against columns (2) and (3) suggests that year six pupils in the first schools to become converter academies perform significantly better in the pre-treatment period than those attending the converter academies included in this paper's main analysis sample. The average KS2 reading standardised test mark in the earliest converters in the pre-treatment period is 0.17 standard deviations compared to 0.05 standard deviations for the converter academies in the treatment and control groups. The earliest converter academies also educate more advantaged pupils (based on eligibility for FSM) than later converter academies. Additionally, unreported results show that the trends in these variables differ in the pre-treatment period between the earliest converters academies and the academies in this paper's sample. This table demonstrates that the first two waves of primary converter academies differ to more recent primary converter academies in observable dimensions in the pre-treatment period. As such, there are reasonable grounds to suspect that they may also differ in unobserved attributes.

Consequentially they are not appropriate to include in a research design which exploits time variation in conversion to converter academy status.

2.6 Results

2.6.1 KS2 attainment

Table 2.5 contains estimates from difference-in-differences (DiD) models with a single post-treatment effect. In columns (1) to (3), the outcome is KS2 maths standardised test mark. KS2 reading standardised test mark is the outcome variable for columns (4) to (6). Columns (1) and (4) feature estimates from a DiD model without any control variables. I add control variables in columns (2) and (5), and then add prior attainment in each subject in columns (3) and (6) to create a value-added model. The converter academy coefficient estimate (found in the first row) corresponds to the estimated effect of academy conversion. The estimates are relatively consistent as control variables and then KS1 attainment are added, ranging between 0.017 to 0.007 standard deviations. None of the estimates are statistically different from zero at the ten per cent significance level. This contrasts with the control variable coefficients which are uniformly estimated with high precision and are statistically different from zero. These estimates do not provide evidence of a converter academy status effect on KS2 attainment. This finding is not sensitive to the measure of KS2 attainment. Appendix Table 2.1 shows there is no academy status effect when the dependent variable is the point score corresponding to the National Curriculum (NC) level achieved by the pupil (levels range from 1 to 6), or a binary variable indicating if the expected NC level (level 4) is achieved.

Appendix Table 2.2 shows that this finding is insensitive to an alternative treatment definition and an alternative model specification. The treatment effect estimate may be subject to bias caused by mismeasurement in the treatment variable. The main mismeasurement threat comes from schools not operating as academies until they enter a full school year as one, despite possibly legally becoming a converter academy midway through the previous school year. In Panel A, I calculate the treatment variable such that schools are coded as exposed to converter academy status only if they have that status at the start of the school year. This does not alter the conclusions that can be drawn from Table 2.5. In Panel B, I add school specific linear time trends to investigate whether the results are being driven by differential trends in KS2 performance between treated and control schools. Estimates of the academy status effect are not sensitive to the inclusion of these trends. In the following tables and figures, I present estimates from the preferred specification (columns (3) and (6)) only; estimates are not sensitive to specification choice.[15]

Figure 2.1 and Figure 2.2 plot estimates from models with pre- and post-treatment effects. I estimate the effect of being in the treatment group in the years leading up to and following treatment. This allows the treatment effect to vary by length of exposure and can also be used to assess the validity of the common trends assumption. There should be no "effect" from being in the treatment group before treatment. If an "effect" is consistently found before treatment, then this raises concerns about the research design. In Figure 2.1 and Figure 2.2, the coefficient estimates for school year 0 correspond to the estimated academy status effect during the conversion year. Coefficient estimates for school years less than 0 correspond to pre-treatment effect estimates. Figure 2.1 plots the estimated treatment effects on KS2 maths standardised test mark, while the effect on KS2 reading standardised test mark is depicted in Figure 2.2. The findings from Figure 2.1 and Figure 2.2 are consistent with those from Table

[15] Full tables are available upon request.

2.5; no statistically significant treatment effect is found for attainment in either subject in any treated school year conditional on the control variables and prior attainment. The F-test statistic corresponding to the null hypothesis that the pre-treatment coefficient estimates are jointly insignificantly different from zero is 1.44 and 1.31 for the maths and reading models respectively. Therefore, there is no evidence of differential trends between the control and treatment groups before treatment. This suggests that the common trends assumption holds.

It is plausible that academy conversion effects on KS2 attainment exist for sub-populations of pupils and schools, despite the seeming lack of an effect for the average school or pupil. Table 2.6 presents estimates from two models which accommodate heterogeneous treatment effects for disadvantaged pupils, and academies which were relatively autonomous before conversion.

Estimates from models allowing the academy conversion effect to vary by FSM eligibility are presented in Panel A. This is the best available indicator of whether the pupil's background is disadvantaged. Columns (1) and (2) suggest that the KS2 maths and reading attainment of FSM ineligible pupils are not affected by academy conversion. However, there is evidence of a small positive academy conversion effect (0.03 standard deviations) on maths and reading attainment for FSM eligible pupils. This effect is statistically different from zero at the five per cent significance level.

Panel B investigates school level heterogeneity; the reported model allows the academy conversion effect to vary between former community and voluntary-controlled schools, which had the least autonomy before becoming an academy, and voluntary-aided and foundation schools which were relatively more autonomous. The academy conversion effect on KS2 maths attainment is insignificantly different from zero regardless of the school's previous structure. However, pupils in former voluntary-controlled and community school academies gain 0.036 standard deviations in KS2 reading on average. This effect is statistically different from zero at the five per cent significance level. There is no effect for pupils from former voluntary-aided or foundation schools.

The estimated academy conversion effect will be biased if enrolment in the converting school is sensitive to the conversion. Parents may interpret the conversion decision as a school quality signal and may alter their child's enrolment accordingly. I estimate DiD models based on school-level data to investigate whether becoming a converter academy influences the composition of the year six cohorts. Table 2.7 shows that there is no academy conversion effect on the observed average characteristics of the year six cohorts. Evidence that enrolment decisions are not sensitive to academy conversion is found in column (1), which reports that there is no academy conversion effect on the percentage of year six pupils who completed KS1 (year two) in the same school.

2.6.2 KS1 attainment

Table 2.8 presents estimates of the effect of academy conversion on KS1 maths attainment (see the first two columns) and KS1 reading attainment (see the last two columns). Since KS1 is the first formal assessment of pupils, there is no opportunity to implement a value-added model. This increases the scope for bias from unobserved confounders relative to the KS2 value-added models. Additionally, KS1 attainment is recorded using teacher assessments which are inherently more subjective. However, there is still a good cause to investigate KS1 outcomes. The KS2 value-added models show a strong relationship between attainment at KS1 and KS2. Moreover, the relationship between attainment at KS1 and KS4 (age 16) is far from trivial. The raw correlation between KS1 maths NC curriculum level and GCSE maths point score is 0.624; the correlation between age seven and 16 English attainment is 0.597[16].

The estimates of the converter academy coefficient are stable following the inclusion of control variables but are insignificantly different from zero both statistically and economically; whereas every control variable coefficient is precisely estimated at the one per cent level. No evidence is found of an academy conversion effect on KS1 attainment. This finding is consistent with an unreported dynamic DiD model, in which pre- and post-treatment effect estimates are insignificantly different from zero.

Again, it is possible that the zero average treatment effect on KS1 attainment is masking non-zero treatment effects for school and pupil sub-populations. In an unreported exercise, I investigate heterogeneous treatment effects at the pupil level (by FSM eligibility) and the school level (by predecessor school type). Similar to the KS2 analysis, I find evidence of a small positive converter academy status effect on reading and maths attainment for FSM eligible pupils, but no effect for FSM ineligible pupils. I also find evidence of a slight, but statistically significant, positive effect on KS1 reading and maths attainment in schools which had the least autonomy before conversion.

2.6.3 Entry-year intake

Finally, I explore whether the composition of schools' entry-year intake changes following academy conversion. Table 2.9 reports the findings from a rudimentary DiD model estimated on school-level data where the outcome variables are the percentage of the entry-year cohort: eligible for FSM; with SEN; whose first language is English, and who are white. The academy coefficients in columns (1) to (3) are insignificantly

[16] Author's own calculations.

different from zero suggesting the composition of the entry-year intake for schools is not affected by becoming an academy in three of the four characteristics investigated. However, column (4)'s estimate suggests that academies experience a 0.6 percentage point decline in the proportion of their entry-year intake that is white. 81 per cent of entry-year pupils are white in the sample. It is unusual that the composition of the new intake would change in this dimension only. Given that the size of the effect is modest at best, I opt to place little emphasis on this finding.

2.7 Conclusion

This chapter attempts to quantify the causal effect of the voluntary conversion of English state primary schools into converter academies on pupil attainment, and the composition of the entry-year intake. To this end, the staggered nature of academy conversions across schools and the availability of a rich administrative dataset are exploited in a battery of difference-in-differences models.

Estimates from these models consistently find no evidence of an academy conversion effect on KS2 maths and reading test point scores for the average pupil. However, heterogeneous effects models do find evidence of a small positive, but statistically significant, KS2 attainment effect for FSM eligible pupils. There is also evidence of a small positive effect in KS2 reading attainment for schools that had the least autonomy before becoming a primary converter academy. KS1 teacher assessments and the composition of the entry-year intake are seemingly unaffected by academy conversion.

Although these results are consistent with prior research into primary converter academies, studies of secondary sponsored academies have found academy status effects on attainment. Numerous reasons may explain this discrepancy. Firstly, converter academy pupils tend to be more advantaged and academically meritorious than their sponsored academy peers. If the marginal effect of school inputs is diminishing, and academy status improves school inputs comparably in converter and sponsored academies, then academy status will be more effective in sponsored academies where pupils' attainment is at a lower base level.

However, academy status means different things for sponsored and converter academies. First-generation sponsored academies often enjoyed new or extensively refurbished facilities, which is likely to affect pupil attainment positively or at least not negatively. Additionally, these academies were highly susceptible to leadership changes following conversion (Eyles and Machin, 2018). Converter academies are not more likely to undergo leadership changes following their conversions (Eyles *et al.*, 2017). Leadership changes may partially explain the difference in the effectiveness of converter and sponsored academy conversions. Suppose underperforming schools are unattractive to effective head teachers. If sponsored academy status increases the attractiveness of an underperforming school to effective head teachers, then sponsored academies may improve pupil attainment through attracting a higher calibre of head teacher. Converter academies might already be attractive to quality school leaders due to their record of good performance. These schools may not attract better leaders following conversions, and, therefore, might not experience attainment improvements.[17]

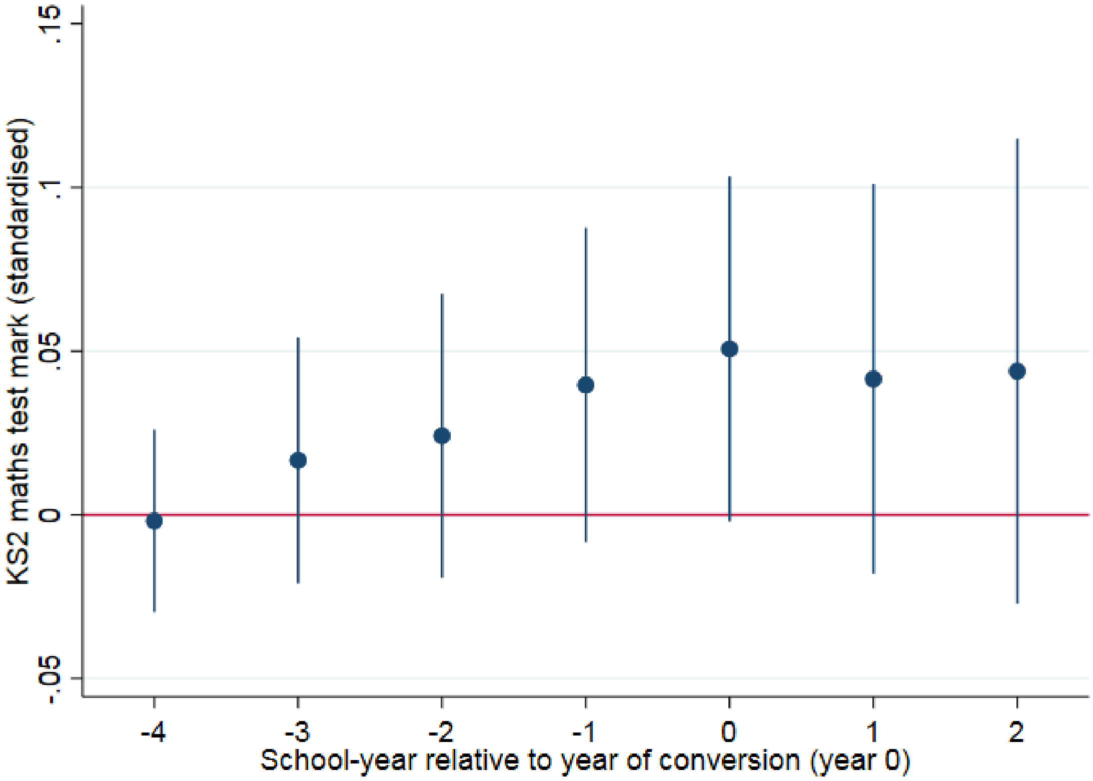
Differences in the stages of schooling may explain the difference in estimated academy status effects. Primary schools are usually smaller than secondary schools, implement different teaching methods, and have different educational goals. The freedom of academies to set their own curriculum may be more consequential for attainment in

[17] If this hypothesis is true, then the effectiveness of sponsored academy status should diminish as the sponsored academy sector expands.

secondary schools since secondary pupils are formally assessed in a broader range of subjects (partially determined by the school); whereas, primary school pupils are predominately assessed in numeracy and literacy. Secondly, if the financial benefit from becoming an academy results in increased availability of useful school resources, then academy status may be more effective at secondary level, as these schools face greater per-pupil costs than primary schools.

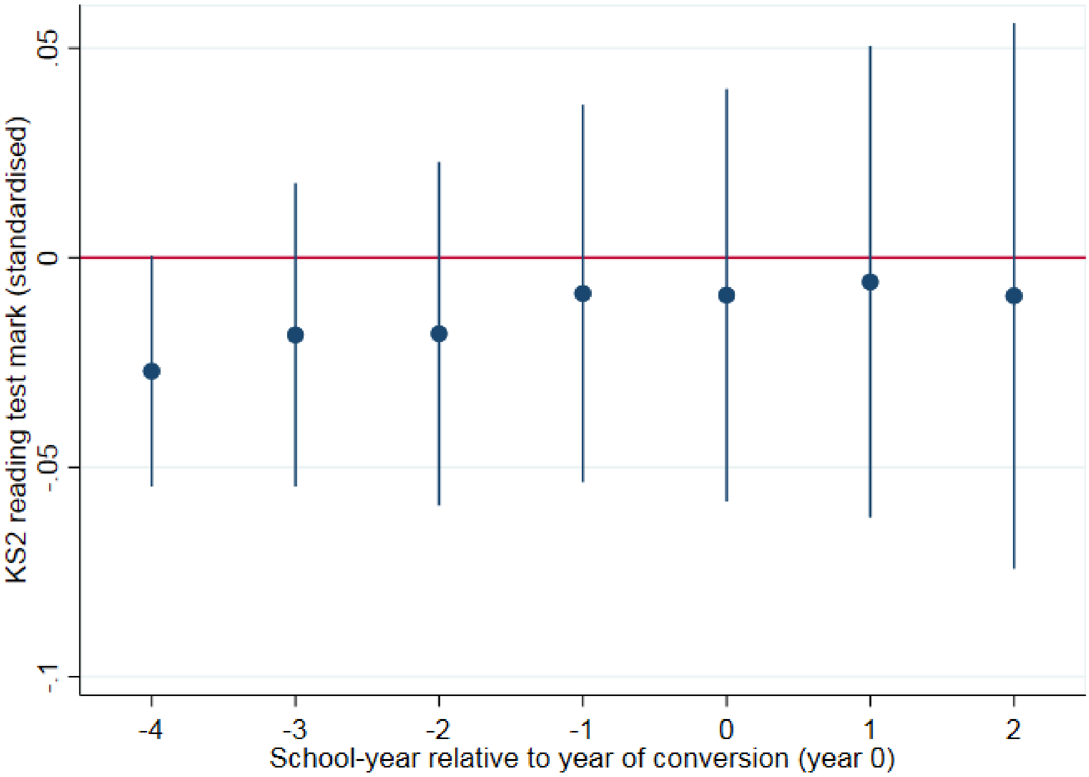
Irrespective of the mechanisms driving the differences between the effectiveness of sponsored and converter academy status, the lack of evidence of an improvement in the attainment of primary converter academies suggests that increasing school autonomy is not a panacea in and of itself. This is an important finding given the considerable cost of the academies programme.

Figure 2.1: Pre- and post-treatment effect estimates for KS2 maths test mark



Notes: filled circle denotes coefficient estimate from leads and lags difference in differences model. Vertical bars denote 95% confidence interval.

Figure 2.2: Pre- and post-treatment effect estimates for KS2 reading test mark



Notes: filled circle denotes coefficient estimate from leads and lags difference in differences model. Vertical bars denote 95% confidence interval.

Table 2.1: The composition of the English state primary school sector at the start of the school year

	2008	2009	2010	2011	2012	2013	2014	2015
Converter academy	0	0	6	265	647	1,069	1,462	1,859
Sponsored academy	0	0	0	5	115	391	685	898
Free school	0	0	0	2	34	37	92	117
Community school	9,893	9,803	9,727	9,491	9,111	8,624	8,166	7,842
Foundation school	911	913	911	863	819	768	734	698
Voluntary aided school	3,747	3,738	3,730	3,684	3,606	3,479	3,326	3,148
Voluntary controlled school	2,465	2,459	2,455	2,427	2,384	2,313	2,234	2,155
Grand Total	17,016	16,913	16,829	16,737	16,716	16,681	16,699	16,717

Notes: each column shows the number of schools of each type open on September 1st of that year. *Source:* author’s analysis of EduBase data.

Table 2.2: Primary converter academy schools by school year of conversion to academy status and predecessor school type

	2010/11	2011/12	2012/13	2013/14	2014/15	2015/16	Total
Community school	77	226	252	232	166	203	1,156
Voluntary controlled school	7	41	43	47	50	58	246
Voluntary aided school	5	82	110	125	128	93	543
Foundation school	31	55	42	18	21	20	187
Multiple or no predecessor school	2	0	0	0	0	2	4
Total	122	404	447	422	365	376	2,136

Notes: school-year is defined as 1st August to 31st July the following calendar year. *Source:* author’s analysis of EduBase data and DfE’s ‘Open Academies’ monthly data release.

Table 2.3: Tests for mean equality between treatment and control groups for attainment and pupil control variables

	(A) School averages in 2011/12			(B) Change in school averages between 2011/12 and 2007/08		
	Control	Treatment	Difference (SE)	Control	Treatment	Difference (SE)
KS2 maths test mark	0.033	0.065	-0.032	-0.019	0.018	-0.036
			(0.024)			(0.027)
KS2 reading test mark	0.078	0.059	0.019	0.012	0.003	0.009
			(0.026)			(0.028)
KS1 maths TA points	15.840	15.853	-0.013	-0.131	-0.159	0.027
			(0.089)			(0.096)
KS1 maths TA points	15.681	15.686	-0.006	0.188	0.189	-0.001
			(0.106)			(0.105)
% female	0.487	0.497	-0.010	0.004	0.004	-0.001
			(0.008)			(0.010)
% English is first language	0.891	0.897	-0.006	-0.023	-0.024	0.001
			(0.013)			(0.005)
% White ethnicity	0.848	0.861	-0.013	-0.027	-0.019	-0.008
			(0.015)			(0.005)
% FSM eligible	0.159	0.163	-0.004	0.027	0.031	-0.004
			(0.010)			(0.007)
% with SEN	0.242	0.240	0.001	0.008	0.008	0.000
			(0.009)			(0.010)
Cohort size	34.643	36.941	-2.298	-3.398	-2.703	-0.694
			(1.474)			(0.586)
Observations	269	1,062		269	1,062	

Notes: variables are school-level averages for the year six cohorts. ***, ** and * denote statistical significance at the 1 per cent, 5 per cent and 10 per cent levels respectively.

Table 2.4: Mean attainment and pupil control variables averaged over pre-treatment period by school type

	(1) Converter academies '11 and '12 openers	(2) Converter academies '13, '14 & '15 openers: treat. group	(3) Converter academies '16 openers: control group	(4) Sponsored academies	(5) Community schools	(6) Voluntary-controlled schools	(7) Voluntary-aided schools	(8) Foundation schools	(9) All schools
KS2 maths test mark	0.191	0.049	0.027	-0.341	-0.012	0.076	0.111	0.016	0.003
KS2 reading test mark	0.173	0.055	0.054	-0.349	-0.026	0.126	0.148	0.029	0.009
KS1 maths TA points	16.237	15.931	15.882	14.990	15.684	16.207	16.134	16.011	15.833
KS1 maths TA points	16.047	15.692	15.648	14.472	15.374	16.050	16.019	15.781	15.577
% female	0.491	0.493	0.488	0.490	0.488	0.491	0.499	0.490	0.491
% English is first language	0.912	0.911	0.902	0.846	0.861	0.961	0.887	0.896	0.883
% White ethnicity	0.870	0.872	0.862	0.801	0.816	0.937	0.839	0.851	0.841
% FSM eligible	0.116	0.147	0.146	0.260	0.186	0.093	0.131	0.124	0.162
% with SEN	0.220	0.243	0.240	0.311	0.257	0.229	0.225	0.228	0.248
Cohort size	46.301	38.325	36.422	39.251	39.522	25.057	30.113	45.922	36.153
Observations	447	1,062	269	910	6,660	1,615	2,707	215	13,885

Notes: variables are school-level averages for the year six cohorts between 2007/08 and 2011/12.

Table 2.5: KS2 maths and reading test mark DiD models with common treatment effect

	(1) KS2 maths mark	(2) KS2 maths mark	(3) KS2 maths mark	(4) KS2 reading mark	(5) KS2 reading mark	(6) KS2 reading mark
Converter academy	0.0141	0.0144	0.0169	0.0087	0.0104	0.0067
	(0.0119)	(0.0119)	(0.0121)	(0.0116)	(0.0116)	(0.0113)
Female		-0.1314***	-0.0882***		0.2061***	0.0769***
		(0.0034)	(0.0029)		(0.0036)	(0.0029)
English is first language		-0.0985***	-0.1677***		0.0484***	-0.0643***
		(0.0111)	(0.0085)		(0.0106)	(0.0081)
White ethnicity		-0.0317***	-0.0248***		-0.0400***	0.0096
		(0.0088)	(0.0064)		(0.0084)	(0.0063)
FSM eligible		-0.3848***	-0.2098***		-0.3620***	-0.1679***
		(0.0075)	(0.0054)		(0.0071)	(0.0050)
Cohort size		-0.0026***	-0.0020***		-0.0018***	-0.0011**
		(0.0007)	(0.0006)		(0.0005)	(0.0005)
KS1 math level			1.0740***			
			(0.0048)			
KS1 reading level						0.8965***
						(0.0042)
Constant	0.0819***	0.5990***	-1.9516***	0.0880***	0.2676***	-1.7700***
	(0.0069)	(0.0365)	(0.0371)	(0.0069)	(0.0315)	(0.0322)
School fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
School year effects	Yes	Yes	Yes	Yes	Yes	Yes
No. of academies	1,062	1,062	1,062	1,062	1,062	1,062
No. of future academies	269	269	269	269	269	269
Observations	326,835	326,835	326,835	324,369	324,369	324,369
Adj. R-Square	0.070	0.104	0.414	0.078	0.116	0.398

Notes: test marks are standardised to mean zero, standard deviation one. Columns (2), (3), (5), and (6) include month of birth effects. Robust standard errors clustered at school level in parentheses. ***, ** and * denote statistical significance at the 1 per cent, 5 per cent and 10 per cent levels respectively.

Table 2.6: KS2 maths and reading test mark DiD models with heterogenous treatment effects

	(1) KS2 maths mark	(2) KS2 reading mark
Panel A: Heterogeneity by FSM eligibility		
Converter academy	0.0121	0.0019
	(0.0119)	(0.0111)
Converter academy x FSM eligible	0.0300^{**}	0.0303^{**}
	(0.0120)	(0.0119)
FSM eligible	-0.2176^{***}	-0.1758^{***}
	(0.0063)	(0.0059)
Adj. R-Square	0.414	0.398
Panel B: Heterogeneity by predecessor school type		
Converter academy	0.0013	-0.0192
	(0.0162)	(0.0150)
Converter academy x community or voluntary-controlled predecessor school	0.0217	0.0361^{**}
	(0.0170)	(0.0156)
Adj. R-Square	0.414	0.398
Control variables	Yes	Yes
Value-added model	Yes	Yes
School fixed effects	Yes	Yes
School year effects	Yes	Yes
No. of academies	1,062	1,062
No. of future academies	269	269
Observations	326,835	324,369

Notes: test marks are standardised to mean zero, standard deviation one. Control variables are the same as in Table 2.5. Robust standard errors clustered at school level in parentheses. ***, ** and * denote statistical significance at the 1 per cent, 5 per cent and 10 per cent levels respectively.

Table 2.7: Year 6 cohort composition DiD models on school-level data

	(1) Same school since KS1	(2) FSM Eligible	(3) SEN	(4) English is first language	(5) White ethnicity	(6) KS1 maths points	(7) KS1 reading points
Converter academy	0.0023	-0.0004	-0.0118**	-0.0014	0.0012	0.0139	0.0290
	(0.0057)	(0.0034)	(0.0047)	(0.0027)	(0.0027)	(0.0443)	(0.0498)
Constant	0.7398***	0.1314***	0.2322***	0.9199***	0.8789***	16.0040***	15.4963***
	(0.0042)	(0.0019)	(0.0029)	(0.0014)	(0.0016)	(0.0280)	(0.0300)
School fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
School year effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
No. of academies	1,062	1,062	1,062	1,062	1,062	1,062	1,062
No. of future academies	269	269	269	269	269	269	269
Observations	9,317	9,317	9317	9,317	9,317	9,317	9,317
Adj. R-Square	0.761	0.768	0.385	0.927	0.935	0.511	0.554

Notes: robust standard errors clustered at school level in parentheses. ***, ** and * denote statistical significance at the 1 per cent, 5 per cent and 10 per cent levels respectively.

Table 2.8: Mid primary school (KS1) maths and reading point score DiD models with common treatment effect

	(1) KS1 maths points	(2) KS1 maths points	(3) KS1 reading points	(4) KS1 reading points
Converter academy	0.0236	0.0238	0.0048	-0.0001
	(0.0385)	(0.0378)	(0.0382)	(0.0375)
Female		-0.0093		1.1031***
		(0.0118)		(0.0125)
English is first language		0.3777***		0.8330***
		(0.0349)		(0.0429)
White ethnicity		-0.2238***		-0.6356***
		(0.0274)		(0.0330)
FSM eligible		-1.4576***		-1.7383***
		(0.0223)		(0.0269)
Cohort size		0.0017		0.0018
		(0.0016)		(0.0017)
Constant	15.9076***	16.9645***	15.7741***	16.2225***
	(0.0236)	(0.0864)	(0.0255)	(0.0917)
School fixed effects	Yes	Yes	Yes	Yes
School year effects	Yes	Yes	Yes	Yes
No. of academies	1,146	1,146	1,146	1,146
No. of future academies	292	292	292	292
Observations	432,669	432,669	432,659	432,659
Adj. R-Square	0.071	0.131	0.077	0.151

Notes: columns (2) and (4) include month of birth dummies. Robust standard errors clustered at school level in parentheses. ***, ** and * denote statistical significance at the 1 per cent, 5 per cent and 10 per cent levels respectively.

Table 2.9: Entry-year intake composition DiD models with school-level data

	(1)	(2)	(3)	(4)
	FSM Eligible	SEN	English is first language	White ethnicity
Converter academy	-0.0056	-0.0031	-0.0017	-0.0059**
	(0.0035)	(0.0027)	(0.0051)	(0.0029)
Constant	0.1008***	0.0453***	0.8971***	0.8702***
	(0.0023)	(0.0016)	(0.0032)	(0.0020)
School fixed effects	Yes	Yes	Yes	Yes
School year effects	Yes	Yes	Yes	Yes
No. of academies	1,137	1,137	847	961
No. of future academies	288	288	214	245
Observations	11,400	11,400	8,488	9,648
Adj. R-Square	0.715	0.333	0.750	0.898

Notes: robust standard errors clustered at school level in parentheses. ***, ** and * denote statistical significance at the 1 per cent, 5 per cent and 10 per cent levels respectively.

Table A2.1: Common treatment effect DiD models with alternative KS2 attainment measures

	(1) KS2 maths NC point score	(2) KS2 reading NC point score	(3) KS2 maths level 4+	(4) KS2 reading level 4+
Converter academy	0.0451	0.0185	0.0025	0.0028
	(0.0547)	(0.0473)	(0.0033)	(0.0023)
Constant	14.2102***	19.1126***	0.4742***	0.7318***
	(0.1713)	(0.1410)	(0.0134)	(0.0074)
Control variables	Yes	Yes	Yes	Yes
Value-added model	Yes	Yes	Yes	Yes
School fixed effects	Yes	Yes	Yes	Yes
School year effects	Yes	Yes	Yes	Yes
No. of academies	1,062	1,062	1,062	1,062
No. of future academies	269	269	269	269
Observations	333,974	331,185	324,653	319,049
Adj. R-Square	0.529	0.433	0.146	0.092

Notes: the outcome variable in columns (1) and (2) is the point score equivalent to the national curriculum (NC) level achieved in the correspondent subject: level 1 is equal to 9 points, level 6 is 39 points (one level corresponds to 6 points). Dependent variable in columns (3) and (4) are equal to one if the pupil achieves NC level 4 in maths/reading and zero otherwise. Control variables are the same as in Table 2.5. Robust standard errors clustered at school level in parentheses. ***, ** and * denote statistical significance at the 1 per cent, 5 per cent and 10 per cent levels respectively.

Table A2.2: Robustness checks

	(1) KS2 maths test mark	(2) KS2 reading test mark
Panel A: Only consider full school-years as an Academy as exposure to academy status		
Converter academy	0.0051	0.0109
	(0.0123)	(0.0113)
Adj. R-Square	0.414	0.398
Panel B: Introduce school-specific trends to the model		
Converter academy	0.0179	0.0018
	(0.0120)	(0.0114)
Adj. R-Square	0.429	0.409
Control variables	Yes	Yes
Value-added model	Yes	Yes
School fixed effects	Yes	Yes
School year effects	Yes	Yes
No. of academies	1,062	1,062
No. of future academies	269	269
Observations	326,835	324,369

Notes: test marks are standardised to mean zero, standard deviation one. Control variables are the same as in Table 2.5. Robust standard errors clustered at school level in parentheses. ***, ** and * denote statistical significance at the 1 per cent, 5 per cent and 10 per cent levels respectively.

Chapter Three:

The Impact of Boycotting Standardised Tests on Subsequent School Outcomes

Standardised tests are a common, yet contentious, feature of many countries' schooling systems. In April 2010, two UK teacher unions called for a boycott of mandatory age eleven standardised tests due to be sat in two weeks' time. One-quarter of the cohort was affected. This chapter uses a difference-in-differences strategy to estimate the effect of preparing for, but ultimately not completing, standardised tests on subsequent measures of attainment. This chapter reports evidence of a statistically significant small adverse effect on age 14 teacher assessed attainment and age 16 secondary school qualification attainment. However, there is substantial heterogeneity in treatment effect estimates between sub-populations of pupils. Potential mechanisms are discussed, particularly the role of target setting.

3.1 Introduction

Standardised tests are a near-ubiquitous feature of schooling systems in developed nations today. The tests, often sat by entire populations of pupils at critical stages of their schooling, are usually an integral component of a school accountability system. In England, for example, aggregated results from SATs tests are the headline measure of primary school performance reported in school league tables.

A vast literature on school accountability systems has emerged finding evidence that such systems have positive effects on pupil performance (Deming *et al.*, 2016; Neal and Schanzenbach, 2010; Rouse *et al.*, 2013). In the UK context, Burgess *et al.* (2013) report evidence that increased school accountability, via the publication of school league tables, has a significant positive causal effect on secondary school qualification attainment. Although, the attainment effects of the English accountability system are heterogeneous, and not beneficial for all pupils (Burgess *et al.*, 2005). While, in the US literature, there are significant concerns about teachers cheating (Jacob and Levitt, 2003) or gaming accountability systems (Figlio, 2006; Figlio and Winicki, 2005).

Overall, however, the test-based accountability literature suggests that there is an external benefit from participation in standardised tests. The information provided by participating pupils allows regulators and parents to identify under-performing schools, while standardised testing simultaneously incentivises schools to maximise their teaching efforts. Both mechanisms benefit individual pupils irrespective of whether they sit the standardised test.

The effect of not participating in standardised tests on subsequent pupil attainment, given that pupils can free-ride from other pupils' participation in such tests is not well understood. That is, it is largely unknown whether there is a private cost or benefit to an individual's participation in a standardised test.

The contribution of this chapter is to quantify the causal effect of participation in standardised tests for individual pupils, while the school accountability system remains active and unchanged. To this end, this chapter exploits as a natural experiment a widespread but hastily arranged, head teacher-led boycott of one standardised test series in England. The boycott prevented one-quarter of the affected cohort from sitting mandatory age eleven SATs tests.

The chapter applies difference-in-difference models to a matched panel of primary schools that did and did not participate in the boycott. I find evidence of a small adverse boycott participation effect on attainment as measured by age 14 teacher assessments and age 16 secondary school qualification achievement. Effects are precisely estimated between -0.014 and -0.019 standard deviations on maths and science attainment. Effects on English attainment are about half this size but are not estimated with sufficient precision to be statistically significant. However, the magnitude and direction of these estimated effects exhibit substantial variation across sub-groups of pupils. This chapter also presents evidence that the boycott caused a change in pupils' subject choices at age 14 and resulted in a slight reduction in absenteeism at secondary school.

Because the boycott was officially called just two weeks before the SATs tests were due to take place, this chapter argues that the treatment effect of participating in the boycott merely is that the pupil does not sit the SATs tests, nor receives any feedback on their test performance. It is argued there are limited to no systematic differences in test preparation between pupils who sit the SATs tests and those that boycott them. Primary survey data which supports this proposition is presented.

The models reveal that pupils enrolled in schools that boycotted SATs received inflated teacher assessments at the time of the boycott, but only in subjects for which tests were boycotted. This may be because boycotting primary schools' performance was evaluated using this data for the first time, as the usual test data is necessarily missing. The chapter will present an argument that, because of their role in determining secondary school qualification grade targets, these inflated teacher assessments drive the adverse boycott participation effects on attainment at ages 14 and 16. The inflated teacher assessments meant that the boycott affected pupils were expected by the government to achieve a higher level of attainment at secondary school than otherwise similar non-boycott affected pupils. If secondary schools wish to maximise their success rate at meeting grade targets, which were then a feature of secondary school league tables, then theoretically teachers should allocate effort away from boycott affected pupils for whom the likelihood of achieving their artificially high grade target is reduced. Other potential mechanisms are evaluated.

3.2 Literature review

School accountability systems – the process of evaluating schools based on pupil performance measures – is increasingly commonplace (Figlio and Loeb, 2011). In test-based systems standardised tests are used to measure pupil performance. With effective accountability, principal-agent problems are overcome, and schools are incentivised to improve pupil performance.

In 'consequential' test-based accountability systems, schools are explicitly sanctioned or rewarded based on their performance in standardised tests (Hanushek and Raymond, 2005). In England, for example, primary schools may be forced to become academies if their SATs test performance is below the 'floor standard', and OFSTED, the statutory school inspectorate, consider test performance when determining the ‘overall effectiveness’ rating of a school. Implicit sanctions and rewards are also an inevitable feature of accountability systems. School test scores (Black, 1999; Gibbons and Machin, 2003), as well as overall school ratings (Figlio and Lucas, 2004), are capitalised in house prices, indicating the presence of a parental response to school performance measures. Charitable donations have also been shown to be affected by school performance measures (Figlio and Kenny, 2009).

The bulk of research into the attainment effects of test-based accountability systems is based on the US’s experience with No Child Left Behind (NCLB), the federal accountability system introduced in 2002. For example, Dee and Jacob (2011) use a long panel to compare state-level achievement since the implementation of NCLB between states that previously operated their own accountability systems and those that did not. They find evidence of modest (up to 0.2 standard deviations) positive effects of test-based accountability on maths achievement.

Evidence of positive attainment effects is not confined to NCLB studies. Hanushek and Raymond (2005), for instance, report positive achievement effects from the introduction of consequential accountability across US states in the 1990s. While, Rouse *et al.* (2013) combines administrative data with survey data, and exploits discontinuities in Florida’s accountability system to show that low school ratings spur on future achievement gains and changes in school policy. Recent evidence reports that the introduction of test-based accountability increased both the likelihood pupils achieve a four-year degree and their earnings by the age of 25 (Deming *et al.*, 2016).

Research based on UK data is comparatively sparse, although Burgess *et al.* (2013) test the effects of discontinuing school league tables in Wales. The paper uses a difference-in-differences strategy in which English schools, whose performance continued to be published publicly, are the control group. The authors find that the reform reduces pupil attainment by 0.08 standard deviations.

The literature is rarely able to exploit variation in pupils' participation in standardised tests while the accountability system is held constant. The present chapter exploits such variation, as does Andersen and Nielsen (2016). Their paper investigates the impact of participating in national standardised tests on future school attainment in Denmark. The tests were computer-based, and compulsory for certain ages. A technical breakdown in the IT system meant that pupils signed up to complete the test in the breakdown window unexpectedly gained exemption. The authors exploit this exogenous shock to test participation and adopt an instrumental variable strategy. They estimate substantial positive benefits of testing for pupils, which are larger for pupils enrolled in schools with low grades. However, unlike SATs tests, the Danish national tests are formative assessments (focused on pupil development) rather than summative (focused on pupil attainment). Furthermore, the Danish tests were low-stakes from the perspective of schools and teachers, and teachers could re-enrol unexpectedly exempt pupils in the tests if they so wished (an endogenous response to an exogenous shock).

Standardised testing has increasingly been subject to criticism from teachers, parents and academics due to the side effects of such testing. One concern is that children are subject to unnecessary and unhealthy "test anxiety" or "exam stress" when standardised testing is used at early stages of schooling (Connor, 2001, 2003). In England, pupils first sit SATs tests at age seven. Undue pressure on pupils can originate from teachers, who will be judged on their pupils' performance, and parents who may overestimate the importance of test performance on future attainment (Putwain *et al.*, 2012). Standardised tests have also been identified as a source of demotivation among teachers, and a contributing factor to the current teacher recruitment and retention difficulties in England (Day and Smethem, 2009).

The chief criticism of standardised testing arguably is that, when used in high-stakes contexts, schools and teachers are incentivised to engage in strategic behaviour which may harm pupils' learning or wellbeing. Tests that have high stakes for schools encourage "teaching to the test" whereby untested knowledge, skills (or even subjects) are disregarded in favour of testable content. Furthermore, schools may focus on test skills rather than the underlying knowledge of the testable content. On this note, research in the US has compared longitudinal gains observed in high-stakes standardised tests, with gains in low-stakes tests. There is evidence that gains in high-stakes tests outweigh gains in low-stakes tests (Klein *et al.*, 2000; Koretz, 2002; Koretz and Barron, 1998). This may be interpreted as evidence of teaching narrowly to the high stakes tests. Evidence from the US and UK has also emerged of schools allocating their effort towards pupils whose test scores are most consequential for school performance measures (Burgess *et al.*, 2005; Neal and Schanzenbach, 2010; Reback, 2008). Surveys of primary and secondary school teachers in the UK document evidence of selective coaching and mentoring of borderline pupils (West and Pennell, 2000; Wiggins and Tymms, 2002).

There are many other strategic responses to standardised testing. The literature reports evidence of schools: changing suspension patterns around the time of tests consistent with boosting test-takers' average scores (Figlio, 2006); changing their meal programs around the time of tests (Figlio and Winicki, 2005); reassigning teachers to different grades based on which grades are tested (Boyd *et al.*, 2008), and misclassifying pupils as having special educational needs (Jacob, 2005).

However, teaching to the test may be beneficial in specific circumstances. Lazear (2006) theoretically demonstrates that well-defined high-stakes tests are beneficial if teachers have low intrinsic motivation, or when pupils are “high cost” learners. Teaching to the test may also be beneficial if test scores are a reliable gauge of productivity-enhancing skills (Hanushek, 2011).

3.3 Institutional background

3.3.1 Standardised assessment in England

The precursor to the mandatory standardised assessment of pupils in English state schools was the adoption of the Education Reform Act in 1988. The Act harmonised the curriculum and organisational structure of schools across England. The legislation introduced the national curriculum. All state schools were expected to deliver the curriculum which defined four key stages (KS) of schooling: grades 1 and 2 (KS1); grades 3 to 6 (KS2); grades 7 to 9 (KS3), and grades 10 and 11 (KS4). KS1 and KS2 are typically taught at primary schools and KS3 and KS4 at secondary schools.

Before the Act, pupils in English schools were formally assessed only at the end of secondary schooling (also the end of KS4) when they would sit examinations in nationally recognised O-Level or CSE qualifications. However, to support the national curriculum, the Act specified that pupils should be assessed at the end of each key stage “for the purpose of ascertaining what they have achieved in relation to the attainment targets for that stage” (Education Reform Act 1988, 1, 2 (2)).

The Task Group on Assessment and Teaching (TGAT) was responsible for developing the new assessment system, which they determined must satisfy several distinct purposes. The system should: provide information on the achievement of pupils; enable teachers to plan the next stage(s) for pupils; provide information on the aggregated achievement of pupils (to evaluate the functioning of schools and teachers); and provide information to parents to inform school choice decisions (Whetton, 2009).

KS1 assessment was introduced in 1991, while KS2 and KS3 assessment followed in 1994. KS4 attainment would continue to be measured by achievement in secondary school qualifications. At that time, SATs consisted of standardised tests and teacher assessments (which had notionally equal status) at all key stages[1]. The tests are externally marked and are subject to stringent procedures for maintaining standards. Teacher assessments are subject to an external moderation procedure.

Until recently, the assessments measure attainment using integer grades known as national curriculum levels, which range from one to eight. Pupils were expected to be working at level 2 at the end of KS1. Pupils should make two levels of progress between each key stage, meaning that pupils should achieve levels 4 and 6 at the end of KS2 and KS3, respectively. Secondary school qualifications use different grading systems but are mapped onto national curriculum levels to measure pupils' progress at secondary schools. In the first three key stages, pupils are assessed in the core subjects of English, maths and science. English assessment consists of an overall national curriculum level, and separate levels for reading, writing, speaking and listening, and spelling, punctuation and grammar depending on the key stage and policy at the time of assessment.

The arrangements for SATs were mostly unchanged until 2005 when the government reformed KS1 assessment and dropped tests in favour of more detailed teacher assessments. Testing at KS3 met the same fate in 2009 when the government concluded

[1] SATs are officially known as National Curriculum Assessments but are rarely referred to as such.

that parents obtained the same information from GCSE exam results as KS3 test results. KS2 testing continued, although from 2009 onwards only a subset of schools is required to administer science tests to monitor national standards.

3.3.2 The KS2 SATs boycott

The teaching profession in the United Kingdom is highly unionised. Some teacher unions have been vocal critics of standardised testing, and, on occasion, opposition to SATs tests led to threats to boycott them. In a 1993 National Union of Teachers (NUT) ballot, 90 per cent of teachers said they would support a boycott of KS3 SATs tests. The government bowed to pressure and cancelled that year's SATs tests (Whetton, 2009). In April 2003, at their annual conference, the NUT voted unanimously to hold an official ballot for a boycott of SATs for the 2003/04 school year. However, as the turnout of an indicative ballot the following December was only 34 per cent, the boycott never materialised. Although, 86 per cent of those who voted were in favour of a boycott. In May 2006, another union, the National Association of Headteachers (NAHT) debated asking parents to take their eleven-year-olds out of school on the days of SATs tests as part of their campaign against KS2 SATs. The union ultimately did not adopt this position.

Despite their multiple failed attempts to boycott SATs tests, the two unions would successfully lead a boycott of KS2 (age eleven, end of primary school) SATs tests during the 2009/10 school year. At their annual conference in April 2009, NUT delegates voted to ballot members on whether they would be willing to refuse to administer the SATs tests. One month later, 94 per cent of delegates at the NAHT conference instructed their union to continue to campaign against SATs including, as a last resort, holding a ballot on whether to boycott SATs. Both motions stated that a boycott would be a last resort and that dialogue with the Department for Children

Schools and Families (DCSF) was preferred. In November that year, the two unions jointly asked their entire membership if they supported a boycott of SATs. This was not an official ballot, under industrial legislation, but was an exercise in gauging the mood of their membership. 75 per cent of respondents backed a boycott, but turnout for this vote was perceived to be low (around 25 per cent of the unions' membership, rising to 35 per cent of those in leadership positions). In late January, the two unions announced that an official ballot compliant with salient industrial legislation would be held.

Only primary school head teachers, their deputies and assistant head teachers were balloted. The two unions combined represented the leadership of 80 per cent of England's primary schools. Head teachers were asked: "In order to protect your terms and conditions of employment, are you prepared to take industrial action short of strike action to frustrate the administration of national curriculum tests in English and Mathematics?"

The ballot opened on 15th March 2010 and closed one month later. Results were immediately announced. 61 per cent of NAHT and 75 per cent of NUT voters were in favour of a boycott. Turnout was 50 per cent and 34 per cent for NAHT and NUT members respectively. 28.8 per cent of eligible voters cast a vote in favour of the boycott. On 21st April 2010, the unions confirmed that they would be pressing ahead with industrial action in the form of the boycott. The KS2 SATs were scheduled for the week commencing the 10th May 2010, meaning there was a very short window between the boycott being confirmed by the unions and the tests being sat.

Participation in the boycott was the sole prerogative of head teachers since legally only they could oversee certain aspects of the tests' administration. 26 per cent of primary schools boycotted the KS2 tests in the 2009/10 school year. Figure 3.1 shows the proportion of pupils without a valid KS2 test mark in English and maths for eight cohorts of year six pupils. In other school years, compliance with SATs tests was high. Of the unaffected cohorts shown in Figure 3.1, fewer than 0.5 per cent of pupils unexpectedly fail to sit both English and maths KS2 SATs tests.

It is essential to understand what the average implications of the SATs boycott were for the affected pupils. Answers to two questions will reveal the nature of the treatment. Firstly, when during the school year did a head teacher decide to participate in the boycott? Secondly, how did the decision to participate in the boycott affect primary schools' conduct?

Suppose head teachers decided to participate in the boycott at the last possible moment: on the morning of the first test. In this scenario, treatment merely is that pupils do not sit the tests and do not receive a test mark. Alternatively, suppose that the decision to boycott does not affect the conduct of the school in any tangible fashion. In this second scenario, the nature of the treatment is identical. However, finally suppose that head teachers decide to participate in the boycott at the start of school year and that the decision to boycott meaningfully alters the behaviour of the school. In this third scenario, the treatment differs. Treatment is that pupils do not sit the test, nor receive a test mark, and are exposed to a school year's worth of altered school inputs (perhaps a reduced emphasis on "teaching to the test"). These are the two limiting cases of treatment.

These two questions are empirical. Unfortunately, data did not exist to answer them. I rely on institutional circumstances to argue that schools were likely to delay their boycott participation decision and that schools were unlikely to change their behaviour in anticipation of the boycott.

Head teachers and their employers, school governing bodies, had statutory obligations to ensure that the KS2 SATs test arrangements were implemented. Head teachers that failed to carry out this responsibility were in breach of their employment contracts. The local government employers' association advised school governing bodies to deduct pay from head teachers who participated in the boycott. The decision to participate in the boycott was costly to head teachers. I argue that this cost was prohibitively high before the NUT and NAHT's call for their members to take part in industrial action. Head teachers were only protected from dismissal from their positions after the unions officially called for industrial action on 21st April 2010. For this reason, I contend that head teachers would have decided to participate in the boycott late in the school year, making any significant test preparation or "teaching to the test" effects unlikely.

I further argue that it is unlikely that primary schools who participated in the boycott were less prone to teach to the test in anticipation of the boycott. Primary schools can teach to the test for two reasons: they may feel it is a valid pedological approach, or they may want to maximise the performance data on which they are judged. If boycott participation were to induce a primary school to stop teaching to the test, then it must be that the school teaches to the test for the latter reason; the pedological value of teaching to the test is independent of the boycott's existence. If a primary school stopped teaching to the test before they were confident that the boycott would go ahead, and arguably before they knew whether other local primary schools would participate, then they would be risking the performance data which they have shown a keen interest to maximise by teaching to the test in the first place. Primary schools could only be confident that an organised boycott would take place from late April. Even at this point the government was seeking legal advice on the legality of the proposed boycott. If the government obtained an injunction, the unions could not have endorsed the boycott or else they would have faced sequestration of funds. It is also important to note that the industrial action did not cover classroom teachers. They had no union protection to refuse any instruction from school leaders to prepare their pupils for SATs tests.

With these arguments in mind, I propose that the nature of the treatment from participation in the boycott merely is that affected pupils do not participate in the test. Consequentially, pupils do not receive a mark but are similarly prepared for the test as those that do participate (until 12 school days before the test).

Other arguments exist, although primary survey data provides a degree of support for this argument. The incumbent head teacher of every primary school still operating which participated in the boycott was invited by email to complete a short online survey. Email addresses were sourced from the Department for Education (DfE). Over 95 per cent of emails were successfully delivered. The response rate was low (3.2 per cent). However, this was expected as nearly eight years had elapsed since the boycott. In total, 119 head teachers responded to the survey. Head teachers were asked to participate if they were confident they could recall the circumstances of the boycott. 63 per cent of head teachers confirmed that their school decided to participate in the boycott only once the ballot was officially called, compared to 17 per cent of headteachers who indicated that the decision to boycott was made at the beginning of the school year. Furthermore, 59 per cent of head teachers said that their participation in the boycott did not result in a "change in the emphasis placed on preparing pupils for the SATs tests". Only 14 per cent of head teachers said that there was "definitely" less emphasis placed on preparing pupils for the SATs tests. Of this last group, most head teachers had also indicated that the decision to participate in the boycott was made after 21st April 2010.

3.4 Data

The data source used in this chapter is an extract from the DfE's National Pupil Database (NPD). The NPD is an administrative dataset oft used by researchers, within its scope are all state schools and their pupils in England.

I combine pupil level School Census records which identify the school attended and contains demographic data such as gender, ethnicity, first or native language, special education needs (SEN) status, free school meal (FSM) eligibility and month and year of birth. Using each pupil's unique identifier, these records are linked to attainment data for KS1 to KS4. Attainment is measured in maths, English and science at KS1 to KS3, with a much broader set of attainment measures available at KS4.

The primary analysis uses data on the entire population of eight successive cohorts of English state school pupils. The sample begins with the cohort who completed KS2 in the 2003/04 school year and ends with those who finished KS2 in the 2010/11 school year. Pupils would complete compulsory schooling (KS4) five years later. Throughout this chapter, cohorts are referred to per their final school year of KS2. The KS2 SATs test boycott affected the 2009/10 cohort. In the sample, this cohort is preceded by six unaffected cohorts, and succeeded by another unaffected cohort.

Boycott participation was determined at the primary school level. For the affected cohort and within each primary school, I calculate the number of pupils who do not have a valid KS2 test mark as a percentage of the number of pupils expected to have a valid KS2 test mark. This is calculated separately for English and maths tests. A school is defined as participating in the boycott if, of their pupils expected to have a valid KS2 test mark, at least 90 per cent do not have a valid mark in either English or maths. There are only four primary schools for which this percentage is calculated as greater than 90 per cent but less than 100 per cent. This indicator of boycott participation exactly matches the boycott participation flag produced by the DfE. Every primary school that boycotted the KS2 maths tests also boycotted the English tests. However, a limited number of primary schools, equivalent to 0.4 per cent of the population of primary schools, boycott only the KS2 English tests.

Pupils are exempt from participating in the KS2 SATs tests if their teacher can objectively conclude that they are working below the level assessed by the tests. If a primary school's entire cohort of pupils is exempt from the KS2 SATs, then the school cannot take a stance on whether to participate in the boycott. No mainstream schools were in this situation. However, many primary schools catering exclusively for pupils with special educational needs were. These primary schools do not feature in the analysis. Primary schools were also dropped from the analysis if they did not have at least one observed pupil in each of the cohorts in the sample. Pupils attending all other primary schools feature in the analysis.

This chapter investigates the boycott participation effect on various measures of attainment. These measures are: KS2 teacher assessments, KS3 teacher assessments, and KS4 GCSE (and equivalents) attainment. Pupils receive separate teacher assessment in maths, English and science at the end of KS2 and KS3. Teacher assessments are graded using integer grades known as national curriculum levels (and sublevels in the case of KS1 teacher assessments).

The KS4 attainment measures are: the pupil's highest point score achieved in a maths GCSE qualification; highest point score achieved in an English GCSE qualification; the highest point score achieved in a science GCSE (or equivalent) qualification; an indicator of whether the pupil achieved five or more GCSEs (or equivalents) at grades

A* to A, and an indicator of whether the pupil achieved five or more GCSEs (or equivalents) at grades A* to C.

All attainment outcomes apart from the binary “threshold” GCSE attainment measures are standardised across the cohort to zero mean, unit standard deviation. Each attainment outcome is recorded for all cohorts except KS3 attainment data which is not available for the 2010/11 cohort as the DfE stopped collecting this information from schools.

KS1 teacher assessments in English, maths and science, conducted when the pupil is aged seven (four years before the KS2 SATs test and teacher assessments), are the sole measure of prior ability recorded. KS1 teacher assessment in the relevant subject is included as a covariate in the DiD models facilitating a value-added approach.

Two non-attainment outcomes are also considered: subject choice and school absenteeism. It is common for pupils to choose which subjects to study for secondary school qualifications throughout KS4. Although some subjects – such as English, maths and science – are compulsory. I measure subject choice in three ways. Firstly, I construct a dummy variable ‘EBacc’ which is equal to 1 if the pupil studies and enrols in GCSE qualifications for English, mathematics, history or geography, two sciences and a modern or accident language, and 0 otherwise. The EBacc (English Baccalaureate) subjects are believed by the government to be important for young people to study at KS4 level. Research suggests that studying an EBacc compatible curriculum is associated with improved further and higher education prospects (Moulton *et al.*, 2018). Secondly, I construct another dummy variable ‘STEM’ which is equal to 1 when a pupil works towards three or more GCSEs in STEM subjects, and 0 otherwise. Thirdly, I construct another dummy variable ‘Vocational’ which equals 1 if the pupil enrols in at least one vocational GCSE, and 0 otherwise. These three measures were proposed by Henderson *et al.* (2018).

Regarding school absenteeism, I measure the number of the sessions missed in a school year due to: authorised absences, unauthorised absences and both types of absences. Each school day consists of two sessions (a morning and an afternoon session), and there are at least 190 school days in a year. Unlike all other outcomes which are measured at pupil level, absentee data is available at a pupil by school year level. Data is available on absenteeism for each year of secondary school and the last two years of primary school, for pupils in the 2006/07 to 2010/11 KS2 cohorts.

3.5 Methodology

3.5.1 Attainment and subject choice outcomes

The impact on a pupil's attainment and subject choice from their primary school's participation in the SATs boycott is estimated using a series of difference-in-differences (DiD) models. I estimate the following baseline equation:

$$y_{isc} = \alpha_s + \alpha_c + \beta Boycott_s \times 2009/2010_c + \gamma' X_i + \varepsilon_{isc} \tag{3.1}$$

where i, s and c are pupil, primary school and cohort identifiers respectively. y_{isc} refers to a measure of attainment or subject choice. α_s represents primary school fixed effects: the mean difference in pupil attainment (or subject choice) between primary schools. This accounts for all time invariant primary school specific impacts on pupil attainment (or subject choice). α_c signifies cohort effects and controls for shocks that are common to all pupils in each cohort. $Boycott_s$ is an indicator variable equal to 1 if primary school s participated in the boycott, and 0 otherwise. $2009/2010_t$ is a second indicator variable equal to 1 if cohort c is the 2009/10 year six cohort, and 0 otherwise. β is the parameter of interest and represents the estimated average causal effect of treatment on the treated (ATT). X_i is a vector of the following pupil level control variables: female indicator; white ethnicity indicator; English is first language indicator; FSM eligibility indicator; month of birth effects. For models of attainment outcomes, X_i also contains KS1 teacher assessment point score (either the average teacher assessment, or the teacher assessment for the appropriate subject). This vector ostensibly controls for student sorting into primary schools based on observable pre-determined characteristics. ε_{ist} is the idiosyncratic error term containing all remaining variation in pupil attainment.

The DiD β estimator will be unbiased under three conditions. Firstly, in the absence of the boycott, outcomes for pupils enrolled in boycotting primary schools must share a common time (i.e. cohort) trend with pupils enrolled at other schools – the parallel trends assumption. That is, $\alpha_c^0 = \alpha_c^1 \; \forall c$ where superscripts denote membership of the control group (0) and treatment group (1). Secondly, for the cohort impacted by treatment, there are no unobserved shocks that are common to either group. That is, $E[\varepsilon_{is}^0 | c = 2009/10] = E[\varepsilon_{is}^1 | c = 2009/10]$. The final assumption, as is required when the DiD estimator is applied to repeated cross-sections, is that there are no unobserved compositional changes within the groups between cohorts.

The parallel trends assumption cannot be formally tested due to the missing counterfactual problem. However, robust evidence in its favour can be uncovered by testing for differential trends between the treatment and control groups for cohorts unaffected by the boycott. I estimate the following equation:

$$y_{isc} = \alpha_s + \alpha_c + \Sigma_{c=2004/05}^{2010/11} \beta_c Boycott_s \times c + \gamma' X_i + \varepsilon_{isc} \qquad (3.2)$$

where all identifiers and variables are defined as per equation (3.1). β_c denotes the estimated difference between the attainment of pupils in the treatment and control group relative to the baseline cohort (2003/04) conditional on a common cohort effect and the other covariates and school-fixed effects. If $\hat{\beta}_c \neq 0 \; \exists c \neq 2009/10$, then this means that there is an "effect" from being in the treatment group for a non-treated cohort, implying a departure from the parallel trends assumption. $\hat{\beta}_c = 0 \; \forall c \neq 2009/10$ indicates that the parallel trends assumption holds for non-treated cohorts but provides no information regarding its applicability for the treated cohort.

Since the identifying assumptions cannot be explicitly tested, they must be justified through a well-reasoned choice of control group. The assumptions are most credible when there are no differences between the mean observed and unobserved characteristics of the treatment and control groups, other than exposure to treatment (participation in the boycott).

3.5.2 Selecting the control group

Identifying an appropriate control group is challenging in this case. All primary schools led by a head teacher who held membership of the NUT or NAHT unions (over 80 per cent of English state primary schools) were eligible to participate in the boycott. It is not possible to observe the trade union affiliation of head teachers. In any case, head teacher trade union membership was not a definitive assignment rule as fewer than half of the primary schools eligible to participate in the boycott elected to do so.

Given the unobserved nature of the treatment assignment mechanism, a data-driven approach is used to identify a suitable control group. For each school, I estimate the propensity score of treatment and use a simple matching procedure to select the control group[2]. The propensity score is estimated using a Probit model. The model features, as independent variables, time-invariant school characteristics, and school-cohort characteristics separately for each of the pre-treatment years in the model. The time-invariant school characteristics are an urban setting indicator; a local authority-controlled school indicator; indicators for the school's most recent OFSTED *Overall Effectiveness* rating and local authority fixed effects. The school-cohort characteristics are: proportion that are white ethnicity; proportion that are female; proportion that speak English as their first language; proportion that are eligible for FSM; average prior attainment as measured separately by KS1 English, maths and science teacher assessments; school size; cohort size, and pupil-teacher ratio. The propensity score model does not include any of the dependent variables in the DiD models.

The left panel of Figure 3.2 depicts kernel density estimates of the propensity score of primary schools by boycott participation. Unsurprisingly, participating schools are estimated to have a much higher probability of involvement in the industrial action, with most of the non-participating schools having estimated propensity scores of less than 20 per cent. I attempt to match each boycotting primary school to a non-boycotting primary school but enforce a restrictive calliper of 0.001 to allow only close matches. Non-boycotting primary schools are not replaced once matched which maximises the size of the control group and ensures schools in the matched sample are equally weighted. There are many participating and non-participating schools, and the overlap of the estimated propensity scores between these two groups is significant. Of the 3,824 boycotting primary schools in the full sample 3,179 are matched to a non-boycotting primary school (a match rate of 83 per cent). The right panel of Figure 3.2 presents the

[2] For the avoidance of doubt, this is not a matched difference-in-difference estimator; this is a difference in difference estimator where matching is used to select a control group from the population of potential control group constituents.

estimated propensity score distribution of the matched sample, which are nearly identical between the boycott participators and non-participators.

Table 3.1 presents the mean of several school characteristics and school-cohort characteristics (averaged over the 2003/04 to 2008/09 cohorts) for primary schools that participated in the boycott and those that did not. The difference in the means between these groups of primary schools is estimated. The first three columns include the full sample of primary schools, and the last three columns include the matched sample. Raw attainment measures are reported, not their standardised counterparts.

In the full sample of schools, pupils enrolled in boycotting primary schools have lower attainment than other pupils at every key stage, and these differences are statistically different from zero at the one per cent significance level for all 16 attainment measures. However, the economic significance of these differences is often quite small but does increase throughout the key stages.

Overall, boycotting primary schools educated more disadvantaged pupils relative to non-boycotting primary schools. The proportion of the cohort eligible for FSM is 5.5 percentage points higher in primary schools that participated in the boycott (from a base of 23 per cent). Boycotting primary schools also educate 4.1 and 5.2 percentage points more pupils who, respectively, do not speak English as a first language and are ethnic minorities.

Addressing school characteristics, urban schools were much more likely to participate in the boycott than rural schools. The NAHT speculates that this is because it was easier for head teachers of urban schools to coordinate with other local head teachers and collectively decide whether to participate in the boycott. The proportion of boycotting primary schools rated ‘Outstanding’ or ‘Good’ by OFSTED was slightly larger than the corresponding proportion for primary schools that did not participate in the industrial action. This may be because these schools could rely on their OFSTED rating to signal their quality to parents and thus KS2 test data was less crucial to the primary school.

The final column shows the difference in the means for boycotting primary schools and non-boycotting primaries within the matched sample. None of the differences in the attainment measures is statistically significant at the one per cent significance level, nor are any of the differences in school characteristics. Two of the school cohort characteristics are statistically different from zero at the ten per cent significance level. The table indicates that the matching procedure has achieved its aim of achieving covariate balance between the treatment and control groups in the pre-treatment periods. One assumes that covariate balance between the groups is most likely to lead to balance in unobservable characteristics between the groups. The parallel trends assumption is most credible when there are no systematic observable or unobservable differences between the groups.

3.5.3 School absenteeism outcomes

Because school absenteeism outcomes are available on a yearly basis for pupils, a difference-in-difference-in-differences model can be estimated. I estimate the following baseline equation:

$$y_{iscg} = \alpha + \lambda_{sc} + \lambda_{gc} + \lambda_{sg} + \Sigma_{g=5,7,8,9,10,11}\, \beta_g ByctSch_s \times ByctChrt_c \times Grade_g + \gamma X_{iscg} + \varepsilon_{iscg} \tag{3.3}$$

Where i, s, c, g are pupil, primary school, cohort and grade identifiers respectively. α is the constant term. λ_{sc} is a primary school cohort fixed effect, λ_{gc} is a grade by cohort fixed effect and λ_{sg} is a primary school by grade fixed effect. $ByctSch_s$ equals 1 if primary school s participated in the boycott and 0 otherwise. $ByctChrt_c$ is equal to 1 if cohort c is the 2009/10 KS2 cohort (the pupils affected by the boycott) and otherwise is 0. $Grade_g$ is equal to 1 if the observation relates to attendance in year/grade g, and 0 otherwise. ε_{icsg} is an error term clustered at the primary school level. β_7 to β_{11} are the coefficients of interest and measure how absenteeism in years/grades 7 to 11 are affected by participation in the boycott. β_5 indicates how absenteeism in year/grade 5 is affected by the boycott. This coefficient estimate should be statistically zero, as this grade will have been completed in advance of the boycott.

3.6 Results

3.6.1 Main results

Table 3.2 represents the principal analysis of the impact of the SATs boycott on subsequent measures of pupil attainment. The table reports the point and standard error estimates of the parameter of interest from eleven different DiD models on the matched sample. Each model has a different outcome variable. The estimates found in columns (1) to (3) correspond to models in which the outcome variable is a subject-specific KS2 (age eleven) teacher assessment; the next three columns model KS3 (age 14) teacher assessments, and the final five columns consider KS4 (age 16, secondary school qualification) attainment measures. Models are estimated on a common sample of pupils. To facilitate comparison of the parameter of interest estimates across outcomes, the outcomes in columns (1) to (9) are standardised to zero mean, unit standard deviation. The outcomes in the last two columns are indicator variables.

Columns (1) and (2) indicate that pupils from school-cohorts that boycotted SATs tests were assessed to be higher achievers by their own teachers than previous and future cohorts in their school. KS2 teacher assessments in maths and English were, respectively, 0.023 and 0.046 standard deviations higher for pupils affected by the boycott. These estimates are statistically different from zero at the one per cent significance level. KS2 teacher assessments in science were seemingly unaffected by the boycott. KS2 teacher assessments are carried out at the time of the KS2 SATs boycott.

There are several reasons why these assessments might be inflated for pupils affected by the boycott. First, given that school-level test attainment data was necessarily missing for schools that participated in the boycott, the importance of teacher assessment data to the boycotting schools increased. OFSTED used school level teacher assessment data in the absence of test data. Teacher assessment data was also available to the public via school league tables. Therefore, primary schools that boycotted had a greater incentive to maximise KS2 teacher assessment outcomes in the boycott year than they ever did before or subsequently. Secondly, primary schools that engaged in the industrial action had an incentive to demonstrate it was a positive decision for their pupils. One mechanism of achieving this is through favourable teacher assessments. The boycott affected cohort were never due to sit a KS2 test in science, and there is no boycott effect on KS2 science teacher assessments. This is consistent with both arguments above. KS2 teacher assessments are important to consider as they are consequential for targets, set by the government, for pupils' secondary school attainment in English and maths (where KS2 test data is missing).

Columns (4) to (6) show the estimated boycott participation effect on KS3 teacher assessments. These teacher assessments are conducted three years after pupils should have completed the KS2 tests by teachers unconnected to the industrial action. Despite being assessed at a higher level of attainment in primary school, these subsequent assessments found affected pupils to be doing significantly worse. Evidence of an adverse boycott effect is found at the one per cent significance level in maths and science. Effect magnitudes are modest: respectively, 0.014 standard deviations and 0.024 standard deviations. The point estimate for KS3 English teacher assessment is -0.012 but is not precisely estimated. We would anticipate more significant impacts on maths and science as attainment in these subjects is more persuadable to school inputs, whereas English attainment is more susceptible to non-school inputs. KS3 teacher assessments are worth considering since, unlike KS4 outcomes, they are low stakes from the perspective of schools, and should not reflect test skills which qualification attainment might.

Columns (7) to (9) display the estimated effect of boycott participation on GCSE attainment in maths and English, and GCSE (or equivalent) attainment in science. The estimated effects of the boycott on GCSE attainment are comparable to the estimated effect on KS3 teacher assessment. Participation in the boycott is estimated to reduce maths attainment at this level by 0.019 standard deviations. This estimate is statistically different from zero at the one per cent significance level. The estimated effects on English attainment (-0.010 standard deviations) and science attainment (-0.012 standard deviations) are not precisely estimated. Column (10) indicates that boycott participation did not affect the likelihood of pupils achieving five or more GCSEs/equivalents at grades A* to A. However, Column (11) shows an adverse effect of 0.6 percentage points on the likelihood of pupils achieving five or more GCSEs/equivalents at grades A* to C.

Any differential trend in attainment outcomes by boycott participation would bias the estimates presented in Table 3.2. To gauge whether differential trends existed between the two groups of schools for cohorts unaffected by the boycott, Figures 3.3-3.6 presents the estimated effect of participating in the boycott on outcomes for cohorts affected and unaffected by the cohort. In more precise terms, the treatment group indicator interacts with cohort effects for all cohorts other than the first cohort in the sample.[3]

Figure 3.3 reports the estimated ‘effect’ of participating in the boycott on unaffected and affected cohorts for KS2 outcomes. There is no evidence of a differential trend in English, maths and science KS2 teacher assessments between the treatment and control schools for any of the unaffected cohorts. Figure 3.4 shows that there is no differential trend in KS3 maths teacher assessment. For both KS3 English and science teacher assessments, one of the five boycott participation effect estimates for unaffected cohorts is statistically non-zero at the five per cent significance level.

Figure 3.5 shows that none of the boycott participation effect point estimates on unaffected cohorts is statistically different from zero for GCSE maths and English. The boycott participation effect estimate on GCSE science on the cohort two-years prior affected cohort is statistically different from zero. However, the trend in this outcome between the treatment and control groups is otherwise identical. Finally, Figure 3.6 applies the flexible trends analysis for the threshold measures of overall GCSE attainment. All twelve point estimates of the boycott participation effect on unaffected cohorts are not statistically different from zero. For all outcomes studied, Figures 3.3 to 3.6 provides compelling, but not definitive, evidence in favour of the parallel trends assumption. Appendix Table 3.1 reproduces Figures 3.3 to 3.6 in tabular form.

The large sample size makes it possible to estimate heterogenous effects of boycott participation precisely. Treatment effect heterogeneity is investigated at pupil and school level.

[3] The interaction term with the first cohort is excluded to avoid perfect collinearity.

Panel A of Table 3.3 investigates heterogeneity in the estimated boycott participation effect by gender. At KS2 level, treatment effect heterogeneity exists only for the English teacher assessments. The boycott induced inflation in English teacher assessments is larger for boys than girls. The situation reverses at KS3 level, boycott participation is estimated to hurt boys' Maths and Science teacher assessments, but there is no corresponding effect for girls. Meanwhile, there is a negative effect on girls' English teacher assessments but no effect on boys'. At KS4 level, there is no statistical difference in the estimated effect on GCSE maths point scores and the likelihood of achieving five or more GCSEs at grades A* to C between boys and girls. However, boycott participation has a sizeable negative effect on girls' English attainment, but a slight positive effect on boys' English attainment. The situation reverses for science: boys are negatively impacted by the boycott, whereas there is not a statistically significant effect for girls.

Panel B considers treatment effect heterogeneity between white ethnicity pupils and ethnic minorities. For some outcomes, there are apparent differences in the boycott participation effect by ethnicity. At KS2 level, English and maths teacher assessments are most inflated for ethnic minorities. At KS3 level, white pupils are harmed more by the boycott than non-white pupils. At KS4 level, the adverse effect of boycott participation is generally greater for white pupils than ethnic minority pupils.

Panel C examines how the treatment effect estimate differs between pupils eligible for FSM and ineligible pupils. The inflation of KS2 teacher assessments in English and maths is much greater for pupils eligible for FSM. A consistent pattern is also found at KS3 and KS4 levels. Pupils who are ineligible for FSM are more adversely affected by boycott participation than eligible pupils. FSM eligible pupils are estimated to benefit from boycott participation.

FSM eligibility, ethnicity and to a lesser extent gender are associated with attainment. Treatment effect heterogeneity in these outcomes may be reflecting heterogeneity by ability. Ability is unobserved but can be proxied using prior achievement (as measured by KS1 teacher assessments). Panel D reports the estimates of a model that allows the boycott participation effect to vary between high and low prior achievers. Pupils whose average KS1 teacher assessment point score is greater than the expected level (17 points) are coded as high ability; all other pupils are coded as low ability. The KS2 teacher assessments of pupils with low prior attainment are inflated more than high prior attainment students. At KS3 and KS4 level, generally, the high prior attainment pupils are more adversely affected by boycott participation than low KS1 attainment pupils. Some of the differences in the estimated effects by prior attainment are large. For example, the boycott participation effect on KS3 English teacher assessment is -0.075 standard deviations for high KS1 attainment pupils, but +0.015 for low KS1 attainment pupils. For KS4 English attainment, the estimated effect is -0.096 standard deviations for prior high achievers and +0.32 for low prior achievers.

Table 3.4 presents estimates from models allowing the treatment effect to vary at the primary school level. In Panel A, the treatment effect varies between urban and rural primary schools. The positive effects of boycott participation on KS2 teacher assessments are greater for urban primary schools than rural primary schools. At KS3 level, the adverse effect of boycott participation exists for pupils who attended rural primary schools, but there is a zero effect for pupils who attended urban primary schools. Similarly, at KS4 level, the adverse effects of boycott participation are found for pupils who were enrolled at rural primary schools, but not those who attended urban primary schools.

In Panel B, the boycott participation effect is permitted to vary between primary schools that were given an ‘Outstanding’ or ‘Good’ overall effectiveness rating by OFSTED and those that received a ‘Satisfactory’ or ‘Inadequate’ rating. Primary schools that were highly rated by OFSTED did not inflate the KS2 maths teacher assessments for the boycott affected cohort, unlike lowly rated primaries. There was evidence of inflation in KS2 English teacher assessments in all schools, but this inflation was greater in poorly rated schools. A clear pattern exists at KS3 and KS4 level; there is no evidence of a boycott participation effect for pupils who attended poorly rated primary schools. However, evidence of a negative boycott participation effect is consistently found for pupils who attended ‘Outstanding’ or ‘Good’ primary schools.

Attention turns towards the additional (non-attainment) outcomes investigated. Table 3.5 investigates the impact of boycott participation on KS4 subject choice. Column (1) indicates that pupils were one percentage point less likely to make subject choices that were compatible with the English Baccalaureate specification if they were affected by the boycott. This equates to 3.3 per cent of the number of pupils in the sample who met the EBacc requirements. This effect is statistically significant at the one per cent significance level. Column (2) indicates that pupils were 0.6 percentage points less likely to enrol in three or more qualifications for STEM subjects. However, this effect is only statistically significant at the ten per cent significance level. This equates to 0.8 per cent of the number of pupils who did enrol in three or more STEM qualifications. Column (3) indicates that there is no evidence that pupils were either more likely or less likely to study at least one vocational GCSE qualification as a result of being affected by the boycott.

Table 3.6 presents results of the difference-in-difference-in-differences model for measures of school absenteeism. The dependent variables in columns (1) to (3) are, respectively, the number of sessions missed in a school year due to overall absence, authorised absence and unauthorised absences. The first row of coefficients contains the estimated effects of being affected by the boycott on absence during year/grade 5. This is the grade immediately before the KS2 SATs tests take place. Therefore, it is expected that these coefficients are not statistically different from zero. In grades 7 onwards (after the KS2 SATs tests should take place), there are negative estimated effects of boycott participation on overall absence and authorised absences. This indicates that pupils who were affected by the boycott are absent for fewer school sessions and that this is driven by fewer authorised absences rather than fewer unauthorised absences. However, the magnitude of the coefficient estimates are relatively small. For example, pupils are estimated to miss 0.191 fewer school sessions in grade 7 if they were affected by the boycott. This corresponds to 1.1 per cent of the average number of sessions missed per school year in the sample.

3.6.2 Robustness

The principal identifying assumption of the difference-in-differences estimator is that the attainment trends of pupils enrolled in primary schools that participated in the boycott is shared with pupils enrolled in non-boycotting primary schools in the absence of treatment.

To test the creditability of this assumption, I add primary school-specific linear cohort trends to the baseline model. These trends account for the average growth in outcomes of pupils from that school over the entire period, and the parameter of interest now represents the deviation from the predicted growth path caused by the boycott participation. If the parallel trends assumption holds, then the predicted growth paths should not differ systematically between the primary schools that did and did not participate in the boycott. This implies that the parameter of interest should be invariant to the inclusion of school-specific linear cohort trends. Panel A of Table 3.7 reports estimates from the school-specific linear cohort trend augmented model. The estimated boycott participation effect on KS3 teacher assessments are the only effects to be statistically different from the corresponding estimates from the preferred model at conventional significance levels.

There is a relationship between the quality of the primary and secondary schools attended by a pupil. Under the assumption that successive cohorts of pupils leaving primary school attend the same secondary schools in the same proportions, then the primary school fixed effect adequately captures the average secondary school experience of pupils from each primary school. This assumption is strong, however, and it is unlikely that the distribution of destinations for primary school leavers is fixed over time. If the secondary schools attended by boycott affected pupils systematically differed compared to the secondary school destinations of pupils from the same primary school in other cohorts, then this could bias the boycott participation estimate.

While there is no reason to suspect that this may have occurred, this identification threat can be dismissed by adding secondary school fixed effects to the model. By adding secondary school fixed effects, the estimated effect is identified by within secondary school variation in boycott participation, as well as within primary school variation in participation. Panel B of Table 3.7 reports estimates from such an enlarged model. The boycott participation effect estimates from this model are nearly identical to those from the main model and are estimated with similar precision. The similarity of the estimates between models with and without secondary school fixed effects, suggests that the mechanism driving these effect is not dependent on the time-invariant characteristics of the secondary school attended.

The sample includes a cohort of pupils that sat their KS2 tests in the school year after the boycott impacted cohort. In this school year, there was no industrial action, and over 99 per cent of pupils took the tests as expected. If there was a systematic effect of participation in the boycott on the primary school in the school years after the school year of the boycott, then it is unwise to regard the post-boycott cohort as unaffected by the boycott. For example, teachers may notice an observable difference in the behaviour of their pupils when the pressure of SATs tests is lifted by the boycott and may decide to place less emphasis on the tests in the future. If persistent boycott participation effects exist within primary schools, then this will bias the boycott participation effect for the affected cohort downwards if the persistent boycott participation effect on future cohorts is positive, and vice-versa. Panel C of Table 3.7 details estimates form a model estimated on a reduced sample excluding the 2010/11 cohort. The boycott participation effect estimates are again almost identical to the estimates of the main model.

When DiD is applied to repeated cross-sectional data, then the parallel trends assumption is necessary but not sufficient. The composition of the pupils in the treatment group must not systematically vary over time relative to pupils in the control group. Such compositional variations may confound the estimated treatment effect.

For each pupil, I calculate the mean of several pupil characteristics for their primary school cohort excluding themselves. These primary school cohort means are then added to the main model to control for the effects of observable variation in the composition of primary school cohorts over the school years. I calculate primary school cohort "means but one" for all included pupil control variables, except month of birth indicators. The boycott participation effect estimated from a model including these cohort variables are presented in Panel D of Table 3.7. Again, the coefficient estimates are very similar. Adverse effects of boycott participation on KS4 English and KS4 maths were estimated in the main model but lacked precision, but they are estimated with suitable precision in Panel D to be statistically different from zero at the five per cent significance level.

A school-level DiD model applied to the matched sample is used to formally test whether the composition of the affected cohort in boycotting primary schools systematically differs to these primary schools' other cohorts. The effect of boycott participation is estimated on ten different pupil characteristics averaged over each school's year six cohort. Effect estimates are presented in Table 3.8. For most of these characteristics, there is no difference between the treated and untreated cohorts of boycotting primary schools. Although, the treated cohort of boycotting primary schools are estimated to have lower levels of prior (KS1) attainment than untreated cohorts of these schools (columns (8) to (10)). However, the economic significance of the effect estimates is limited. One national curriculum level translates into six points. Therefore, the estimated difference in the prior maths attainment between the treated and untreated cohorts (of boycotting primary schools) is equivalent to 1 in 70 pupils receiving one national curriculum level lower. Given the limited magnitude of the effect size, it is argued that this does not present a worrisome identification threat.

3.7 Potential mechanisms

The preceding section presented robust evidence of a small adverse boycott participation effect on age fourteen teacher assessment attainment as well as age sixteen secondary school qualifications attainment. Attention now turns towards explaining why these effects are found.

Potential mechanisms can be categorised as occurring within primary schools or secondary schools. Alternatively, in equivalent chronological terms, mechanisms can operate before or during the boycott, or they can function post-boycott.

There are two potential primary school mechanisms: 1) not preparing, or revising, for the test as pupils might have done in the absence of the boycott, 2) not gaining as much experience of high stake formal examinations as pupils who did complete KS2 SATs. The salience of the first potential mechanism crucially depends on when schools stopped preparing for the SATs. I earlier argued that boycotting primary schools were unlikely to stop teaching to the test until late in the school year. It is possible that changes in test preparation in the two-week lead between the unions calling for industrial action and the tests taking place contribute to the estimated effects. Teaching to the test and revising for the test may be useful ways of learning, or at least good ways of learning the tested material. This will benefit future attainment if primary and secondary school curriculums are well aligned (as the National Curriculum intends).

Practising completing high stakes formal examinations is beneficial if exams measure exam skills as well as knowledge and understanding of the testable content. This is likely to be true. This mechanism is unlikely to explain the adverse boycott participation effect on KS3 teacher assessments – a measure of attainment which should not depend on exam skills. Furthermore, when the pupils in this sample were at school, GCSE qualifications were modular meaning that pupils would have sat very many examinations during KS4. Given that pupils sit many exams at secondary school, the lack of exam experience at primary school caused by boycott participation may be of limited importance.

Having discussed potential primary school mechanisms, I now discuss three potential secondary school mechanisms. The first relates to secondary school qualification grade targets. An important metric for secondary schools, and their teachers, is the percentage of their pupils who are deemed to have made the level of progress expected of them between KS2 and KS4 in English and maths. This success rate features prominently in secondary school league tables. A pupil has made expected progress if the 'gain' between their KS2 test grade and their KS4 GCSE grade is sufficiently large. For example, a pupil awarded grade (or level) 4 on their KS2 English SATs test, would be considered as having made expected progress if they secured a grade C in GCSE English. Whereas, a pupil who is awarded level 3 on their KS2 maths SATs test, would only be required to achieve a grade D at GCSE to have made expected progress.

Pupils affected by the boycott do not have SATs test levels. The levels awarded to these pupils on KS2 teacher assessments were used instead. Teacher assessments are inherently a more subjective measure of attainment than externally marked tests and are usually overstated relative to KS2 test data, plus the results section shows that teacher assessments were inflated because of the boycott. Table 3.9 shows using a difference-in-difference model that the KS2 maths (Column (1)) and English (Column (2)) level used as the baseline measure of attainment for expected progress calculations was inflated for boycott affected pupils. This means that boycott affected pupils would have to achieve better grades in GCSE English and maths to be considered as having made the expected level of progress than if they were not affected by the boycott.

If secondary schools wish to maximise their expected progress success rates, then they should reallocate their effort from pupils with boycott-induced high (difficult to obtain) expected progress related GCSE targets to non-boycott affected pupils for whom it should be easier to achieve the expected progress GCSE targets. Columns (3) and (4) of Table 3.9 show that boycott affected pupils were indeed less likely to make the progress expected of them in English and maths.

The heterogeneity analysis indicated that the adverse effects at KS3 and KS4 are greater for high prior ability pupils. This pattern is consistent with the expected progress mechanism. Figure 3.7 shows the proportion of pupils achieving expected progress by KS2 attainment. Expected progress in maths and English is shown to be easier to achieve for more able pupils. Therefore, secondary schools that wish to maximise their expected progress success rates are likely to be focus on high ability pupils rather than low ability pupils. If the boycott induced secondary school teachers to reallocate their effort, then it is more likely to be reallocated from high ability boycott affected pupils to high ability non-boycott affected pupils.

Two further secondary school, or post-boycott, mechanisms are that missing KS2 test data: 1) significantly affected the ability of secondary schools to stream pupils by ability, and 2) impaired the ability of parents to gauge their child's attainment, and thus lessened the allocative efficiency of parental inputs. However, a compelling argument against these propositions is that KS2 test data is just one of many available signals of pupil attainment and ability. Secondary schools and parents could draw upon KS2 teacher assessment data in place of KS2 test data for affected pupils. A common practice in English secondary schools is to run formative assessment tests (particularly in the entry-year). These tests, as well as teacher observation and reporting, could be used to stream pupils and to inform parents about their child's progress.

3.8 Conclusions

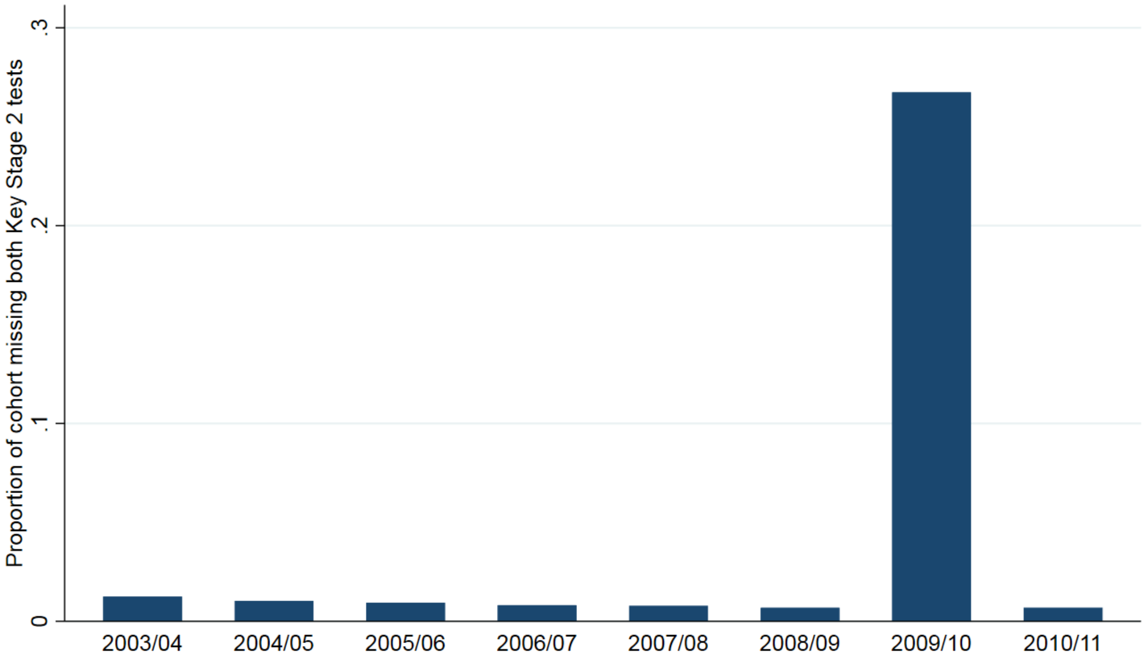
This chapter investigates whether there is a private benefit or cost to pupils from their own participation in mandatory standardised tests on their future teacher assessment and secondary school qualification attainment, independent of the existing school accountability arrangements.

Industrial action by head teachers in the form of a widespread, but hastily called, boycott of the 2010 KS2 SATs test is exploited as a natural experiment. I find evidence that pupils who were prevented from participating in the age eleven SATs tests by the boycott were assessed to be working at a lower level than their peers by their teachers at age fourteen. The GCSE maths attainment of boycott affected pupils are estimated to be 0.019 standard deviations lower than unaffected pupils. These findings provide evidence that there is a small private benefit to pupils from participation in mandatory standardised tests.

However, this private attainment benefit, may not arise through enhanced human capital accumulation. The chapter demonstrates that as a side effect of boycott participation, pupils are set artificially high secondary school expected progress targets. It is speculated that since secondary schools are incentivised to maximise their expected progress success rates, teacher effort may be reallocated away from boycott affected pupils to pupils without inflated expected progress targets. Other potential mechanisms exist, and arguments in their favour are evaluated.

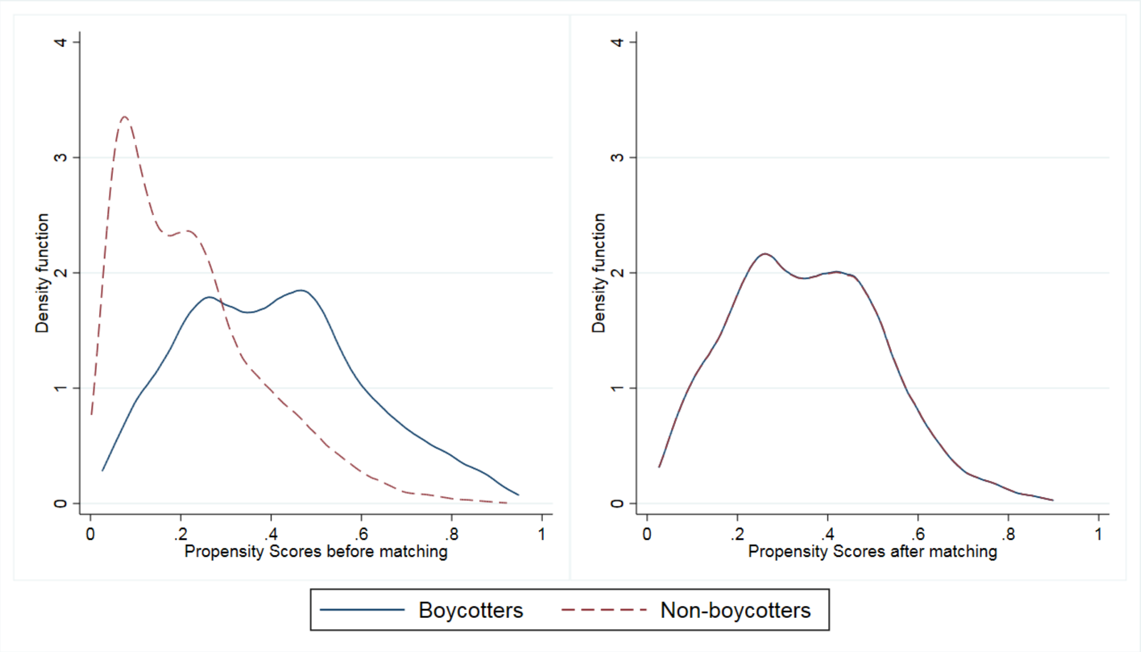
Separately, evidence is also reported that the boycott reduced the likelihood a pupil selected secondary school qualifications that were compatible with the English Baccalaureate. Boycott participation is also shown to slightly reduce the number of school sessions missed due to unauthorised absences.

Figure 3.1: Proportion of cohort unexpectedly without a valid mark in KS2 English and maths



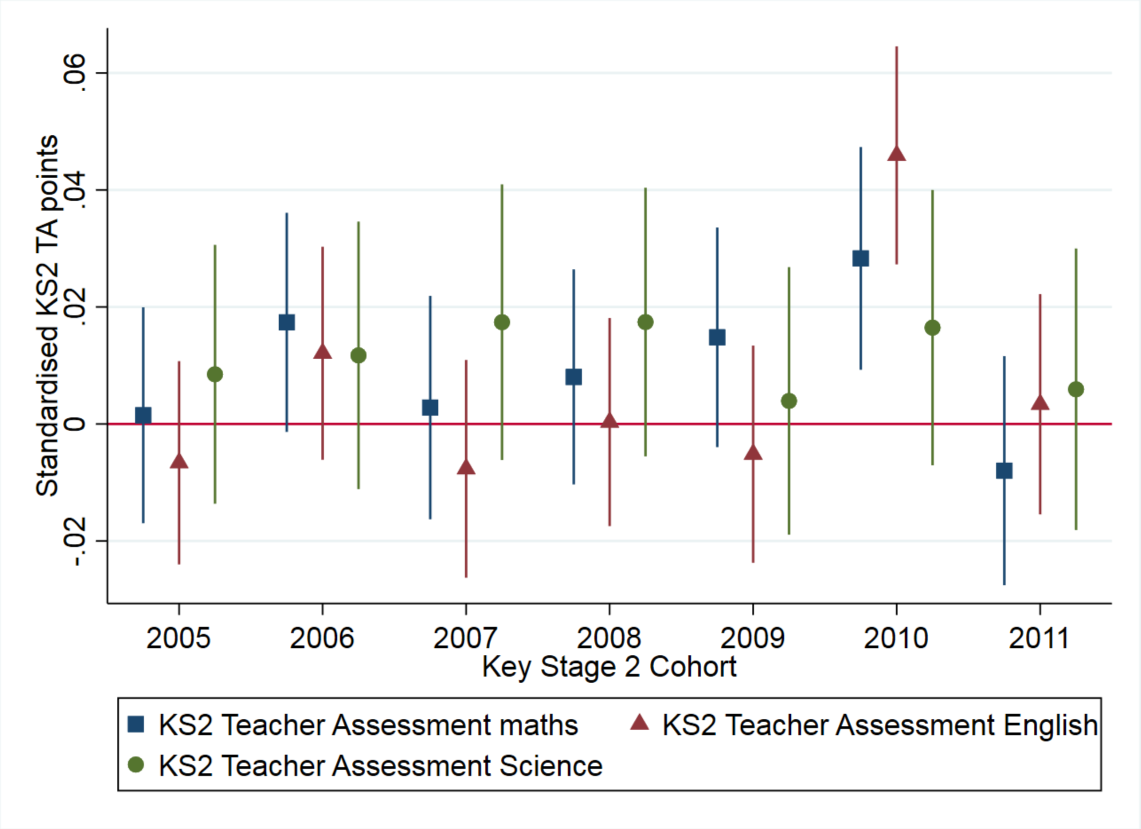
Notes: bar represents the proportion of the corresponding school year's year six cohorts unexpectedly missing a valid mark in Key Stage 2 English and maths SATs tests.

Figure 3.2: Estimated propensity score of boycott participation



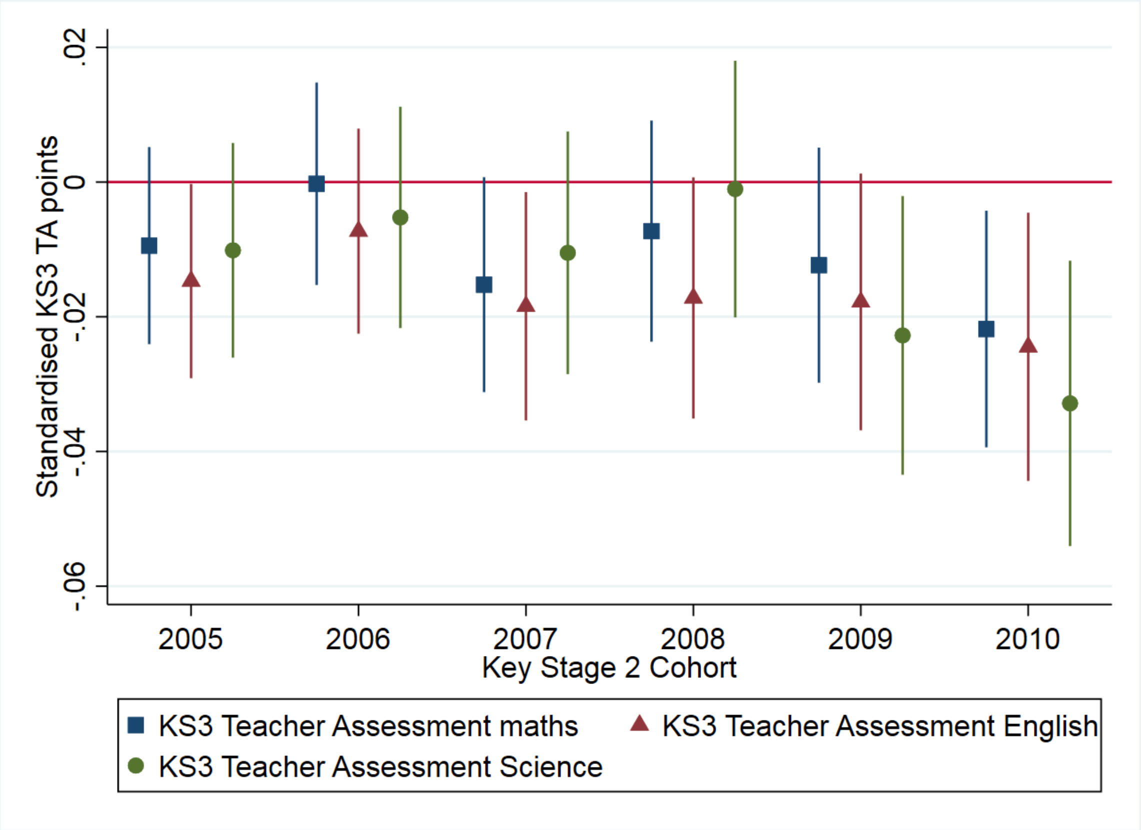
Notes: propensity score is estimated using a Probit model. The included time-invariant school characteristics are: urban setting indicator; local authority-controlled school indicator; indicators for the school's most recent Ofsted Overall Effectiveness rating and local authority fixed effects. The included school-cohort characteristics are: proportion of school cohort that has white ethnicity; proportion of school cohort that is female; proportion of school cohort that speaks English as their first language; proportion of school cohort that is eligible for free school meals; average prior attainment as measured separately by KS1 English, maths and science teacher assessments; school size; cohort size, and pupil-teacher ratio. These characteristics are included separately for each pre-treatment cohort. One to one matching without replacement and a 0.001 calliper. Match rate of 83 per cent.

Figure 3.3: Estimated difference in trends between boycotting and non-boycotting primary schools for Key Stage 2 teacher assessment outcomes



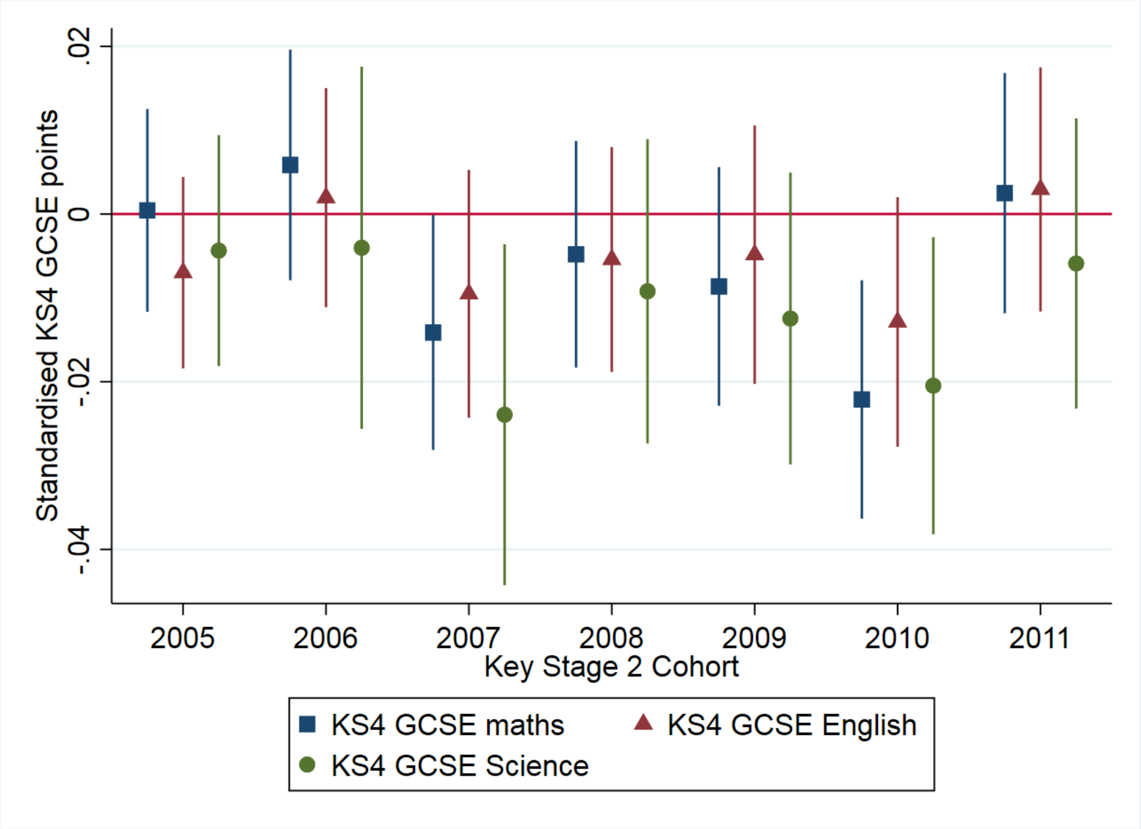
Notes: filled shape represents the point estimate of the difference in outcomes between pupils who attended boycotting primary schools (treatment group) and non-boycotting primaries (control group) relative to the 2003/04 cohort (conditional on a common cohort (time) effect). Vertical lines indicate 95 per cent confidence intervals**.** All specifications include cohort (time) effects, school fixed effects, student characteristics and prior attainment as measured by KS1 teacher assessment. Student characteristics are: female, white ethnicity, FSM eligible, English is first language, cohort size and month of birth effects. Robust standard errors clustered at primary school level are estimated.

Figure 3.4: Estimated difference in trends between boycotting and non-boycotting primary schools for Key Stage 3 teacher assessment outcomes



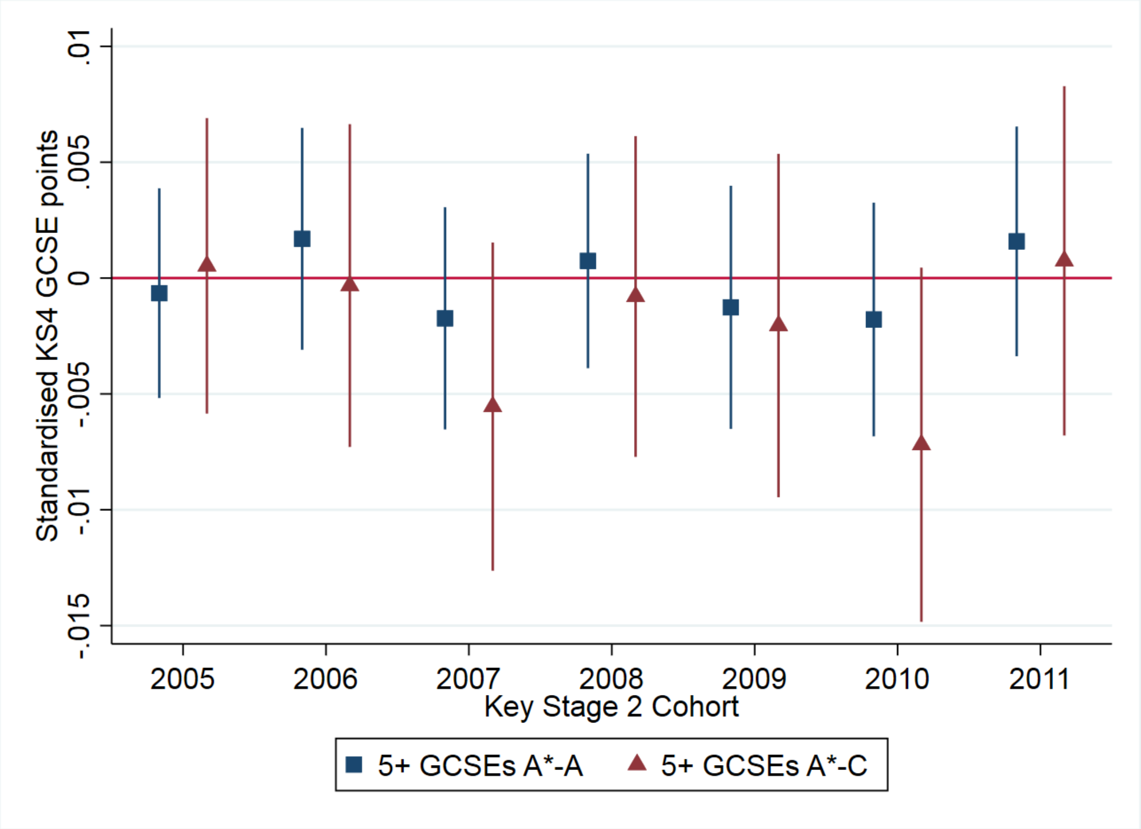
Notes: filled shape represents the point estimate of the difference in outcomes between pupils who attended boycotting primary schools (treatment group) and non-boycotting primaries (control group) relative to the 2003/04 cohort (conditional on a common cohort (time) effect). Vertical lines indicate 95 per cent confidence intervals**.** All specifications include cohort (time) effects, school fixed effects, student characteristics and prior attainment as measured by KS1 teacher assessment. Student characteristics are: female, white ethnicity, FSM eligible, English is first language, cohort size and month of birth effects. Robust standard errors clustered at primary school level are estimated.

Figure 3.5: Estimated difference in trends between boycotting and non-boycotting primary schools for Key Stage 4 outcomes (part 1)



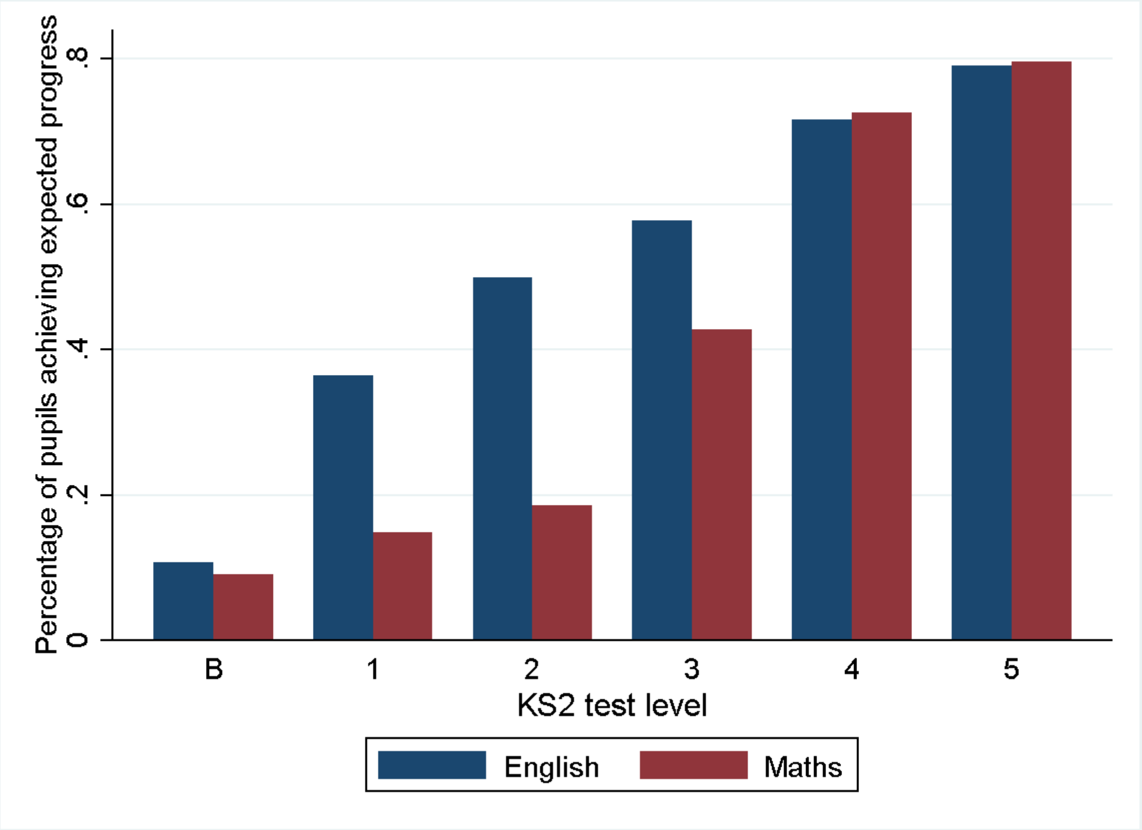
Notes: filled shape represents the point estimate of the difference in outcomes between pupils who attended boycotting primary schools (treatment group) and non-boycotting primaries (control group) relative to the 2003/04 cohort (conditional on a common cohort (time) effect). Vertical lines indicate 95 per cent confidence intervals**.** All specifications include cohort (time) effects, school fixed effects, student characteristics and prior attainment as measured by KS1 teacher assessment. Student characteristics are: female, white ethnicity, FSM eligible, English is first language, cohort size and month of birth effects. Robust standard errors clustered at primary school level are estimated.

Figure 3.6: Estimated difference in trends between boycotting and non-boycotting primary schools for Key Stage 4 outcomes (part 2)



Notes: filled shape represents the point estimate of the difference in outcomes between pupils who attended boycotting primary schools (treatment group) and non-boycotting primaries (control group) relative to the 2003/04 cohort (conditional on a common cohort (time) effect). Vertical lines indicate 95 per cent confidence intervals**.** All specifications include cohort (time) effects, school fixed effects, student characteristics and prior attainment as measured by KS1 teacher assessment. Student characteristics are: female, white ethnicity, FSM eligible, English is first language, cohort size and month of birth effects. Robust standard errors clustered at primary school level are estimated.

Figure 3.7: Percentage of pupils who achieve the expected level of progress between KS4 and KS2 in English and maths by KS2 attainment



Notes: calculated using data on the year six cohorts of 2005/06 to 2008/09. "B" denotes pupils who were working below national curriculum level 1 at KS2.

Table 3.1: Means of school and school-cohort characteristics for boycotting and non-boycotting primary schools in full and matched samples

	(1)	(2)	(3)	(4)	(5)	(6)
		Full sample			Matched sample	
	Non-boycotters	Boycotters	Difference (SE)	Non-boycotters	Boycotters	Difference (SE)
Key Stage 1 (age 7)						
Maths teacher assessment	15.872	15.726	0.173***	15.750	15.757	-0.008
			(0.023)			(0.030)
English teacher assessment	15.312	15.134	0.201***	15.176	15.174	-0.001
			(0.025)			(0.033)
Science teacher assessment	15.803	15.646	0.191***	15.684	15.687	-0.013
			(0.025)			(0.034)
Key Stage 2 (age 11)						
Maths teacher assessment	4.085	4.050	0.033***	4.062	4.056	0.005
			(0.005)			(0.007)
English teacher assessment	4.036	3.993	0.042***	4.007	4.002	0.004
			(0.005)			(0.007)
Science teacher assessment	4.216	4.185	0.033***	4.193	4.190	0.003
			(0.005)			(0.007)
English test raw mark	58.105	57.090	0.931***	57.506	57.245	0.251*
			(0.108)			(0.144)
Maths test raw mark	65.188	64.291	0.797***	64.642	64.378	0.232
			(0.139)			(0.187)
Number of schools	10,668	3,824		3,179	3,179	

Notes: all variables are school means for year six cohorts averaged over the 2003/04 to 2008/09 cohorts (pre-treatment cohorts). Differences are weighted by school size. ***, ** and * denote statistical significance at the 1 per cent, 5 per cent and 10 per cent levels respectively.

Table 3.1 (continued): Means of school and school-cohort characteristics for boycotting and non-boycotting primary schools in full and matched samples

	(1)	(2)	(3)	(4)	(5)	(6)
		Full sample			Matched sample	
	Non-boycotters	Boycotters	Difference (SE)	Non-boycotters	Boycotters	Difference (SE)
Key Stage 3 (age 14)						
Maths teacher assessment	5.788	5.690	0.092*** (0.009)	5.717	5.703	0.013 (0.012)
English teacher assessment	5.360	5.292	0.065*** (0.007)	5.312	5.303	0.010 (0.010)
Science teacher assessment	5.483	5.399	0.077*** (0.008)	5.423	5.414	0.006 (0.010)
Key Stage 4 (age 16)						
Highest point score in GCSE maths	39.006	38.073	0.810*** (0.084)	38.366	38.225	0.111 (0.114)
Highest point score in GCSE English	39.581	38.941	0.575*** (0.073)	39.116	39.051	0.053 (0.098)
Highest point score in GCSE (or equivalent) Science	37.867	36.528	1.136*** (0.115)	37.014	36.755	0.181 (0.152)
5 or more GCSEs A* to A	0.160	0.133	0.020*** (0.002)	0.142	0.137	0.002 (0.003)
5 or more GCSEs A* to C	0.585	0.547	0.032*** (0.004)	0.559	0.553	0.004 (0.005)
Number of schools	10,668	3,824		3,179	3,179	

Notes: all variables are school means for year six cohorts averaged over the 2003/04 to 2008/09 cohorts (pre-treatment cohorts). Differences are weighted by school size. ***, ** and * denote statistical significance at the 1 per cent, 5 per cent and 10 per cent levels respectively.

Table 3.1 (continued): Means of school and school-cohort characteristics for boycotting and non-boycotting primary schools in full and matched samples

	(1)	(2)	(3)	(4)	(5)	(6)
		Full sample			Matched sample	
	Non-boycotters	Boycotters	Difference (SE)	Non-boycotters	Boycotters	Difference (SE)
School cohort characteristics						
Proportion female	0.491	0.492	0.000	0.492	0.491	0.002**
			(0.001)			(0.001)
Proportion FSM eligible	0.123	0.159	-0.036***	0.150	0.150	0.000
			(0.003)			(0.004)
Proportion SEN	0.219	0.226	-0.009***	0.221	0.223	-0.001
			(0.002)			(0.003)
Proportion English is first language	0.074	0.106	-0.040***	0.096	0.096	-0.002
			(0.005)			(0.006)
Proportion white ethnicity	0.871	0.831	0.050***	0.840	0.841	0.003
			(0.006)			(0.008)
Cohort size	35.514	37.566	1.671**	36.651	36.968	0.419
			(0.789)			(0.996)
School size	240.403	259.738	-10.731***	252.882	252.877	1.780
			(3.489)			(4.455)
Pupil/teacher ratio	21.102	21.484	-0.044	21.462	21.462	0.127*
			(0.055)			(0.075)
Number of schools	10,668	3,824		3,179	3,179	

Notes: all variables are school means for year six cohorts averaged over the 2003/04 to 2008/09 cohorts (pre-treatment cohorts). Differences are weighted by school size. ***, ** and * denote statistical significance at the 1 per cent, 5 per cent and 10 per cent levels respectively.

Table 3.1 (continued): Means of school and school-cohort characteristics for boycotting and non-boycotting primary schools in full and matched samples

	(1)	(2)	(3)	(4)	(5)	(6)
		Full sample			Matched sample	
	Non-boycotters	Boycotters	Difference (SE)	Non-boycotters	Boycotters	Difference (SE)
School characteristics						
Urban location	0.681	0.794	-0.076*** (0.007)	0.764	0.774	-0.000 (0.008)
OFSTED Outstanding	0.136	0.129	0.010 (0.008)	0.127	0.123	0.006 (0.010)
OFSTED Good	0.501	0.519	-0.028** (0.011)	0.509	0.517	-0.012 (0.015)
OFSTED Requires improvement	0.333	0.338	-0.002 (0.011)	0.347	0.344	0.002 (0.014)
OFSTED Inadequate	0.030	0.014	0.020*** (0.003)	0.017	0.017	0.003 (0.004)
Local authority-controlled school	0.712	0.764	-0.050*** (0.008)	0.759	0.757	0.006 (0.011)
Number of schools	10,668	3,824		3,179	3,179	

Notes: all variables are school means for year six cohorts averaged over the 2003/04 to 2008/09 cohorts (pre-treatment cohorts). Differences are weighted by school size. ***, ** and * denote statistical significance at the 1 per cent, 5 per cent and 10 per cent levels respectively.

Table 3.2: Estimates of the effect of participation in the SATs boycott on pupil attainment from DiD models

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	Key Stage 2 teacher assessments			Key Stage 3 teacher assessments			Key Stage 4 GCSE and equivalents attainment				
	Maths	English	Science	Maths	English	Science	Maths	English	Science	5+ A*-A	5+ A*-C
Boycott_{st}	0.023^{***}	0.046^{***}	0.007	-0.014^{**}	-0.012	-0.024^{***}	-0.019^{***}	-0.010^{*}	-0.012^{*}	-0.002	-0.006^{**}
	(0.006)	(0.006)	(0.008)	(0.006)	(0.007)	(0.008)	(0.005)	(0.006)	(0.007)	(0.002)	(0.003)
Adjusted R^2	0.462	0.520	0.332	0.467	0.458	0.319	0.409	0.381	0.294	0.198	0.319
Observations	1,510,019	1,510,019	1,510,019	1,444,863	1,444,863	1,444,863	1,510,019	1,510,019	1,510,019	1,510,019	1,510,019

Notes: all specifications include cohort (time) effects, school fixed effects, student characteristics and prior attainment as measured by KS1 teacher assessment. Student characteristics are: female, white ethnicity, FSM eligible, English is first language, cohort size and month of birth effects. Robust standard errors clustered at primary school level in parentheses. ***, ** and * denote statistical significance at the 1 per cent, 5 per cent and 10 per cent levels respectively.

Table 3.3: Pupil level heterogeneity in the estimated effect of participation in the SATs boycott on pupil attainment

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	Key Stage 2 teacher assessments			Key Stage 3 teacher assessments			Key Stage 4 GCSE and equivalents attainment				
	Maths	English	Science	Maths	English	Science	Maths	English	Science	5+ A*-A	5+ A*-C
Panel A: Estimated boycott participation effect by pupil gender											
Boycott$_{st}$	0.024^{***}	0.058^{***}	0.004	-0.029^{***}	-0.004	-0.050^{***}	-0.020^{***}	0.010^{*}	-0.043^{***}	-0.006^{***}	-0.008^{***}
	(0.007)	(0.007)	(0.008)	(0.007)	(0.008)	(0.009)	(0.006)	(0.006)	(0.008)	(0.002)	(0.003)
Boycott$_{st}$ ×	-0.001	-0.022^{***}	0.007	0.030^{***}	-0.014^{***}	0.052^{***}	0.001	-0.041^{***}	0.062^{***}	0.009^{***}	0.005^{*}
Female	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.006)	(0.005)	(0.005)	(0.006)	(0.002)	(0.003)
Female	-0.050^{***}	0.083^{***}	0.000	-0.005^{***}	0.210^{***}	0.047^{***}	0.026^{***}	0.177^{***}	0.077^{***}	0.046^{***}	0.078^{***}
	(0.001)	(0.001)	(0.002)	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)
Adjusted R^2	0.462	0.520	0.332	0.467	0.458	0.319	0.409	0.381	0.295	0.198	0.319
Panel B: Estimated boycott participation effect by pupil ethnicity											
Boycott$_{st}$	0.081^{***}	0.070^{***}	0.023^{*}	0.011	-0.002	0.003	-0.027^{***}	-0.022^{***}	0.019^{*}	0.005	0.003
	(0.010)	(0.010)	(0.012)	(0.010)	(0.010)	(0.012)	(0.008)	(0.008)	(0.010)	(0.003)	(0.004)
Boycott$_{st}$ ×	-0.072^{***}	-0.029^{***}	-0.020^{*}	-0.032^{***}	-0.011	-0.034^{***}	0.010	0.016^{**}	-0.038^{***}	-0.008^{**}	-0.011^{**}
White ethnicity	(0.009)	(0.009)	(0.011)	(0.009)	(0.009)	(0.011)	(0.007)	(0.008)	(0.010)	(0.003)	(0.004)
White ethnicity	-0.045^{***}	-0.029^{***}	-0.025^{***}	-0.082^{***}	-0.091^{***}	-0.070^{***}	-0.143^{***}	-0.111^{***}	-0.170^{***}	-0.033^{***}	-0.067^{***}
	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)	(0.003)	(0.004)	(0.002)	(0.002)
Adjusted R^2	0.463	0.520	0.332	0.467	0.458	0.319	0.409	0.381	0.294	0.198	0.319
Observations	1,510,019	1,510,019	1,510,019	1,444,863	1,444,863	1,444,863	1,510,019	1,510,019	1,510,019	1,510,019	1,510,019

Notes: all specifications include cohort (time) effects, school fixed effects, student characteristics and prior attainment as measured by KS1 teacher assessment. Student characteristics are: female, white ethnicity, FSM eligible, English is first language, cohort size and month of birth effects. Robust standard errors clustered at primary school level in parentheses. ***, ** and * denote statistical significance at the 1 per cent, 5 per cent and 10 per cent levels respectively.

Table 3.3 (continued): Pupil level heterogeneity in the estimated effect of participation in the SATs boycott on pupil attainment

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	Key Stage 2 teacher assessments			Key Stage 3 teacher assessments			Key Stage 4 GCSE and equivalents attainment				
	Maths	English	Science	Maths	English	Science	Maths	English	Science	5+ A*-A	5+ A*-C
Panel C: Estimated boycott participation effect by FSM eligibility											
Boycott$_{st}$	0.015^{**}	0.038^{***}	0.001	-0.020^{***}	-0.019^{**}	-0.030^{***}	-0.028^{***}	-0.021^{***}	-0.010	-0.001	-0.009^{***}
	(0.006)	(0.006)	(0.008)	(0.006)	(0.008)	(0.008)	(0.005)	(0.006)	(0.007)	(0.002)	(0.003)
Boycott$_{st}$ ×	0.066^{***}	0.061^{***}	0.051^{***}	0.042^{***}	0.048^{***}	0.036^{***}	0.064^{***}	0.081^{***}	-0.014	-0.007^{**}	0.023^{***}
FSM eligible	(0.008)	(0.008)	(0.009)	(0.008)	(0.008)	(0.010)	(0.009)	(0.009)	(0.010)	(0.003)	(0.004)
FSM eligible	-0.156^{***}	-0.128^{***}	-0.214^{***}	-0.213^{***}	-0.186^{***}	-0.254^{***}	-0.272^{***}	-0.230^{***}	-0.314^{***}	-0.025^{***}	-0.119^{***}
	(0.002)	(0.002)	(0.003)	(0.002)	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.001)	(0.001)
Adjusted R^2	0.462	0.520	0.332	0.467	0.458	0.319	0.409	0.381	0.294	0.198	0.319
Panel D: Estimated boycott participation effect by pupil prior KS1 attainment											
Boycott$_{st}$	0.029^{***}	0.079^{***}	0.017^{**}	-0.017^{**}	0.015^{**}	-0.014^{*}	-0.007	0.032^{***}	-0.023^{***}	-0.009^{***}	0.005
	(0.007)	(0.007)	(0.008)	(0.007)	(0.008)	(0.008)	(0.006)	(0.006)	(0.007)	(0.002)	(0.003)
Boycott$_{st}$ × High	-0.016^{***}	-0.100^{***}	-0.030^{***}	0.008	-0.090^{***}	-0.034^{***}	-0.039^{***}	-0.128^{***}	0.033^{***}	0.022^{***}	-0.034^{***}
KS1 attainment	(0.006)	(0.006)	(0.007)	(0.006)	(0.007)	(0.007)	(0.005)	(0.005)	(0.007)	(0.003)	(0.003)
High KS1	0.183^{***}	0.185^{***}	0.401^{***}	0.267^{***}	0.213^{***}	0.437^{***}	0.140^{***}	0.184^{***}	0.459^{***}	0.143^{***}	0.050^{***}
attainment	(0.003)	(0.002)	(0.003)	(0.003)	(0.002)	(0.003)	(0.002)	(0.002)	(0.003)	(0.001)	(0.002)
Adjusted R^2	0.466	0.524	0.351	0.475	0.464	0.340	0.411	0.386	0.317	0.212	0.320
Observations	1,510,019	1,510,019	1,510,019	1,444,863	1,444,863	1,444,863	1,510,019	1,510,019	1,510,019	1,510,019	1,510,019

Notes: all specifications include cohort (time) effects, school fixed effects, student characteristics and prior attainment as measured by KS1 teacher assessment. Student characteristics are: female, white ethnicity, FSM eligible, English is first language, cohort size and month of birth effects. Robust standard errors clustered at primary school level in parentheses. ***, ** and * denote statistical significance at the 1 per cent, 5 per cent and 10 per cent levels respectively.

Table 3.4: Primary school level heterogeneity in the estimated effect of participation in the SATs boycott on pupil attainment

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	Key Stage 2 teacher assessments			Key Stage 3 teacher assessments			Key Stage 4 GCSE and equivalents attainment				
	Maths	English	Science	Maths	English	Science	Maths	English	Science	5+ A*-A	5+ A*-C
Panel A: Estimated boycott participation effects for rural and urban primary schools											
Boycott$_{st}$	-0.031***	0.023**	-0.014	-0.062***	-0.066***	-0.073***	-0.041***	-0.042***	-0.060***	-0.010**	-0.023***
	(0.011)	(0.011)	(0.014)	(0.011)	(0.016)	(0.014)	(0.009)	(0.011)	(0.013)	(0.004)	(0.005)
Boycott$_{st}$ ×	0.063***	0.027**	0.024*	0.055***	0.063***	0.056***	0.025***	0.038***	0.056***	0.010**	0.020***
Urban location	(0.011)	(0.011)	(0.014)	(0.011)	(0.016)	(0.015)	(0.009)	(0.011)	(0.013)	(0.004)	(0.005)
Adjusted R^2	0.462	0.520	0.332	0.467	0.458	0.319	0.409	0.381	0.294	0.198	0.319
Panel B: Estimated boycott participation effect for primary schools with a good or outstanding OFSTED Overall Effectiveness rating and those without											
Boycott$_{st}$	0.056***	0.074***	0.019*	0.012	0.010	-0.004	0.000	0.008	-0.008	-0.002	0.002
	(0.009)	(0.009)	(0.011)	(0.009)	(0.010)	(0.011)	(0.007)	(0.007)	(0.010)	(0.002)	(0.004)
Boycott$_{st}$ × Good	-0.052***	-0.043***	-0.019	-0.041***	-0.035***	-0.032***	-0.031***	-0.029***	-0.006	-0.000	-0.012***
/Outstanding sch.	(0.009)	(0.009)	(0.012)	(0.009)	(0.010)	(0.012)	(0.008)	(0.008)	(0.010)	(0.003)	(0.004)
Adjusted R^2	0.463	0.520	0.332	0.467	0.458	0.319	0.409	0.381	0.294	0.198	0.319
Observations	1,510,019	1,510,019	1,510,019	1,444,863	1,444,863	1,444,863	1,510,019	1,510,019	1,510,019	1,510,019	1,510,019

Notes: Urban Location and Good/Outstanding school indicators are absorbed by the primary school fixed effect. All specifications include cohort (time) effects, school fixed effects, student characteristics and prior attainment as measured by KS1 teacher assessment. Student characteristics are: female, white ethnicity, FSM eligible, English is first language, cohort size and month of birth effects. Robust standard errors clustered at primary school level in parentheses. ***, ** and * denote statistical significance at the 1 per cent, 5 per cent and 10 per cent levels respectively.

Table 3.5: Estimates of the effect of participation in the SATs boycott on subject choice from DiD models

	(1)	(2)	(3)
	EBacc	3+ STEM	Any Applied
Boycott$_{st}$	-0.010***	-0.006*	0.001
	(0.003)	(0.003)	(0.005)
Dependent variable mean	0.307	0.750	0.158
Observations	1,834,708	1,834,708	1,834,708
Adj. R-Square	0.139	0.139	0.123

Notes: All specifications include cohort (time) and school effects and student characteristics. Student characteristics are: female, white ethnicity, FSM eligible, English is first language, cohort size and month of birth effects. Robust standard errors clustered at primary school level in parentheses. ***, ** and * denote statistical significance at the 1%, 5% and 10% levels respectively.

Table 3.6: Estimates of the effect of participation in the SATs boycott on school absences from DiDiD models

	(1) Overall Absence	(2) Authorised Absences	(3) Unauthorized Absences
Boycott school × boycott cohort × grade 5	-0.097	-0.028	-0.069
	(0.098)	(0.093)	(0.045)
Boycott school × boycott cohort × grade 7	-0.191**	-0.194**	0.003
	(0.093)	(0.086)	(0.047)
Boycott school × boycott cohort × grade 8	-0.126	-0.188**	0.062
	(0.104)	(0.091)	(0.058)
Boycott school × boycott cohort × grade 9	-0.245**	-0.183*	-0.061
	(0.112)	(0.095)	(0.066)
Boycott school × boycott cohort × grade 10	-0.154	-0.081	-0.073
	(0.124)	(0.097)	(0.080)
Boycott school × boycott cohort × grade 11	-0.126	-0.184	0.058
	(0.160)	(0.118)	(0.104)
Dependent variable mean	16.938	13.717	3.220
Observations	7,703,942	7,703,942	7,703,942
Adj. R-Square	0.081	0.056	0.077

Notes: All specifications include cohort (time) and school effects and student characteristics. Student characteristics are: female, white ethnicity, FSM eligible, English is first language, cohort size and month of birth effects. Robust standard errors clustered at primary school level in parentheses. ***, ** and * denote statistical significance at the 1%, 5% and 10% levels respectively.

Table 3.7: Robustness checks

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	Key Stage 2 teacher assessments			Key Stage 3 teacher assessments			Key Stage 4 GCSE and equivalents attainment				
	Maths	English	Science	Maths	English	Science	Maths	English	Science	5+ A*-A	5+ A*-C
Panel A: Add primary school-specific linear time trends to the preferred model											
Boycott_{st}	0.026^{***}	0.045^{***}	0.007	-0.006	0.000	-0.014	-0.019^{***}	-0.012^{**}	-0.009	-0.002	-0.006^{**}
	(0.006)	(0.006)	(0.008)	(0.007)	(0.008)	(0.009)	(0.005)	(0.005)	(0.006)	(0.002)	(0.003)
Adjusted R^2	0.476	0.533	0.352	0.475	0.469	0.331	0.418	0.391	0.302	0.322	0.737
Observations	1,510,019	1,510,019	1,510,019	1,444,863	1,444,863	1,444,863	1,510,019	1,510,019	1,510,019	1,510,019	1,510,019
Panel B: Add secondary school fixed effects to the preferred model											
Boycott_{st}	0.025^{***}	0.047^{***}	0.008	-0.015^{***}	-0.010	-0.025^{***}	-0.019^{***}	-0.009^{*}	-0.011^{*}	-0.002	-0.007^{***}
	(0.006)	(0.006)	(0.008)	(0.006)	(0.007)	(0.007)	(0.005)	(0.005)	(0.006)	(0.002)	(0.003)
Adjusted R^2	0.474	0.528	0.349	0.497	0.500	0.371	0.450	0.442	0.345	0.244	0.340
Observations	1,509,742	1,509,742	1,509,742	1,443,901	1,443,901	1,443,901	1,509,742	1,509,742	1,509,742	1,509,742	1,509,742

Notes: all specifications include cohort (time) effects, school fixed effects, student characteristics and prior attainment as measured by KS1 teacher assessment. Student characteristics are: female, white ethnicity, FSM eligible, English is first language, cohort size and month of birth effects. Robust standard errors clustered at primary school level in parentheses. ***, ** and * denote statistical significance at the 1 per cent, 5 per cent and 10 per cent levels respectively.

Table 3.7 (continued): Robustness checks

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	Key Stage 2 teacher assessments			Key Stage 3 teacher assessments			Key Stage 4 GCSE and equivalents attainment				
	Maths	English	Science	Maths	English	Science	Maths	English	Science	5+ A*-A	5+ A*-C
Panel C: Estimate the preferred model on the sample excluding the post-boycott cohort											
Boycott_{st}	0.021^{***}	0.047^{***}	0.007	-0.014^{**}	-0.012	-0.024^{***}	-0.018^{***}	-0.008	-0.012	-0.001	-0.006^{**}
	(0.007)	(0.007)	(0.009)	(0.006)	(0.007)	(0.008)	(0.005)	(0.006)	(0.007)	(0.002)	(0.003)
Adjusted R^2	0.456	0.518	0.336	0.467	0.458	0.319	0.406	0.385	0.296	0.198	0.321
Observations	1,310,208	1,310,208	1,310,208	1,444,863	1,444,863	1,444,863	1,310,208	1,310,208	1,310,208	1,310,208	1,310,208
Panel D: Add primary school cohort control variables to the preferred model											
Boycott_{st}	0.019^{***}	0.042^{***}	0.002	-0.019^{***}	-0.015^{**}	-0.029^{***}	-0.023^{***}	-0.013^{**}	-0.016^{**}	-0.003	-0.008^{***}
	(0.006)	(0.006)	(0.008)	(0.006)	(0.007)	(0.008)	(0.005)	(0.005)	(0.007)	(0.002)	(0.003)
Adjusted R^2	0.468	0.524	0.338	0.471	0.461	0.324	0.413	0.383	0.298	0.199	0.322
Observations	1,509,961	1,509,961	1,509,961	1,444,802	1,444,802	1,444,802	1,509,961	1,509,961	1,509,961	1,509,961	1,509,961

Notes: all specifications include cohort (time) effects, school fixed effects, student characteristics and prior attainment as measured by KS1 teacher assessment. Student characteristics are: female, white ethnicity, FSM eligible, English is first language, cohort size and month of birth effects. Robust standard errors clustered at primary school level in parentheses. ***, ** and * denote statistical significance at the 1 per cent, 5 per cent and 10 per cent levels respectively.

Table 3.8: DiD models testing the association between primary school cohort composition and exposure to the SATs boycott

	(1) % same school as KS1 school	(2) % white ethnicity	(3) % female	(4) % first language is English	(5) % with SEN	(6) % eligible for FSM	(7) Average cohort size	(8) Average KS1 English	(9) Average KS1 maths	(10) Average KS1 science
Boycott_{st}	-0.005 (0.003)	0.002 (0.002)	-0.000 (0.003)	0.001 (0.001)	0.008^{***} (0.003)	0.003^{*} (0.002)	-0.065 (0.145)	-0.070^{***} (0.025)	-0.086^{***} (0.026)	-0.038 (0.027)
Adjusted R^2	0.865	0.952	0.051	0.927	0.489	0.757	0.938	0.650	0.560	0.566
Observations	50,864	50,864	50,864	50,864	50,864	50,864	50,864	50,864	50,864	50,864

Notes: all specifications include cohort (time) effects and school fixed effects. Robust standard errors clustered at primary school level in parentheses. ***, ** and * denote statistical significance at the 1 per cent, 5 per cent and 10 per cent levels respectively.

Table 3.9: Estimates of the effect of participation in the SATs boycott on the measure of KS2 attainment used for expected progress calculations and the indicator of achieving expected progress

	(1) KS2 maths level (baseline measure) used in expected progress calculations (standardised)	(2) KS2 English level (baseline measure) used in expected progress calculations (standardised)	(3) Expected progress achieved in maths (=1 if expected progress achieved, 0 otherwise)	(4) Expected progress achieved in English (=1 if expected progress achieved, 0 otherwise)
Boycott_{st}	0.046^{***} (0.006)	0.051^{***} (0.006)	-0.024^{***} (0.003)	-0.020^{***} (0.003)
Adjusted R^2	1,292,130	1,292,130	1,083,643	1,083,643
Observations	0.488	0.521	0.189	0.134

Notes: all specifications include cohort (time) effects, school fixed effects, student characteristics and prior attainment as measured by KS1 teacher assessment. Student characteristics are: female, white ethnicity, FSM eligible, English is first language, cohort size and month of birth effects. Robust standard errors clustered at primary school level in parentheses. ***, ** and * denote statistical significance at the 1 per cent, 5 per cent and 10 per cent levels respectively.

Table A3.1: **Estimated difference in trends between boycotting and non-boycotting primary schools for attainment outcomes**

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	Key Stage 2 teacher assessments			Key Stage 3 teacher assessments			Key Stage 4 GCSE and equivalents attainment				
	Maths	English	Science	Maths	English	Science	Maths	English	Science	5+ A*-A	5+ A*-C
Boycott$_s$ × 5yrs	0.001	-0.007	0.008	-0.009	-0.015**	-0.010	0.000	-0.007	-0.004	-0.001	0.001
pre-boycott	(0.009)	(0.009)	(0.011)	(0.007)	(0.007)	(0.008)	(0.006)	(0.006)	(0.007)	(0.002)	(0.003)
Boycott$_s$ × 4yrs	0.017*	0.012	0.012	-0.000	-0.007	-0.005	0.006	0.002	-0.004	0.002	-0.000
pre-boycott	(0.010)	(0.009)	(0.012)	(0.008)	(0.008)	(0.008)	(0.007)	(0.007)	(0.011)	(0.002)	(0.004)
Boycott$_s$ × 3yrs	0.003	-0.008	0.017	-0.015*	-0.018**	-0.011	-0.014**	-0.010	-0.024**	-0.002	-0.006
pre-boycott	(0.010)	(0.009)	(0.012)	(0.008)	(0.009)	(0.009)	(0.007)	(0.008)	(0.010)	(0.002)	(0.004)
Boycott$_s$ × 2yrs	0.008	0.000	0.017	-0.007	-0.017*	-0.001	-0.005	-0.005	-0.009	0.001	-0.001
pre-boycott	(0.009)	(0.009)	(0.012)	(0.008)	(0.009)	(0.010)	(0.007)	(0.007)	(0.009)	(0.002)	(0.004)
Boycott$_s$ × 1yr	0.015	-0.005	0.004	-0.012	-0.018*	-0.023**	-0.009	-0.005	-0.012	-0.001	-0.002
pre-boycott	(0.010)	(0.009)	(0.012)	(0.009)	(0.010)	(0.011)	(0.007)	(0.008)	(0.009)	(0.003)	(0.004)
Boycott$_s$ ×	0.028***	0.046***	0.016	-0.022**	-0.024**	-0.033***	-0.022***	-0.013*	-0.020**	-0.002	-0.007*
Boycott year	(0.010)	(0.010)	(0.012)	(0.009)	(0.010)	(0.011)	(0.007)	(0.008)	(0.009)	(0.003)	(0.004)
Boycott$_s$ × 1yr	-0.008	0.003	0.006				0.002	0.003	-0.006	0.002	0.001
post-boycott	(0.010)	(0.010)	(0.012)				(0.007)	(0.007)	(0.009)	(0.003)	(0.004)
Adjusted R^2	0.462	0.520	0.332	0.467	0.458	0.319	0.409	0.381	0.294	0.198	0.319
Observations	1,510,019	1,510,019	1,510,019	1,444,863	1,444,863	1,444,863	1,510,019	1,510,019	1,510,019	1,510,019	1,510,019

Notes: all specifications include cohort (time) effects, school fixed effects, student characteristics and prior attainment as measured by KS1 teacher assessment. Student characteristics are: female, white ethnicity, FSM eligible, English is first language, cohort size and month of birth effects. Robust standard errors clustered at primary school level in parentheses. ***, ** and * denote statistical significance at the 1 per cent, 5 per cent and 10 per cent levels respectively.

Chapter Four:

The Impact of Low Stakes Standardised Test Grades on Subsequent School Outcomes

Standardised tests often facilitate school accountability, and pupils usually receive grades (or other forms of feedback) based on their test performance. However, providing feedback to pupils is not necessary for school accountability purposes. This chapter evaluates the effect of receiving integer grades based on a series of low-stakes standardised tests taken by eleven-year-olds in England. The chapter uses raw test marks, typically unobserved by pupils, and grade thresholds to implement a sharp regression discontinuity design. The results indicate that just passing the cut-off to achieve a higher grade in these tests leads to an improvement in secondary school qualification attainment. The estimated effect of just crossing a grade cut-off on secondary school attainment is typically larger for economically disadvantaged pupils. The chapter finds no evidence of an effect on school attendance.

4.1 Introduction

Standardised testing has become commonplace in schools as policymakers across the globe introduce test-based school accountability systems. If there are consequences for school-level or class-level test performance, then two pervasive principal-agent problems can be mitigated. The first is the inability of parents to observe the teaching quality of a school directly. The second is the lack of information on teacher effort available to head teachers. The prevailing consensus in the literature is that introducing test-based accountability improves the attainment of pupils (Figlio and Loeb, 2011).[1] Indeed, recent work suggests the pressures of test-based accountability can lead to higher college participation rates and improved earnings for pupils (Deming *et al.*, 2016).

If the purpose of a standardised test is to provide data for an accountability system, then there is no requirement to give pupils any feedback on their achievement in the tests. However, standardised tests, such as those used in England's primary school accountability system, often purportedly serve both a summative and a formative function. These standardised tests are intended to provide parents with information about their child's attainment and assist teachers with their planning for their pupils' subsequent learning (Whetton, 2009). Without this function, parental opposition to the testing of seven and eleven-year-olds might prove politically inhibitive.

The inclusion of the formative function implicitly assumes that providing pupils with feedback on their standardised test performance is useful and harmless. However, if standardised test performance measures a mere subset of meaningful attainment at school or is an otherwise noisy attainment measure, then performance feedback is likely

[1] Albeit achievement on standardised tests usually defines attainment, which may measure a mere subset of desirable skills and knowledge.

to provide limited insight to pupils, their parents and teachers. Furthermore, if performance feedback is coarse, misleading, or prone to misinterpretation, then incorrect inferences may be drawn about a pupil's progress to date. Future decisions taken by pupils, their parents and teachers based on these inferences may differ from those made with accurate information about attainment to date. In other words, the assumption that feedback is beneficial is not trivial.

This chapter evaluates the consequences of the feedback arrangements for a series of standardised tests sat by eleven-year-olds in England's state-funded schools between the late 1980s and 2015. Specifically, this chapter investigates whether the feedback influenced school attendance immediately after the standardised tests and attainment on important secondary school qualifications (GCSEs – General Certificates of Secondary Education – and their equivalents) five years subsequent.

Under the feedback arrangements in question, pupils were typically unaware of their raw mark on the standardised tests but instead received feedback in the form of a discrete numerical grade. Over 94 per cent of pupils received either a 'level 3', 'level 4' or 'level 5' grade, which respectively denotes achievement below, at or above the nationally expected standard for their age. Raw test mark thresholds, unknown by both pupils and markers, exclusively determined which grade each pupil achieved in each subject. These arrangements meant that, in the region of these thresholds, a pupil who scored one mark higher than an otherwise similar pupil could receive markedly different feedback about their progress at school to date. Similarly, a pupil who scores just enough marks to pass the threshold to receive the grade denoting attainment at the expected standard would receive identical feedback to another who narrowly falls short of the mark threshold to achieve the grade representing attainment above that standard.

This chapter combines data on the raw marks scored by pupils on the standardised tests with knowledge of the grade thresholds to implement a sharp regression discontinuity design. This methodology disentangles the effect of marginally passing the mark threshold to receive a grade (henceforth referred to as a ‘grade feedback effect’) from underlying achievement on that test for pupils who scored close to the grade threshold.

This chapter finds evidence that pupils who just pass the age eleven English language test mark cut-off for the grade denoting attainment above the expected standard perform two per cent of a standard deviation better in their English language secondary school qualifications. Those who just pass the maths test mark cut-off for the grade representing attainment at the expected standard perform 3.6 per cent of a standard deviation better in their maths secondary school qualifications. Both effects are statistically different from zero at the one per cent significance level. While small, these effects nonetheless indicate that the age eleven standardised tests are not unambiguously “no stakes” for pupils, as performance feedback influences pupils’ subsequent attainment. There is no evidence of an effect on school attendance during the school year following the standardised test, leading to speculation that the mechanism of the effect on attainment is not through pupil effort, at least not at the extensive margin.

Further analysis reveals that the grade feedback effect varies by pupils’ socioeconomic circumstances. For economically advantaged pupils the estimated effect on English language secondary school qualification achievement of grade feedback on the English language standardised tests is precisely zero. However, for economically disadvantaged pupils, there are modest positive returns to just passing the English language test cut-off for the grades denoting performance at the expected standard and above the expected standard; 2.9 per cent and 6.6 per cent of a standard deviation respectively. Both estimates are statistically significant at the one per cent level. For maths, there is evidence of a positive effect on maths secondary school qualification achievement of crossing the test mark cut-off for the expected standard grade for both groups of pupils. The point estimates are larger and more precise for disadvantaged pupils, but the differences between the groups are not statistically significant.

Not only does this indicate that age eleven standardised test grades polarise subsequent high-stakes secondary school exam performance, but also that there is a socioeconomic gradient in this effect. Pupils who fall just short of the English language test mark cut-off for a better grade will perform less well in important English language secondary school qualifications if they are economically disadvantaged, but not if they are economically advantaged. As such, the results of this chapter speak to the literature on how socioeconomic gaps in educational achievement accumulate (Anders, 2012; Chowdry *et al.*, 2013). Other heterogeneity analysis finds statistically significant feedback effects among white pupils but is unable to estimate a precise effect among non-white pupils.

That two pupils who are identical in all aspects other than that the first scored a single mark below a grade threshold while the other scored one mark more will have, on average, different outcomes in high stakes examinations suggest that the age eleven standardised tests are not indeed low-stakes from the pupil perspective. Furthermore, it is undesirable that this effect be far stronger for socioeconomically disadvantaged pupils than for all other pupils. This implies the stakes are higher for the disadvantaged group, which is troubling since that by age eleven they will have already fallen behind their more advantaged peers on average.

The findings of this chapter do not suggest that providing feedback on standardised test performance impairs human capital accumulation. However, the findings do indicate that it is necessary to consider the nature of standardised test feedback provided to pupils. It should be a concern to policymakers that grades awarded on a purportedly low stakes standardised test series influence pupils' achievement in high stakes secondary school qualifications five years later. There are many factors outside of the control of pupils that can impact attainment on a single standardised test.

For example, primary schools will adopt a variety of approaches to test preparation; some will intensively prepare for the standardised tests, whereas other head teachers will feel less need to focus intensively on the test. A second factor is that pupils cannot be confident how markers will interpret their potential answers; this is particularly true in the English language standardised tests, which include long-form answers where there is more subjectivity in the marking process than in maths. Finally, there is, of course, an element of good or bad luck behind every academic test result. If the pupil has the potential to score near a grade threshold in the standardised test, then having good luck will not only mean they achieve a higher grade on that test but also that they will achieve more in secondary school qualifications five years later. This is not a characteristic of a system that offers every pupil the same opportunity to reach his or her potential.

4.2 Institutional Background

Compulsory schooling typically lasts for twelve years in England. Pupils usually begin the reception grade at age four or five. Grades one to eleven follows reception. Grade retention is rare, so pupils finish compulsory schooling in the year they turn sixteen[2].

[2] Since 2015 pupils in England must, until the age of eighteen, either continue in full-time education, combine part-time education with part-time work/volunteering or enrol in a study at work scheme.

Pupils complete reception and grades one to six in a primary school and the remaining grades in a secondary school.

The introduction of the national curriculum in 1989 split grades one to eleven into four Key Stages (KS). KS1 spans grades one and two; KS2 covers grades three to six; KS3 extends across grades seven to nine, while KS4 covers the final two grades of secondary school. When the government introduced the curriculum, ministers decided to assess pupil attainment at the end of each KS to judge how effectively schools were delivering it. National Curriculum Assessments (NCAs) are the mechanism for measuring achievement. For the cohort studied in this chapter, NCAs included externally marked standardised tests at the end of KS2; informally and more commonly known as SATs tests. NCAs also consist of externally moderated teacher assessments at the end of KS1, KS2 and KS3. Achievement in nationally recognised secondary school qualifications (typically GCSEs – General Certificates of Secondary Education) represents attainment at the end of KS4.

Pupils complete KS2 SATs tests in the final term of primary school, which comprises a series of English and maths tests. There are three maths tests: a test to be completed with a calculator, another to be attempted without and a mental arithmetic test. The English tests include a reading test, a long-form writing test and a short-form writing test (which also assesses spelling).

The testing authority maps the total raw mark for the series of tests for each subject onto National Curriculum Levels (henceforth referred to as levels) by setting several cut-off points throughout the total raw mark distribution. The cut-off points vary slightly year upon year. Levels range from 1 to 8, and themselves correspond to broad subject-specific descriptors of attainment. Teacher assessments at KS1 to KS3 are also denominated using levels. Pupils are expected to be working at level 2 at the end of KS1 and to progress between each key stage at a rate equivalent to two levels. Therefore, the expectation is that pupils achieve level 4 and level 6 at the end of KS2 and KS3 respectively. Pupils and parents are aware of these expectations. The template produced by the Department for Education (DfE) for schools to report KS2 SATs test results explicitly states that level 3 'represents achievement below the nationally expected standard', level 4 'represents achievement at the nationally expected standard', and level 5 'represents achievement above the nationally expected standard'.

Pupils can be awarded levels 2 to 5 based on their achievement in the KS2 SATs tests. Table 4.1 shows the distribution of levels in English and maths across pupils who sat the tests in May 2009. Three per cent of pupils are recorded as working below the level assessed by the English tests (3.5 per cent for the maths tests). These pupils have gained exemption from sitting the tests which is at their head teacher's discretion[3]. Pupils who are recorded as not being awarded a level sit the test but do not perform well enough to be awarded level 2.

The KS2 SATs tests are designed to be accessible for pupils working towards levels 2 to 5; therefore, it is unsurprising fewer than two per cent of pupils are awarded level 2 or no level in the English and maths tests. Approximately fifteen per cent of pupils are awarded level 3 in the English and maths tests. The most commonly awarded level is level 4; 45 per cent of pupils are awarded it in English and 52 per cent in maths. However, a sizeable proportion of pupils receive the highest available level; just over a third of pupils in English and 29 per cent in maths.

[3] Head teachers do not need to ask for permission to make a pupil exempt from the SATs tests, but they must justify their decision in a report copied to the pupil's parents and the chair of the school's governing body. Parents have the right of appeal.

Table 4.2 characterises the distribution of levels across different subgroups of pupils. Roughly the same proportion of free school meal (FSM) eligible and ineligible pupils achieve level 4 in English and maths, but FSM eligible pupils are approximately twice as likely to receive the level 3 grade, and half as likely to receive the level 5 grade. In English, girls are less likely to receive the level 2 and level 3 grades and are much more likely to be graded level 5 relative to boys. However, in maths, girls are more likely to receive grades level 2 and level 3 than boys. Girls are also less likely to achieve the top grade than boys are. Although, overall gender differences are smaller in maths than in English. Regarding differences between ethnicities, white pupils relative to non-white pupils are marginally less likely to be graded level 2, and more likely to be graded level 3. There is little variation between the two groups regarding achieving level 4.

External markers mark the tests. Markers typically do not mark more than one type of test and do not know the raw mark threshold for each level – the testing authority does not set this in advance of the marking process. As such, there is limited scope for raw marks to be manipulated such that pupils do not fall narrowly short of any level threshold. The DfE advises primary schools not to report total raw test marks to pupils. However, parents can access this data through data protection legislation.

Compliance with the statutory obligation to complete the tests is generally high. Despite opposition to SATs tests from some, but not all, teacher unions, schools do not refuse to administer them – with the notable exception of the 2010 boycott studied in the previous chapter. Furthermore, unlike in some US states, there is not a substantial parent-led protest movement against SATs tests, in which parents withdraw their children from schools during test days.

Pupils learn which levels they have achieved in their SATs tests and their internal teacher assessments at the end of the school year in which they complete the tests. The tests are notionally low stakes from the pupils' perspective as their primary function is to facilitate the assessment of school-level performance. The KS2 SATs tests do not influence which secondary school pupils attend. Both expressions of secondary school preferences by parents and allocations of places by the local authority take place well before pupils sit KS2 SATs tests. KS2 SATs tests are also not used to determine whether a child is offered a place at one of England's few remaining academically selective grammar schools.

However, secondary schools are given information on the KS2 SATs test performance of incoming pupils by their primary schools. There is a lack of nationally representative data on how English schools assign pupils to class groups, although both 'setting' and 'streaming' are thought to be prevalent in English secondary schools. Setting refers to schools grouping pupils into subject-specific classes based on their apparent ability in that subject; whereas streaming is when schools group pupils into classes fixed across many subjects based on their perceived level of general ability. SATs test data is the only ability signal always available to secondary schools when pupils join from primary school. SATs test performance may, therefore, be consequential for class assignment, and thus the composition of within secondary school peer groups.

On the other hand, secondary schools are known to treat KS2 SATs test data with suspicion. The prevailing wisdom among secondary teachers is that KS2 SATs test performance is heavily dependent on the primary school pupils attended. Not only is there variation in the effectiveness of primary schools, but there is also variation in the emphasis primaries place on securing good KS2 SATs test performance. Some primary schools rigorously prepare for KS2 SATs tests, whereas others barely acknowledge their existence and adopt a more holistic approach to teaching. Furthermore, secondaries will quickly accumulate additional data on the ability of their incoming pupils. Over two-thirds of English secondaries are known to use cognitive ability tests (CATs) at the point of admission. As such, it is likely KS2 SATs test performance has a minimal impact on class assignment particularly beyond the first term of secondary school.

A more likely reason why KS2 SATs tests are not truly low stakes for pupils is the unequivocal link between the levels awarded for KS2 SATs tests and GCSE achievement targets. At the time that this chapter's sample completed secondary school, the government's preferred, and headline, measure of secondary school pupil progress (and thus secondary school effectiveness) was the school-level percentage of pupils 'who make progress expected of them' in English and maths (Leckie and Goldstein, 2017).

What constitutes expected progress for a pupil is determined solely by the level they received in their KS2 SATs tests. For example, a pupil awarded level 4 on their English KS2 SATs test is deemed to have progressed as expected if they achieve a C grade or better in GCSE English and is deemed not to have made sufficient progress if they achieve a D grade or worse. Similarly, a pupil awarded level 5 on their English KS2 SATs test must achieve a B grade or better at GCSE level to be recorded as having made expected progress.

The expected progress metric effectively places a lower limit on GCSE English and maths target grades. Schools must strive for as many of their pupils as possible to be deemed as having made expected progress, even if other available information suggests that what constitutes expected progress is unrealistic. As the expected progress measure featured prominently in school league tables, schools closely monitored their expected progress rate. As such, secondary schools and their teachers are acutely aware of each pupil's KS2 SATs test levels.

4.3 Literature Review

In recent years, a literature has emerged investigating the causal effect of narrowly passing certain grade (or other feedback) thresholds in test scores on subsequent educational outcomes. The literature predominantly considers the effects of just achieving a 'pass' grade (relative to just falling short of a pass grade) on high school exit exams (HSEE) on outcomes such as high school graduation, choice of major, college enrolment and time to degree completion. This review briefly summarises notable international contributions before discussing three papers using English data.

Partly due to the No Child Left Behind Act, most US states require their high school students to pass state-mandated tests to receive their high school diploma. Several papers have used a regression discontinuity design to compare outcomes between students who narrowly achieve a pass grade on their HSEEs and those that narrowly fall short of passing. The first, Martorell (2004), studied the impacts of the Texan exit exam on a sample of students from the 1990s and found no effect on early high school dropout behaviour but did find evidence of reduced post-secondary attainment. Similarly, Reardon *et al.* (2010) find no evidence of an effect of narrowly crossing the pass threshold on California's HSEEs on various high school outcomes.

However, using data from New Jersey, Ou (2010) finds that students who just fall short of a HSSE pass grade cut-off were more likely to drop out than those who just meet the cut-off. The estimated effect is largest for the maths exam. Effects are also more pronounced for economically disadvantaged and ethnic minority students. Papay *et al.* (2010) report that just falling short of the pass threshold on a first attempt at the English or maths HSEE does not affect the average Massachusetts student's probability of graduating. Although low-income urban students who narrowly fall short of the pass grade threshold for the maths HSEE have an eight-percentage-point lower graduation rate than similar students who narrowly crossed the same threshold. This is sizable given that the graduation rate is 74 per cent for the latter group. The difference in the findings of Martorell (2004) and Reardon *et al.* (2010), and those of Ou (2010) and Papay *et al.* (2010), could be a result of variation in the nature of the tests, the relative 'difficulty' of achieving a pass grade and the student population (Papay *et al.*, 2014).

The literature also considers post-high school outcomes. Students who narrowly fall short of the pass grade threshold in Massachusetts HSEEs are between seven to sixteen per cent less likely to enrol in college within two years of their cohort's high school graduation (Papay *et al.*, 2014). Narrowly meeting the pass grade threshold in the French national HSEE leads to an improvement in average college peer quality of around 13 per cent of a standard deviation (as measured by HSEE attainment) and an increase in the likelihood of enrolment in a postsecondary STEM institution by between 19 and 24 percentage points (Canaan and Mouganie, 2018).

Whereas the papers above consider the impacts of narrowly meeting the pass grade threshold on HSEEs, Papay *et al.* (2016) investigate the impact of passing grade thresholds on low-stakes tests that have no official consequences for students. The authors find that, for urban low-income students, narrowly passing the test mark score cut-off for a grade with positive connotations increases the probability of attending college. The effect is greatest among students who previously reported they did not plan to attend a four-year college. These findings are surprising given that students are aware of their raw test score. That is, the grade itself provides no additional information to students.

The Advanced Placement (AP) programme in the US offers a college-level curriculum to high achieving high school students; and uses examinations to assess participants' attainment. Participants do not know their raw exam score, but they receive an integer grade from 1 to 5 denoting increasing levels of proficiency. Narrowly crossing the raw exam score threshold for an integer grade that grants college credits increases the probability of completing a bachelor's degree within four years of high school graduation (Smith *et al.*, 2016). Just passing the exam score cut-off to obtain a higher integer grade for an AP exam also increases the probability that a student will major in that exam subject at college by five per cent (Avery *et al.*, 2017).

A small number of papers consider the effect of narrowly crossing grade thresholds in the context of English schools. While this chapter considers the effect of just surpassing KS2 SATs test grade cut-offs, Alcott (2017) investigates the effect of just crossing KS3 SATs test grade cut-offs on GCSE attainment and enrolment in A-level and University degree courses. The paper uses Longitudinal Survey of Young People in England (LSYPE) data, which follows 15,770 young people between the ages of 14 and 25. The LSYPE sample attempted KS3 SATs tests in 2004. KS3 SATs tests were abolished after 2008, but the marking and feedback arrangements were directly comparable to those of the KS2 SATs tests studied in this chapter. Adopting a sharp regression discontinuity design, the paper reports evidence that just passing the threshold to obtain the English grade denoting the expected attainment standard has a positive effect on subsequent GCSE attainment, and enrolment in both A-level and University courses for pupils with low socio-economic status; whereas there is no pattern of effects for mid- and high socio-economic status pupils.

Using the same dataset, Sartarelli (2011) investigates the impact of just surpassing grade thresholds on KS2 SATs tests in English, maths and science on a range of self-reported behaviours. The research suggests obtaining the level 4 grade does not have a meaningful impact on pupil behaviour as measured by outcomes such as the likelihood of playing truant, being suspended or permanently excluded from school, or receiving a police caution. Both Sartarelli (2011) and Alcott (2017) are frustrated by the LSYPE dataset whose small sample size constrains the precision of effect estimates. LSYPE also suffers from severe sample attrition which limits the external validity of the findings.

More recent work by Machin *et al.* (2018) explores the effect of just passing the mark threshold for a crucial GCSE grade on post-schooling outcomes. They use a unique dataset on the marks achieved during assessments for a GCSE English language qualification before pupils have the opportunity to appeal as an instrument for whether the pupil achieves a C grade at GCSE – a partial fuzzy regression discontinuity design (Battistin and Rettore, 2008). The authors find that narrowly falling short of the C grade threshold in English language: reduces the likelihood of enrolling in a higher-level qualification by nine percentage points; increases the probability of dropping out of education by age eighteen by four percentage points, and raises the likelihood of being not in education, training or employment by two percentage points. The paper provides suggestive evidence on the potential mechanism; failing to achieve the C grade appears to limit the scope of courses and institutions that individuals can subsequently attend.

4.4 Data

This chapter uses an extract from the DfE’s National Pupil Database (NPD). The NPD is a collection of linked administrative datasets providing detailed information on England’s state schools and their pupils. The School Census links pupils to the school they attend at a given day once per school term. It additionally contains rich demographic information such as gender, ethnicity, first language, as well as month and year of birth. Proxy variables including current and historical eligibility for FSM capture socioeconomic circumstances. Researchers can directly match the School Census data to pupils' attainment records for each key stage.

The extract contains data on the full cohort of pupils who completed KS4 in a state-funded school in England during the 2013/14 academic year. Most of these pupils completed KS2 in the 2008/09 academic year and thus sat KS2 SATs tests in the summer term of 2009. I only include pupils who sat this series of SATs tests in the analysis. Out of the 562,570 pupils recorded as completing secondary school in 2013/14, 535,896 (95.25 per cent) were registered to complete both the English and maths KS2 SATs in 2009/09.

The analysis does not use data on earlier cohorts since before the summer 2009 test season the testing authority practised 'borderlining'. Borderlining is the policy of re-marking all test papers for a subject in which the pupil fell three marks below the level boundaries. Remarking did not take place outside of this range. The immediate consequence of this practice is that the frequency of test scores dips three marks before the threshold and rises straight after. More importantly, however, borderlining effectively disturbs the discrete nature of the grade assignment mechanism. Markers who are remarking papers because of borderlining are likely to internalise that a pupil has narrowly fallen short of a threshold and thus might be inclined to be more generous in their marking to ensure that the pupil achieves the higher grade. Indeed, when the testing authority discontinued the borderlining procedure, the government attributed the fall in the rate of pupils achieving level 4 to this effect (DCSF, 2009).

I use total marks scored on the set of KS2 SATs English tests and math tests to calculate the distance in marks to the mark threshold to be awarded level 4 and level 5. Therefore, these variables are negative if the pupil failed to achieve the level 4 (5) threshold, zero if the pupil exactly achieved the threshold and positive if the pupil surpassed the threshold. These four variables are the set of running variables used in the regression discontinuity design. Zero corresponds to the treatment assignment cut-off for each running variable.

The outcome variables include measures of school attainment and attendance after the KS2 SATs tests. Attainment is measured by achievement in secondary school qualifications (GCSEs and equivalents). The NPD records the letter grades that pupils achieve in a range of qualifications (including both English and maths). GCSEs are graded from A* to G which are mapped to a point score from 58 to 16 with one GCSE grade worth six points. The DfE maps grades achieved in other qualifications onto a point score on the same scale. The attainment outcome variables are the point score in GCSE English and maths (separately) and the point score total from the pupils' best eight GCSE or equivalent qualifications. I standardise these variables such that they have a zero mean and a standard deviation of one.

The attendance outcome variable is the number of school sessions that a pupil missed during the seventh grade (the grade immediately after the KS2 SATs tests). There are two sessions per school day, and the legal minimum length of the school year is 190 school days (380 sessions). The pupils in the sample missed on average 15.9 sessions throughout the seventh grade.

4.5 Methodology

This chapter aims to estimate the causal effects of achieving level 4 and level 5 in KS2 SATs tests in English and maths on subsequent GCSE attainment and school attendance. For simplicity, this section frames the discussion of the econometric approach in the context of estimating the causal effect of achieving level 4 in the KS2 SATs maths test. The methodology, as well as the underpinning institutional features, are common to all subjects and levels investigated.

Consider the following potential outcomes formulation. Treatment is being awarded level 4 or above in the KS2 SATs maths tests. D_i is a binary indicator denoting treatment status, it is valued at 1 if the pupil is treated and 0 otherwise. Let Y_{i0} and Y_{i1} denote the exhaustive set of potential outcomes for pupil i: Y_{i0} is the outcome when the pupil is awarded level 2 or 3 for maths and Y_{i1} is the outcome given that the pupil is awarded level 4 or 5 in maths. Neither Y_{i1} or Y_{i0} can be simultaneously observed meaning the treatment effect for pupil i $(Y_{i1} - Y_{i0})$ is also unobserved. Instead $Y_i = D_i Y_{i1} + (1 - D_i) Y_{i0}$ is observed.

Assignment to treatment is a deterministic function of an observable running variable (X_i), which in this case is the distance between the pupil’s total KS2 maths SATs test score and the minimum score required to receive level 4 or above. A pupil is in the treatment group $(D_i = 1)$ if this distance is greater than or equal to the cut-off value (c), which is set at zero for all pupils. Pupils are untreated $(D_i = 0)$ if the distance variable is less than the cut-off value. The treatment assignment rule is perfectly binding: $D_i = \mathbb{1}(X_i \geq c)$.

Given the sharp discontinuity in treatment assignment at c under the assumptions outlined below then the local average treatment effect at cut-off $\beta = \mathbb{E}[Y_{i1} - Y_{i0}|X_i = c]$ can be identified.

$$\mathbb{E}[Y_i|X_i = c] - \lim_{\varepsilon \to 0^+} \mathbb{E}[Y_i|X_i = c - \varepsilon] \tag{4.1}$$

$$= \mathbb{E}[Y_{i1}|X_i = c] - \lim_{\varepsilon \to 0^+} \mathbb{E}[Y_{i0}|X_i = c - \varepsilon] \tag{4.2}$$

$$= \mathbb{E}[Y_{i1} - Y_{i0}|X_i = c] \tag{4.3}$$

The intuition is that the first term on the left-hand side of (4.2) is observed in the data when $X_i \geq c$ and the second term of (4.2) is observed when $X_i < c$. Therefore given that ε is sufficiently small $\mathbb{E}[Y_{i0}|X_i = c - \varepsilon]$ is an appropriate approximation for the unobservable $\mathbb{E}[Y_{i0}|X_i = c]$.

Three assumptions underpin this interpretation. The first is the unconfoundedness or ignorability assumption: $Y_{i0}, Y_{i1} \perp D_i|X_i$. This condition is trivially satisfied since there is no variation in D_i conditional on X_i. However, this does not mean that potential outcomes are guaranteed to be independent of treatment assignment. It is necessary to assume that pupils lack precise control over their aggregate maths test score. If pupils could precisely control the running variable, then pupils would self-select to either side of the cut-off value. Pupils who set their value of the running variable below the cut-off are likely to be systematically different to those who set their running variable value above the cut-off. Comparisons of outcomes for pupils either side of the cut-off would confound these systematic differences, meaning that the treatment effect is not identified by $\mathbb{E}[Y_{i1} - Y_{i0}|X_i = c]$.

Alternatively, suppose that pupils have imprecise control over the running variable. Correctly answering questions is not an unambiguously trivial exercise. Among those with the ability, work ethic and other characteristics required to score near to the cut-off, it comes down to idiosyncratic variation (or ‘chance’) as to whether they score above or below the cut-off. Pupils know that they need to score more marks to achieve level 4, but they do not know how many marks they need to score nor how the examiner will interpret their answers. This ambiguity is particularly true for the KS2 English SATs tests. The long-form writing task consists of one writing exercise and is worth 31 per cent of the available English marks. There is considerable scope for variation in the marks awarded for this task between different markers. Given that pupils have imprecise control over their aggregate maths test score (X_i) then level 4 (treatment) is “as good as” randomly assigned around the cut-off (Lee, 2008; Lee and Lemieux, 2010). However, I do not invoke the “as good as” random argument to claim identification.

It is also necessary to assume that test markers are unable to exercise precise control over the running variable. Otherwise, if markers have a bias towards pupils with specific characteristics, then it would be possible that there are systematic differences in pupils either side of the cut-off. This assumption is not particularly onerous due to the features of the marking process. Firstly, the running variable is a function of the marks scored on three different maths tests marked by at least three different markers. Secondly, the only information markers have about pupils is their name and the name of their school. Thirdly, markers are subject to a moderation process.

Schools can also request a review of pupils’ test scripts to ensure the original application of the mark scheme is appropriate and that no clerical errors have been made. Schools are more likely to request reviews when a pupil narrowly falls short of a grade threshold. As schools select into requesting reviews, grade feedback could become endogenous.

In 2009, less than 0.2 per cent of maths test scripts were reviewed. 52 per cent of these reviews resulted in the pupil receiving a different level. Less than 0.5 per cent of pupils received a different level following a successful review of their English test scripts (Qualifications and Curriculum Development Agency, 2009). While it may appear that reviews are rare, and thus of limited concern, if these reviews are concentrated around level 4 and 5 grade thresholds then they represent a non-ignorable proportion of scripts in the region of grade thresholds.

Figures 4.1 and 4.2 are histograms of the distributions of the 2009 English and maths KS2 SATs test marks. The vertical lines denote the thresholds for the level 4 and level 5 grades. These histograms clearly indicate bunching to the right of the thresholds. As there is no other plausible explanation for the bunching of the distribution, these histograms indicate that a sizable number of pupils around the grade thresholds have benefited from successful reviews. This bunching does not necessarily invalidate the research design. However, we must further assume that schools' decision to review scripts is exogenous. I argue this is likely since requesting a review is low cost: £6.50 per script (refundable if the review is successful). This is likely to be affordable for even the most under resourced schools. Explicitly, it is necessary to assume that the decision to review a script is orthogonal to the determinants of the dependent variable.

The second assumption is that the conditional expectation of the potential outcomes, $\mathbb{E}[Y_{0i}|X_i]$ and $\mathbb{E}[Y_{1i}|X_i]$ are continuous at c. Without this assumption, it is impossible to distinguish between the local average treatment effect and any discontinuity in $\mathbb{E}[Y_{0i}|X_i]$ and $\mathbb{E}[Y_{1i}|X_i]$ at the cut-off. In other words, I assume that the relationships between the outcome variable and its determinants are continuous at c.

The third assumption is that ε is sufficiently small such that $\mathbb{E}[Y_{i0}|X_i = c - \varepsilon]$ approximates $\mathbb{E}[Y_{i0}|X_i = c]$ well. Since the support of X_i is discrete rather than continuous, the limit of ε is one and not zero as presented above. That is, the closest approximation of $\mathbb{E}[Y_{i0}|X_i = c]$ is $\mathbb{E}[Y_{i0}|X_i = c - 1]$ which will overstate the true $\mathbb{E}[Y_{i1} - Y_{i0}|X_i = c]$. However, identification is achievable assuming that

$$\mathbb{E}\left[Y_{ij}\middle|X_{ij} = x_{ij}\right] = \beta D_{ij} + f\left(x_{ij}\right) \tag{4.4}$$

where j denotes possible values of X_i, $f(\cdot)$ is a continuous function, D_{ij}= $\mathbb{1}\left(X_{ij} \geq c\right)$ and $f(c) = \mathbb{E}[Y_{i0}|X_i = c]$. This specification can be expressed as a regression model suitable for estimation on pupil-level data

$$Y_i = \alpha + f(x_i) + \beta D_i + \epsilon_i \tag{4.5}$$

where Y_i denotes pupil i's GCSE maths attainment, x_i is the distance between KS2 SATs maths test mark and the threshold for the level 4 grade, and ϵ_i is the error term.

OLS produces consistent estimates of equation (4.5) assuming that $f(\cdot)$ is correctly specified. Following Lee and Card (2008), clustering standard errors by X_i became common practice in the empirical literature. However, Kolesár and Rothe (2018) demonstrate that standard errors clustered by X_i lead to confidence intervals with substantially worse coverage properties relative to conventional heteroscedasticity robust standard errors. As such, this chapter departs from the recent conventional wisdom and clusters standard errors at the primary school level.

The analysis relies on an OLS estimator despite the popularisation of non-parametric methods for the estimation of β by Hahn, Todd, and Klaauw (2001). Non-parametric methods do not require the functional form assumptions implicit in parametric estimators. This is particularly advantageous for regression discontinuity designs since functional form mis-specification can generate bias in the β estimator. Local linear regression estimators are preferred since they have been shown to have particularly good properties at the boundary of their support within the class of non-parametric estimators (Fan, 1992).

However, given the discrete nature of the running variable, the asymptotic assumptions which justify non-parametric estimation will never hold (Lee and Card, 2008). Irrespective of the sample size, no observations of the running variable will fall within a small neighbourhood below the cut-off. In the limit, a one-sided non-parametric estimator will only place weight on observations satisfying $c - 1 < X_i < c$; but X_i will not take on any value in this range.

The choice of the functional form for $f(\cdot)$ is crucial. $f(\cdot)$ is modelled as a quadratic function. A hypothesis test based on a goodness-of-fit statistic calculated using the residual sum of squares form a cubic function (restricted model) and an unrestricted model of a full set of dummy variables for each value of X indicates that a cubic functional form is too restrictive.

The choice of bandwidth entails a variance bias trade-off. Variance reduces with a larger sample size (and thus bandwidth), but a larger bandwidth means $f(\cdot)$ must fit a larger and more diverse set of observations, increasing the likelihood of misspecification. I adopt a bandwidth of ± 15 marks around the threshold and show that the effect estimates are insensitive to smaller bandwidth choices (± 5 marks and ± 10 marks).

An attractive quality of the data extract is the large sample size, which permits investigation of heterogeneous treatment effects. This chapter investigates whether the effect of ‘just’ passing the mark cut-off to achieve level 4 in maths varies according to the pupil’s history of eligibility for free school meals (over the last six years), their gender and ethnicity. The heterogeneous analysis consists of splitting the sample into groups based on the characteristic of interest and estimating equation (4.5) on each subsample.

4.6 Results

Table 4.3 contains point estimates of the average treatment effect of marginally passing the raw mark thresholds for the level 4 and level 5 grades relative to narrowly falling short of the respective thresholds. These effects are referred to as the ‘level 4 (5) grade feedback effect’ throughout this section. The level 4 grade represents test performance at the expected standard and level 5 denotes performance above this standard. Column (1) features estimates of the English test level 4 grade feedback effect; Column (2) shows the English test level 5 grade feedback effect. Column (3) and Column (4) respectively show the maths test grade feedback effects for level 4 and level 5. Tables 4.4 to 4.10 also follow this structure. The dependent variable in the first row is the standardised point score in either GCSE English or maths. In the second row, the dependent variable is the standardised point score of a pupil’s best eight GCSEs or equivalent qualifications (henceforth referred to as ‘best eight point score’). Absence, as measured by the number of sessions missed in the seventh grade, is the dependent variable in the final row. Each point estimate is from a different regression model.

The estimate of the English level 4 grade feedback effect on GCSE English point score is an improvement of 0.01σ (henceforth σ denotes the sample standard deviation of the dependent variable). This point estimate is not statistically different from zero at conventional significance levels, nor is the estimate of the corresponding effect on best eight point score; which is similar in magnitude (0.012σ). The maths level 4 grade feedback effect estimates are more consequential. The estimated effects on GCSE maths point score and best eight point score are improvements of 0.036σ and 0.032σ respectively. These point estimates respectively are statistically significant at the one per cent and five per cent levels. The point estimates of the English level 5 grade feedback effect are larger than the estimates of the maths level 5 grade feedback effect. The point estimate of the effect of narrowly passing the level 5 threshold on the English test on GCSE English point score is 0.02σ; twice the magnitude of the corresponding impact for maths.

The grade feedback effect estimates on overall absence in the seventh-grade range from -0.222 to 0.057. These estimates are equivalent to missing one-tenth less of a school day to missing one-twentieth more school days. The standard errors are large relative to the point estimates meaning that none of the estimates are statistically non-zero. Thus, the table presents no evidence of grade feedback effects on school attendance for the average pupil. On the other hand, the table indicates that there is a moderate maths level 4 grade feedback effect on subsequent attainment, and a slightly smaller English level 5 grade feedback effect; yet there is no evidence of an English level 4 or maths level 5 grade feedback effect.

These small local average treatment effect estimates may mask substantial variation in the local average treatment effects for subgroups of the pupil population. Table 4.4 shows grade feedback effect estimates for pupils with a history of FSM eligibility (Panel A), and those who have no FSM entitlement history (Panel B). The table indicates there are noticeable differences in the grade feedback effects between these groups. For illustrative purposes, Figures 4.3-4.6 graph the regression model used to estimate the grade feedback effects for FSM eligible pupils on GCSE English and maths point score.

The estimate of the English level 4 grade feedback effect on GCSE English point score is 0.029σ and is 0.037σ on best eight point score for FSM eligible pupils. For FSM ineligible pupils, the corresponding estimates respectively are 0.002σ and -0.001σ, and neither is statistically different from zero. The estimated English level 5 grade feedback effects for FSM eligible pupils are greater: 0.066σ on GCSE English point score, and 0.045σ on best eight point score. Again, the corresponding point estimates for pupils without a history of FSM eligibility are statistically zero. Thus, English grade feedback effects on subsequent attainment are consequential for the most disadvantaged pupils, and seemingly non-existent for other pupils.

The distinction between the maths test grade feedback effects for FSM eligible pupils and other pupils is less clear than the effects for English tests. The maths level 4 grade feedback effect estimates for FSM eligible pupils is sizable and positive (for example, 0.054σ for GCSE maths point score). However, there are also positive – but substantially smaller – effects for non-eligible pupils (0.025σ on GCSE maths point score and 0.027σ on best eight point score). The effect estimates for non-eligible pupils are relatively less precise, and the difference in the point estimates between the two groups is not statistically significant at the five per cent level. The pattern of larger grade feedback effect estimates on subsequent attainment for FSM eligible pupils does not hold for the maths level 5 grade feedback effect. For FSM eligible pupils, the estimated maths level 5 grade feedback effects on attainment outcomes are not statistically different from zero, whereas for non-FSM eligible pupils the maths level 5 grade feedback effect estimate on GCSE maths point score is 0.011σ (but only statistically significant at the ten per cent level). The magnitude of this estimate is small relative to the English grade feedback effects for FSM eligible pupils.

The grade feedback effect estimates on absence for FSM non-eligible pupils are close to zero (-0.099 to 0.109) and statistically zero. For pupils with a history of FSM eligibility, the point estimates are all negative (ranging from -0.209 to -0.636) indicating that fewer school sessions are missed by pupils who marginally pass a level threshold. However, none of these estimates are statistically different from zero at the ten per cent level.

The heterogeneity analysis continues in Table 4.5, which features separate grade feedback effect estimates for boys and girls. Panel A contains the estimates for boys, while Panel B contains the estimates for girls. Regarding English test levels, the grade feedback effect estimates for girls (on all dependent variables) are not statistically different from zero. However, for boys, there are modest positive English level 5 grade feedback effect estimates (0.33σ for GCSE English point score and 0.22σ for best eight point score). The English level 5 grade feedback effect on absence during the seventh grade is -0.420 for boys. These three coefficient estimates are statistically significant at least at the five per cent level. The English level 4 grade feedback effect estimates on future attainment are modest and positive for boys but are not statistically significant; for girls the corresponding estimates are precise zeros.

Switching focus to the estimates of the maths level 4 grade feedback effects, we find the most substantial effect estimates within the table. For boys, the maths level 4 grade feedback effect is 0.038σ on GCSE maths point score and 0.040σ for best eight point score. These estimates are statistically significant at the five per cent level. For girls, the corresponding estimated effect on GCSE maths point score is smaller (0.030σ) although it is not statistically different from the boys’ estimate. The estimated impact on best eight point score is statistically zero for girls. Akin to the FSM heterogeneity results, the pattern in the results flips for the maths level 5 grade feedback effect estimates. The estimated attainment maths level 5 grade feedback effect is statistically zero for boys. However, for girls, there is a small positive estimated effect.

Table 4.6 contains the final heterogeneous effects results. Panel A contains effect estimates for white pupils, whereas Panel B contains estimates for non-white pupils. For white pupils, three of the four coefficient estimates of the grade feedback effect on GCSE point score are precise and statistically non-zero. The English level 5 and maths level 4 grade feedback effects are largest (0.026σ and 0.036σ respectively), while the English level 4 and maths level 5 grade feedback effects are relatively smaller (respectively 0.016σ and 0.015σ). For white pupils, the estimated effects on best eight point score are similar to the corresponding estimates on GCSE point score in English or maths. However, the former set of estimates are less precise meaning that only the larger estimated effects (English level 5 and maths level 4) are statistically non-zero at conventional significance levels.

For non-white pupils, the point estimates for the attainment models are generally much closer to zero than the corresponding estimates for white pupils; none are statistically different from zero. Regarding the effect on absence estimates, the estimates for non-white pupils are not statistically different from zero. For white pupils, the estimated math level 5 grade feedback effect on absence is -0.239 and is the only statistically non-zero absence estimate at the ten per cent significance level.

Attention now turns towards judging the robustness of the results presented above. Table 4.7 investigates whether any discontinuities exist in predetermined pupil characteristics at the level 4 and 5 grade thresholds for both the English and maths KS2 SATs tests. The predetermined pupil characteristics are four binary variables indicating independently whether a pupil is female, has a history of FSM eligibility, is white and is a native English speaker. The fifth pupil characteristic is the standardised average point score of a pupil's KS1 (age seven) teacher assessments. This latter characteristic is the only measure of attainment available before the KS2 SATs tests. None of the coefficient estimates presented in the table are statistically different from zero at the five per cent significance level. Only two coefficient estimates (the 'effect' of marginally crossing the level 4 threshold in English on speaking English as a first language and on KS1 average teacher assessments) are statistically non-zero at the ten per cent significance level. The table fails to provide evidence of discontinuities in the observable predetermined characteristics at the level 4 and 5 grade thresholds. While this does not guarantee that there are no discontinuities in unobservable attributes at the level thresholds, it does not undermine the assumptions underpinning the identification strategy.

Table 4.8 explores the sensitivity of the results from the full sample of pupils to the inclusion of predetermined pupil characteristics in the regression models. If the coefficient estimates substantially differ between the models with and without controls, then this undermines the results presented thus far. A difference in the coefficient estimates would indicate that the estimated effect of marginally passing the level thresholds is at least partly due to differences in the observable characteristics on either side of the threshold. Adding the control variables removes any bias in the treatment effect estimate caused by differences in observable characteristics on either side of the thresholds. Each panel presents estimates from models with and without pupil characteristics. In columns (2) to (4), the difference in estimates between the models with and without controls is trivial. In Column (1) the estimated effect of just meeting the level 4 threshold for English doubles in magnitude in Panels A and B, however, this difference is not statistically significant.

Table 4.9 appraises the sensitivity of the coefficient estimates to the choice of the bandwidth used in the regression modelling. In the principal analysis, a bandwidth of ±15 marks around the level thresholds is used. The table contains results using alternative bandwidths of ±5 marks, ±10 marks and ±20 marks. Panel A features models in which the dependent variable is the GCSE point score in English and maths. Aside from column (3) in which coefficient estimates range from 0.048 to 0.027, the estimates do not meaningfully differ as the bandwidth varies. In Panel B, the dependent variable is the best eight point score. There is limited variation in the point estimates when the bandwidth is either ±5 marks, ±10 marks or ±15 marks. Finally, in Panel C, which considers absence in the seventh grade, the coefficient estimates are not statistically significant at conventional significance levels irrespective of the bandwidth in columns (1) through (4). Overall, the interpretation of the primary results would not be substantially different if they were based on one of the alternative bandwidths.

Table 4.10 considers whether discontinuities in the dependent variable exist at other points in the raw test mark distribution. The table replicates the regression models but shifts the level thresholds between two marks below and two marks above the actual thresholds. If evidence exists of other discontinuities in the region near the level thresholds, then this would severely impair the validity of the main results. In panels A and B (models for GCSE point score in English/maths and best eight point score), the point estimates of discontinuities at the false level thresholds are generally close to zero. The point estimates of the effects of marginally passing the ‘fake’ maths level 4 thresholds are modest, but like all but one coefficient estimates in panels A and B, they are not statistically different from zero at the ten per cent significance level. In Panel C, only one coefficient estimate is statistically different from zero at the ten per cent level. Thus, the table suggests that discontinuities in the outcome variables are not present at arbitrary points in the neighbourhood of the level thresholds; this further strengthens the causal interpretation given to the results discussed earlier.

4.7 Discussion

Grade based feedback from an ostensibly low-stakes series of standardised tests taken at age eleven has a polarising effect on the secondary school qualification grades of otherwise similar pupils. The polarisation appears to be stronger for economically disadvantaged pupils, and – in the case of English grades – non-existent for other pupils.

The core function of the KS2 SATs tests is to measure primary school performance. The KS2 SATs tests do not provide the basis for the awarding of qualifications. As such, it is undesirable that the coarse feedback pupils receive from their participation impacts their subsequent progress. Pupils need not receive any feedback for the tests to serve their use as a primary school performance metric.

This chapter, therefore, highlights that when test-based school accountability systems are being designed, it is essential to consider the nature of feedback provided to participating pupils. It also suggests that it is necessary to understand how the feedback might interact with the institutional features of the schooling system. For example, the potential for the feedback to be used to set class groups or to fix subsequent attainment targets. Since this chapter shows that feedback from the KS2 SATs tests generates differences in progress between comparable pupils, it is arguable that a review of both the feedback mechanism and the uses of the feedback would be prudent.

Such a review has taken place. Following the introduction of the new national curriculum for England in 2014, the DfE reformed the SATs test feedback mechanism for pupils in 2016. The department replaced the broad national curriculum levels studied in this chapter with scaled scores. Scaled scores represent a less coarse feedback mechanism, while still being anchored to attainment standards and suitable for comparisons across cohorts.

Instead of receiving a level ranging from two to five, pupils receive a scaled score that ranges from 80 to 120. A scaled score of 100 denotes achieving the expected standard precisely, while a score above or below 100 represents attainment above or below the expected standard respectively. This granular approach to feedback ensures that pupils who are working just below the expected standard have a sense of perspective to their apparent underachievement. Under the previous national curriculum levels mechanism, pupils are unaware of the extent of their apparent distance to the expected standard. Pupils who are very far from the expected standard no longer receive the same feedback as those very close (but not close enough) to the expected standard.

Also, in 2016, the DfE replaced expected progress as the headline measure of pupil progress at secondary school with a simple value-added metric, progress eight. A school's progress eight score is the simple average of the difference between its pupils' total point score from GCSE English and maths plus six other subjects and the national average total point score of pupils with the same KS2 SATs performance. The introduction of progress eight means that GCSE English and maths targets are no longer monotonically increasing step functions of KS2 SATs test attainment (with 'jumps' corresponding to each level threshold). As such, the pupil progress metric no longer incentivises schools to exert effort disproportionately on pupils who score 'just' above each level thresholds relative to those 'just' below each level threshold.

There are other potential reforms of the feedback mechanism. Alcott (2017) characterises them on a spectrum ranging from extreme measures to mild. The extreme reform would be to discontinue KS2 SATs test in their entirety, and by implication discontinuing feedback from the tests. However, if school-level KS2 SATs test performance is a useful indicator of school quality for parents and regulators, then this information would be lost.

Alternatively, if the mechanism linking grade feedback to subsequent attainment were autogenous to pupils, then it would be wise not to inform pupils of their test performance. For example, if grade feedback affects pupils' motivation and effort and this is the cause of the polarisation in subsequent attainment, then pupil motivation will be uninfluenced by their performance substantially if they are unaware of their results. If grade feedback affects effort and motivation, then one might expect to see an effect of grade feedback on school attendance. However, there is no evidence of substantial effects on school attendance.

Supposing that the mechanism is institutional, such as through class or target setting, then a wise suggestion would be to anonymise feedback from the tests fully. Teachers and school leaders would not be able to use the test feedback for class or target setting, but they would be able to use alternative, and possibly better suited, data for these purposes.

A moderate approach is to provide feedback on a continuous scale, as is now the case in England. This might complicate the interpretation of scores for both parents and pupils, but the benefits of moving away from broad grades are likely to outweigh this cost (Alcott, 2017). The mildest reform would be to strive to moderate the interpretation of test feedback by parents and pupils. If the source of the grade feedback effect is parents or pupils, then policymakers should encourage parents and pupils to view schooling attainment as malleable, and not consider test feedback as a decisive and irreversible determination of ability.

4.8 Conclusion

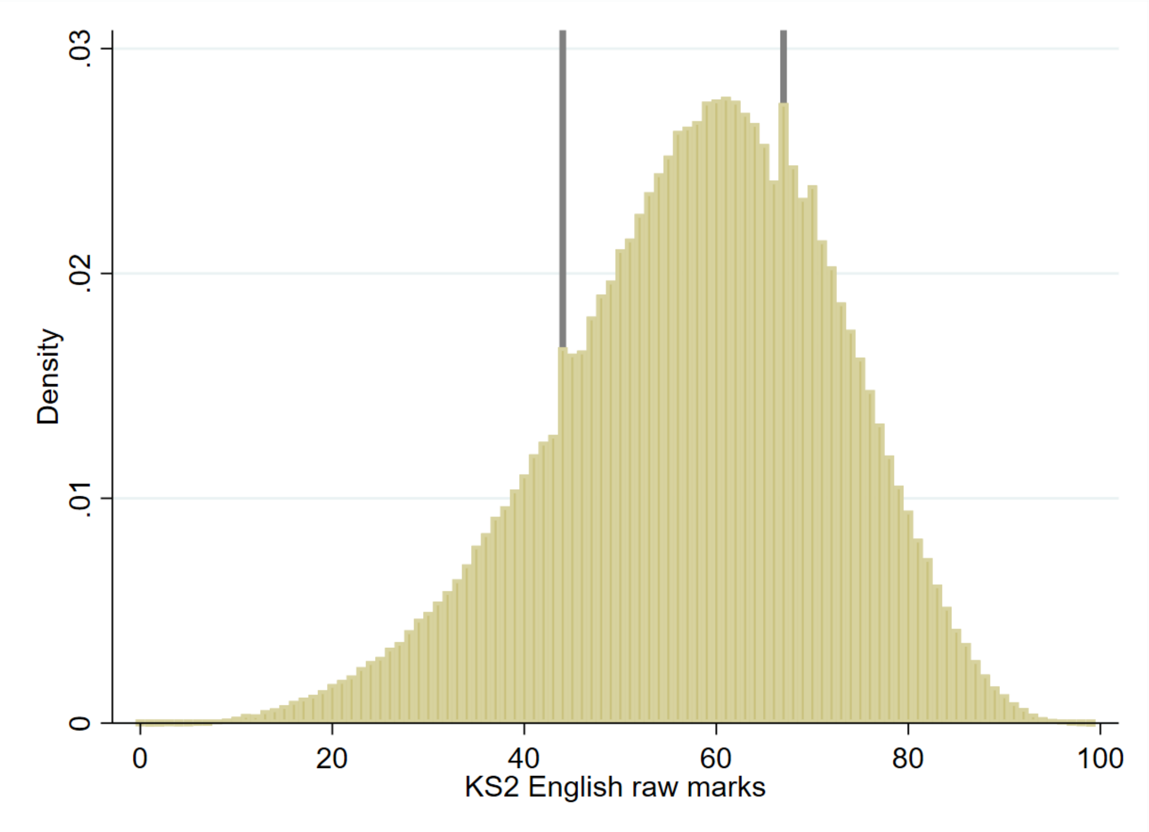
This chapter attempts to quantify the causal effect of marginally passing grade thresholds on standardised tests, attempted at age eleven, on subsequent school attendance as well as attainment at the end of compulsory schooling. To this end, the analysis exploits the assignment of grades to pupils on the sole basis of the raw test scores achieved on the standardised tests with the availability of a rich administrative dataset to estimate a series of models based on a sharp regression discontinuity design (RDD) identification strategy.

The chapter finds evidence that the grades awarded on the supposedly low stakes standardised tests affect future secondary school qualification grades, but no evidence of an impact on attendance during the school year after the pupils sit the test. For the average pupil, just meeting the raw mark threshold for the grade denoting achievement above the expected standard in English at age eleven increases English grades by two per cent of a standard deviation five years later at the end of schooling. In maths, the average effect of narrowly passing the cut-off for grade denoting meeting the expected standard improves the end of school maths grade by 3.6 per cent of a standard deviation.

However, these small average treatment effects mask substantial heterogeneity. English language standardised test grades do not influence subsequent English attainment for pupils who have never had an entitlement to free school meals. On the other hand, ‘just’ achieving the expected standard in English boosts free school meal eligible pupils’ end of schooling English grades by three per cent of a standard deviation. While marginally reaching the raw mark threshold for the above-expected standard grade boosts subsequent English grades by 6.6 per cent of a standard deviation for the same group.

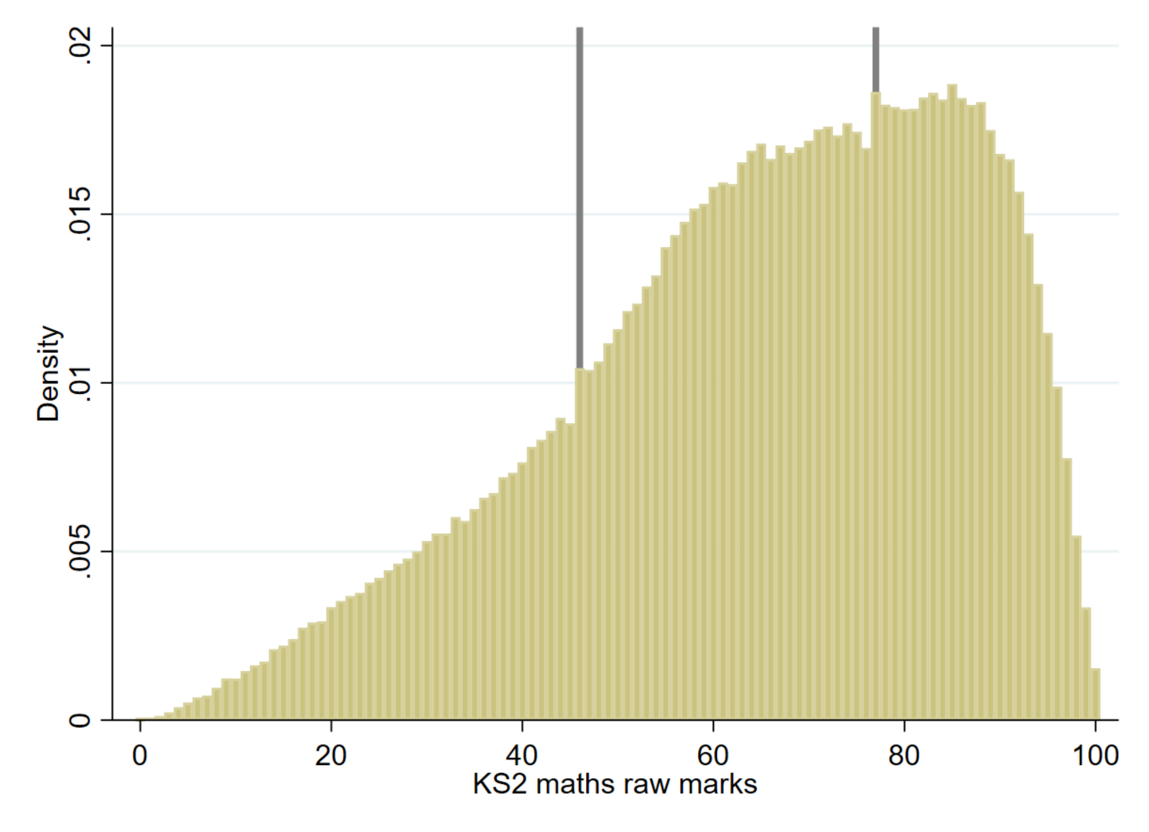
The findings indicate that it would be wise of policymakers to consider the nature of the feedback provided by standardised tests as well as the usage of said feedback when designing a test-based accountability system. Policymakers in England designed SATs tests to be low stakes from the pupil perspective, yet this chapter demonstrates how SATs test grades affect performance on terminal school exams independent of pupils' underlying ability. In addition, the findings of this chapter raise new questions about the differences in school experience between disadvantaged and advantaged pupils. It is undoubtedly troubling that pupils who narrowly miss achieving a grade due to bad luck are penalised more if they are socioeconomically disadvantaged.

Figure 4.1 Histogram of 2009 KS2 English SAT test raw marks



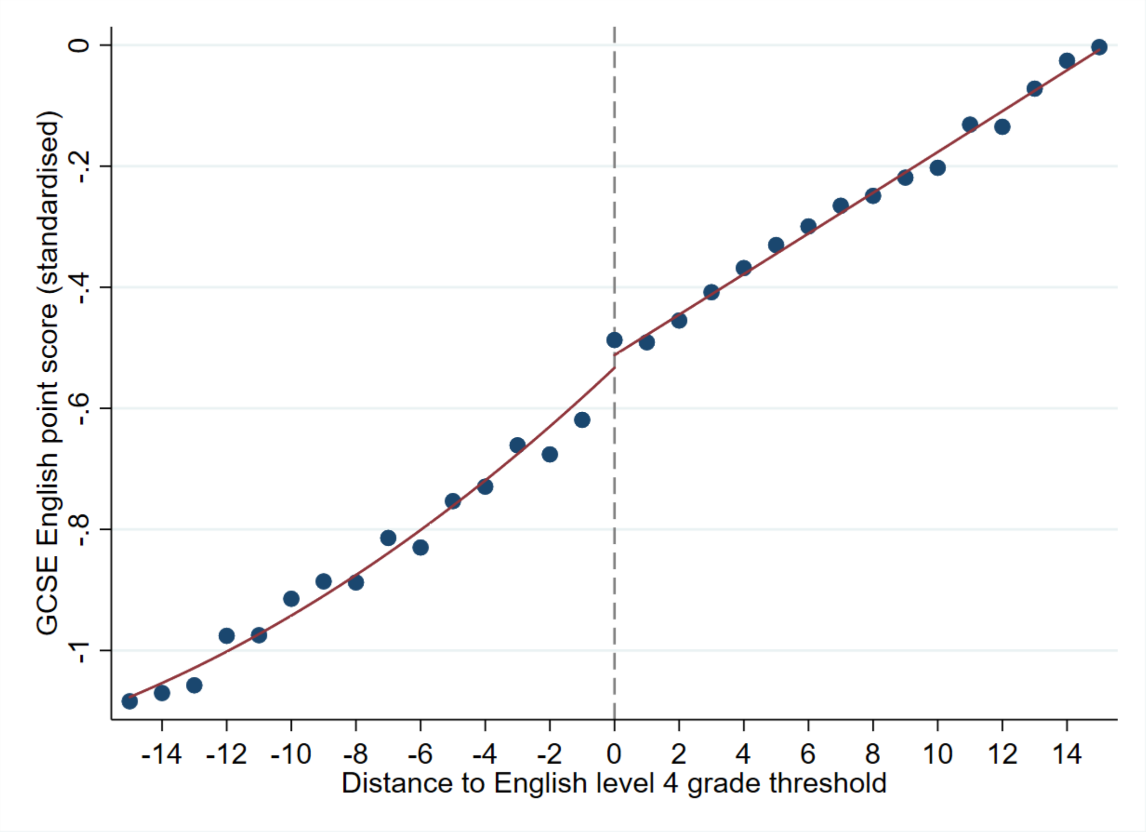
Notes: the vertical grey bars (at 44 and 67 marks) denote the level 4 and level 5 grade thresholds respectively.

Figure 4.2 Histogram of 2009 KS2 maths SAT test raw marks



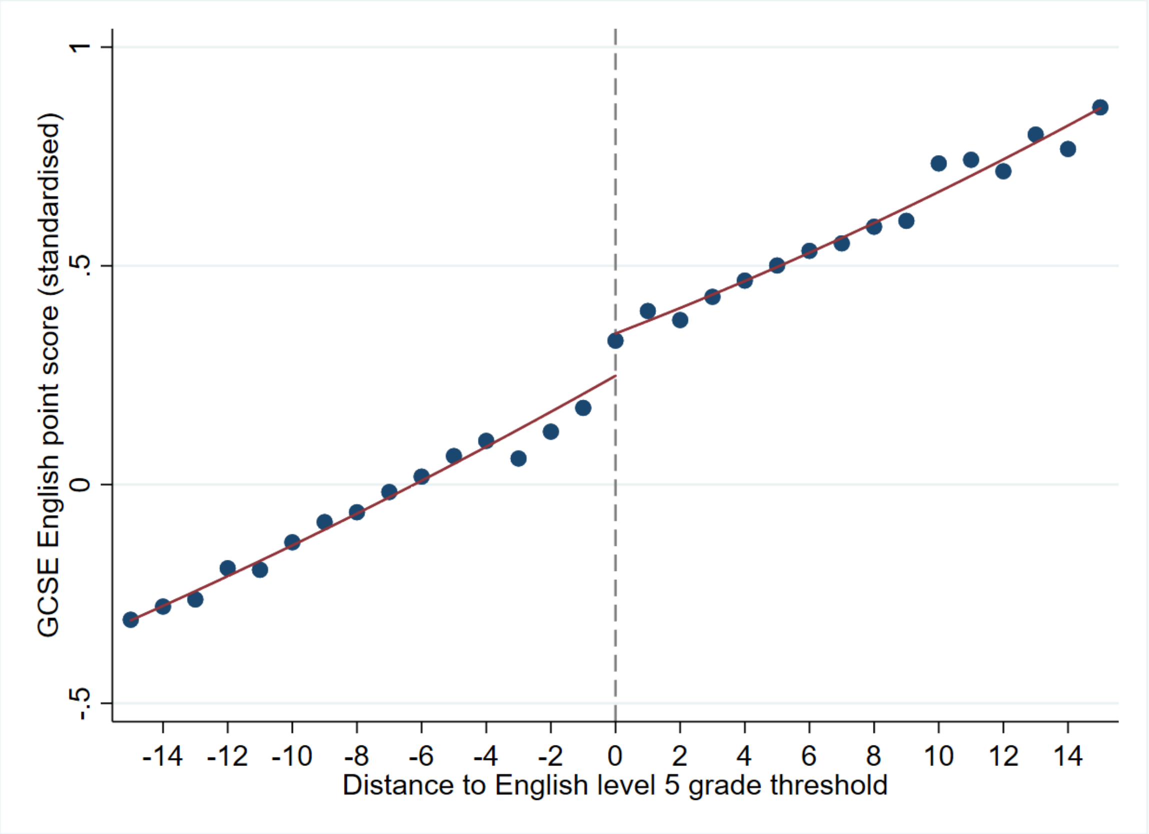
Notes: the vertical grey bars (at 46 and 77 marks) denote the level 4 and level 5 grade thresholds respectively.

Figure 4.3 Effect of marginally passing the English level 4 grade threshold on GCSE English point score (standardised) for FSM eligible pupils



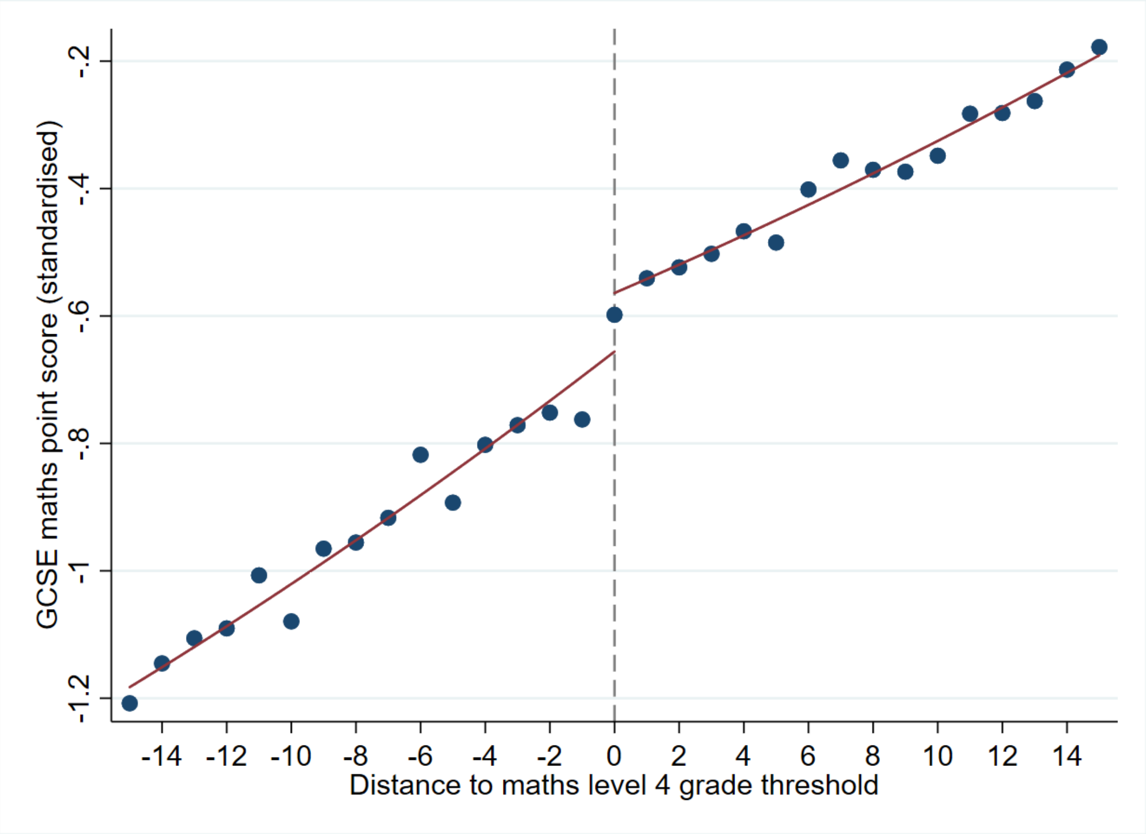
Notes: filled circles represent cell means of GCSE English point score (standardised). See Panel A of Table 4.4 for details of the estimated discontinuity.

Figure 4.4 Effect of marginally passing the English level 5 grade threshold on GCSE English point score (standardised) for FSM eligible pupils



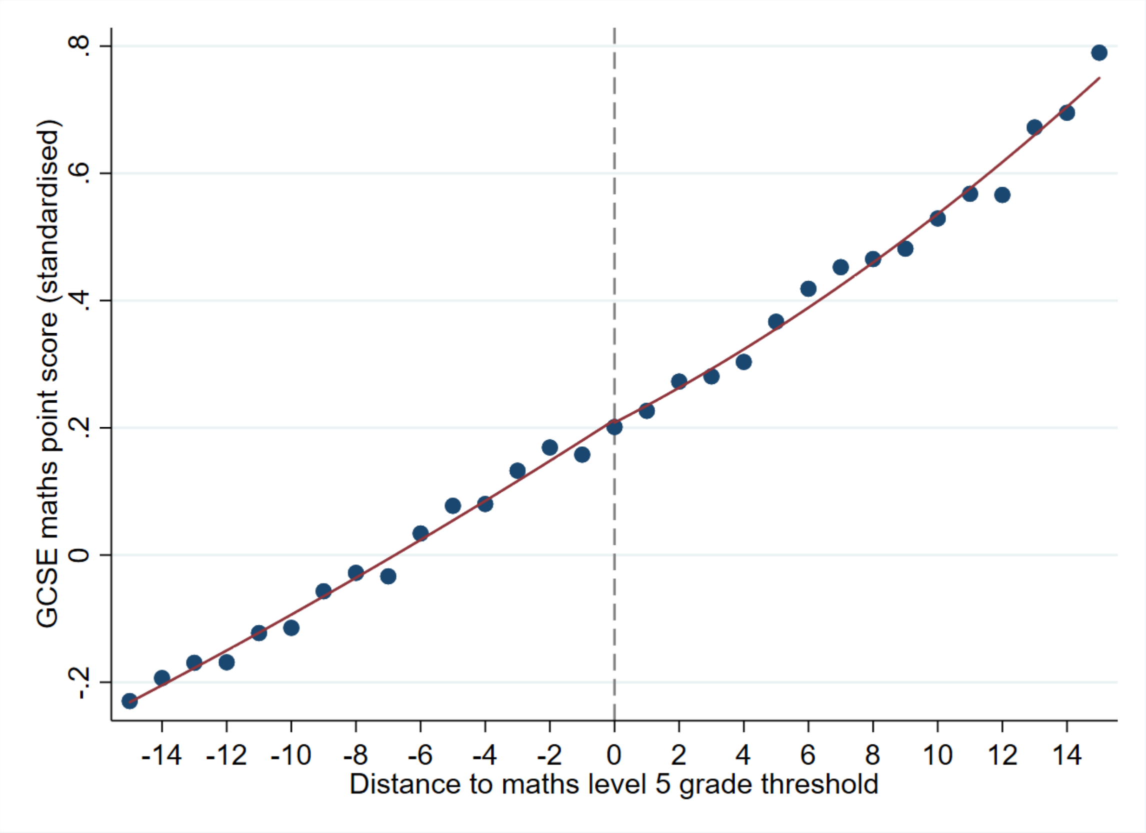
Notes: filled circles represent cell means of GCSE English point score (standardised). See Panel A of Table 4.4 for details of the estimated discontinuity.

Figure 4.5 Effect of marginally passing the maths level 4 grade threshold on GCSE maths point score (standardised) for FSM eligible pupils



Notes: filled circles represent cell means of GCSE maths point score (standardised). See Panel A of Table 4.4 for details of the estimated discontinuity.

Figure 4.6 Effect of marginally passing the maths level 5 grade threshold on GCSE maths point score (standardised) for FSM eligible pupils



Notes: filled circles represent cell means of GCSE maths point score (standardised). See Panel A of Table 4.4 for details of the estimated discontinuity.

Table 4.1: Distribution of national curriculum levels awarded for KS2 tests

	(1)	(2)
	Number of students	Percentage (%)
Panel A: English test levels		
Below the level assessed by the test	15,655	2.92
Not awarded a level	4,255	0.79
Level 2	3,770	0.70
Level 3	82,240	15.35
Level 4	240,468	44.87
Level 5	183,239	34.19
Absent/Maladministration/etc.	4,001	0.75
Panel B: Maths test levels		
Below the level assessed by the test	18,679	3.49
Not awarded a level	6,675	1.25
Level 2	2,925	0.55
Level 3	75,321	14.05
Level 4	276,801	51.65
Level 5	153,921	28.72
Absent/Maladministration/etc.	3,834	0.72

Notes: Based on the NPD extract of all pupils completing KS4 in 2013/14 and KS2 in 2008/09.

Table 4.2: Percentage of pupils awarded each national curriculum level by FSM history, gender and ethnicity

	(1)	(2)	(3)	(4)	(5)	(6)
	FSM history		Gender		Ethnicity	
	No FSM (%)	FSM (%)	Female (%)	Male (%)	White (%)	Non-white (%)
Panel A: English test levels						
Below the level assessed by the test	2.30	6.68	2.35	4.57	3.33	4.17
Not awarded a level	0.48	1.63	0.48	1.10	0.78	0.87
Level 2	0.35	1.07	0.37	0.71	0.52	0.64
Level 3	10.99	22.32	10.83	17.14	13.59	16.04
Level 4	51.61	51.76	50.17	53.08	51.47	52.45
Level 5	33.66	15.38	35.16	22.54	29.54	25.20
Absent/Maladministration/etc.	0.61	1.16	0.64	0.86	0.77	0.63
Panel B: Maths test levels						
Below the level assessed by the test	1.95	5.56	2.34	3.47	2.77	3.57
Not awarded a level	0.82	2.40	1.33	1.16	1.18	1.52
Level 2	0.50	1.26	0.74	0.67	0.67	0.86
Level 3	12.62	22.72	16.47	14.27	14.96	17.01
Level 4	44.12	46.90	47.21	42.63	45.00	44.32
Level 5	39.44	20.22	31.24	37.02	34.69	32.05
Absent/Maladministration/etc.	0.55	0.94	0.67	0.78	0.73	0.67

Notes: Based on the NPD extract of all pupils completing KS4 in 2013/14 and KS2 in 2008/09.

Table 4.3: Estimated effects of achieving level four or five in KS2 SATs English and maths tests

	(1)	(2)	(3)	(4)
	KS2 SATs English		KS2 SATs maths	
	Achieved level four	Achieved level five	Achieved level four	Achieved level five
GCSE point score in subject (standardised)	0.010	0.020***	0.036***	0.010*
	(0.011)	(0.006)	(0.012)	(0.006)
Best 8 GCSE point score total (standardised)	0.012	0.013**	0.032**	0.007
	(0.011)	(0.006)	(0.014)	(0.008)
Absence	0.057	-0.222	-0.211	-0.172
	(0.257)	(0.140)	(0.299)	(0.172)

Notes: all estimates are from separate regressions. A bandwidth of ±15 marks around the level threshold is used. Robust standard errors clustered at primary school level in parentheses. ***, ** and * denote statistical significance at the 1 per cent, 5 per cent and 10 per cent levels respectively.

Table 4.4: Estimated effects of achieving level four or five in KS2 SATs English and maths tests by FSM eligibility history

	(1)	(2)	(3)	(4)
	KS2 SATs English		KS2 SATs maths	
	Achieved level four	Achieved level five	Achieved level four	Achieved level five
Panel A: History of FSM eligibility				
GCSE point score in subject (standardised)	0.029***	0.066***	0.054**	0.002
	(0.011)	(0.018)	(0.023)	(0.016)
Best 8 GCSE point score total (standardised)	0.037**	0.045***	0.039	-0.001
	(0.018)	(0.017)	(0.026)	(0.020)
Absence	-0.209	-0.636	-0.558	-0.395
	(0.489)	(0.422)	(0.617)	(0.487)
Panel B: No history of FSM eligibility				
GCSE point score in subject (standardised)	0.002	0.007	0.025*	0.011*
	(0.012)	(0.007)	(0.014)	(0.006)
Best 8 GCSE point score total (standardised)	-0.001	0.003	0.027*	0.009
	(0.013)	(0.006)	(0.015)	(0.008)
Absence	0.109	-0.099	0.002	-0.098
	(0.269)	(0.136)	(0.304)	(0.169)

Notes: all estimates are from separate regressions. A bandwidth of ±15 marks around the level threshold is used. Robust standard errors clustered at primary school level in parentheses. ***, ** and * denote statistical significance at the 1 per cent, 5 per cent and 10 per cent levels respectively.

Table 4.5: Estimated effects of achieving level four or five in KS2 SATs English and maths tests by gender

	(1)	(2)	(3)	(4)
	KS2 SATs English		KS2 SATs maths	
	Achieved level four	Achieved level five	Achieved level four	Achieved level five
Panel A: Male				
GCSE point score in subject (standardised)	0.017	0.033***	0.038**	0.004
	(0.015)	(0.010)	(0.019)	(0.008)
Best 8 GCSE point score total (standardised)	0.022	0.022**	0.040**	-0.007
	(0.015)	(0.009)	(0.020)	(0.011)
Absence	-0.121	-0.420**	-0.274	-0.070
	(0.326)	(0.206)	(0.448)	(0.249)
Panel B: Female				
GCSE point score in subject (standardised)	0.002	0.009	0.030*	0.016*
	(0.017)	(0.008)	(0.016)	(0.008)
Best 8 GCSE point score total (standardised)	-0.004	0.007	0.015	0.022**
	(0.017)	(0.008)	(0.017)	(0.010)
Absence	0.339	-0.062	-0.139	-0.280
	(0.415)	(0.193)	(0.397)	(0.236)

Notes: all estimates are from separate regressions. A bandwidth of ±15 marks around the level threshold is used. Robust standard errors clustered at primary school level in parentheses. ***, ** and * denote statistical significance at the 1 per cent, 5 per cent and 10 per cent levels respectively.

Table 4.6: Estimated effects of achieving level four or five in KS2 SATs English and maths tests by ethnicity

	(1)	(2)	(3)	(4)
	KS2 SATs English		KS2 SATs maths	
	Achieved level four	Achieved level five	Achieved level four	Achieved level five
Panel A: White				
GCSE point score in subject (standardised)	0.016	0.026***	0.036***	0.015**
	(0.013)	(0.007)	(0.014)	(0.006)
Best 8 GCSE point score total (standardised)	0.012	0.016**	0.036**	0.010
	(0.012)	(0.007)	(0.015)	(0.008)
Absence	-0.079	-0.257	-0.164	-0.239*
	(0.300)	(0.157)	(0.340)	(0.192)
Panel B: Non-white				
GCSE point score in subject (standardised)	-0.003	-0.007	0.029	-0.019
	(0.022)	(0.015)	(0.027)	(0.014)
Best 8 GCSE point score total (standardised)	0.020	0.002	0.007	-0.006
	(0.023)	(0.014)	(0.029)	(0.017)
Absence	0.482	-0.065	-0.350	0.171
	(0.475)	(0.319)	(0.609)	(0.383)

Notes: all estimates are from separate regressions. A bandwidth of ±15 marks around the level threshold is used. Robust standard errors clustered at primary school level in parentheses. ***, ** and * denote statistical significance at the 1 per cent, 5 per cent and 10 per cent levels respectively.

Table 4.7: Tests for discontinuity in predetermined characteristics at level four and five thresholds

	(1)	(2)	(3)	(4)
	KS2 SATs test English		KS2 SATs test Maths	
	Achieved level four	Achieved level five	Achieved level four	Achieved level five
Female	-0.001	-0.000	0.011	-0.000
	(0.007)	(0.005)	(0.008)	(0.006)
History of FSM eligibility	0.006	-0.002	-0.001	-0.004
	(0.007)	(0.004)	(0.008)	(0.005)
White ethnicity	0.007	0.001	-0.005	-0.004
	(0.006)	(0.004)	(0.006)	(0.004)
English is first language	0.009*	-0.000	-0.007	0.001
	(0.005)	(0.003)	(0.005)	(0.004)
KS1 average teacher assessment (standardised)	0.018*	0.002	0.015	0.003
	(0.010)	(0.006)	(0.012)	(0.008)

Notes: all estimates are from separate regressions. A bandwidth of ±15 marks around the level threshold is used. Robust standard errors clustered at primary school level in parentheses. ***, ** and * denote statistical significance at the 1 per cent, 5 per cent and 10 per cent levels respectively.

Table 4.8: Sensitivity of estimated effects to inclusion of control variables

	(1)	(2)	(3)	(4)
	KS2 SATs English		KS2 SATs maths	
	Achieved level four	Achieved level five	Achieved level four	Achieved level five
Panel A: GCSE point score in subject (standardised)				
Estimates without controls	0.010	0.020***	0.036***	0.010*
	(0.011)	(0.006)	(0.012)	(0.006)
Estimates with controls	0.020*	0.019***	0.034***	0.014**
	(0.011)	(0.006)	(0.012)	(0.006)
Panel B: Best 8 GCSE point score total (standardised)				
Estimates without controls	0.012	0.013**	0.032**	0.007
	(0.011)	(0.006)	(0.014)	(0.008)
Estimates with controls	0.021**	0.012**	0.031**	0.011
	(0.010)	(0.006)	(0.013)	(0.007)
Panel C: Absence				
Estimates without controls	0.057	-0.222	-0.211	-0.172
	(0.257)	(0.140)	(0.299)	(0.172)
Estimates with controls	-0.064	-0.204	-0.346	-0.197
	(0.262)	(0.142)	(0.312)	(0.176)

Notes: all estimates are from separate regressions. The control variables specification includes primary school fixed effects, a gender indicator, FSM history indicator, month of birth indicators, ethnicity indicators and first language indicators. A bandwidth of ±15 marks around the level threshold is used. Robust standard errors clustered at primary school level in parentheses. ***, ** and * denote statistical significance at the 1 per cent, 5 per cent and 10 per cent levels respectively.

Table 4.9: Sensitivity of estimated effects to choice of bandwidth

	(1)	(2)	(3)	(4)
	KS2 SATs English		KS2 SATs maths	
	Achieved level four	Achieved level five	Achieved level four	Achieved level five
Panel A: GCSE point score in subject (standardised)				
Bandwidth: ±5 marks	0.010	0.021*	0.048	0.010
	(0.024)	(0.012)	(0.026)	(0.012)
Bandwidth: ±10 marks	0.009	0.025***	0.046***	0.010**
	(0.014)	(0.008)	(0.016)	(0.007)
Bandwidth: ±15 marks	0.010	0.020***	0.036***	0.010*
	(0.011)	(0.006)	(0.012)	(0.006)
Bandwidth: ±20 marks	0.011*	0.016***	0.027***	0.011**
	(0.006)	(0.006)	(0.011)	(0.005)
Panel B: Best 8 GCSE point score total (standardised)				
Bandwidth: ±5 marks	0.019	0.012	0.023	0.010
	(0.023)	(0.011)	(0.019)	(0.006)
Bandwidth: ±10 marks	0.013*	0.015**	0.029*	0.010
	(0.007)	(0.007)	(0.017)	(0.010)
Bandwidth: ±15 marks	0.012	0.013**	0.032**	0.007
	(0.011)	(0.006)	(0.014)	(0.008)
Bandwidth: ±20 marks	0.011*	0.014**	0.023**	0.015**
	(0.010)	(0.006)	(0.010)	(0.007)

Notes: all estimates are from separate regressions. Robust standard errors clustered at primary school level in parentheses. ***, ** and * denote statistical significance at the 1 per cent, 5 per cent and 10 per cent levels respectively.

Table 4.9 (continued): Sensitivity of estimated effects to choice of bandwidth

	(1)	(2)	(3)	(4)
	KS2 SATs English		KS2 SATs maths	
	Achieved level four	Achieved level five	Achieved level four	Achieved level five
Panel C: Absence				
Bandwidth: ±5 marks	0.070	-0.315	0.151	-0.579
	(0.531)	(0.297)	(0.616)	(0.364)
Bandwidth: ±10 marks	0.224	-0.211	0.057	-0.315
	(0.323)	(0.175)	(0.371)	(0.216)
Bandwidth: ±15 marks	0.057	-0.222	-0.211	-0.172
	(0.257)	(0.140)	(0.299)	(0.172)
Bandwidth: ±20 marks	0.104	-0.178	0.047	0.002
	(0.226)	(0.124)	(0.255)	(0.147)

Notes: all estimates are from separate regressions. Robust standard errors clustered at primary school level in parentheses. ***, ** and * denote statistical significance at the 1 per cent, 5 per cent and 10 per cent levels respectively.

Table 4.10: Falsification test – estimated effects at false level thresholds

	(1)	(2)	(3)	(4)
	KS2 SATs English		KS2 SATs maths	
	Achieved level four	Achieved level five	Achieved level four	Achieved level five
Panel A: GCSE point score in subject (standardised)				
Cut-off: threshold -2 marks	0.015	0.005	-0.018	0.001
	(0.015)	(0.008)	(0.016)	(0.007)
Cut-off: threshold -1 marks	0.011	0.006	0.011	-0.000
	(0.014)	(0.007)	(0.014)	(0.007)
Cut-off: threshold +1 marks	0.013	0.003	0.019	0.006
	(0.010)	(0.007)	(0.017)	(0.006)
Cut-off: threshold +2 marks	0.002	-0.006	0.015	0.006
	(0.010)	(0.007)	(0.014)	(0.006)
Panel B: Best 8 GCSE point score total (standardised)				
Cut-off: threshold -2 marks	-0.005	0.005	-0.005	-0.004
	(0.015)	(0.007)	(0.017)	(0.009)
Cut-off: threshold -1 marks	-0.004	0.008	0.025	0.009
	(0.013)	(0.007)	(0.016)	(0.009)
Cut-off: threshold +1 marks	0.014	-0.001	0.024	0.004
	(0.010)	(0.006)	(0.014)	(0.007)
Cut-off: threshold +2 marks	-0.001	0.000	0.023*	0.001
	(0.010)	(0.007)	(0.013)	(0.008)

Notes: all estimates are from separate regressions. A bandwidth of ±15 marks around the level threshold is used. Robust standard errors clustered at primary school level in parentheses. ***, ** and * denote statistical significance at the 1 per cent, 5 per cent and 10 per cent levels respectively.

Table 4.10 (continued): Falsification test – estimated effects at false level thresholds

	(1)	(2)	(3)	(4)
	KS2 SATs English		KS2 SATs maths	
	Achieved level four	Achieved level five	Achieved level four	Achieved level five
Panel C: Absence				
Cut-off: threshold -2 marks	-0.210	-0.321*	-0.169	0.263
	(0.338)	(0.166)	(0.382)	(0.209)
Cut-off: threshold -1 marks	-0.178	-0.006	-0.402	0.104
	(0.312)	(0.157)	(0.352)	(0.196)
Cut-off: threshold +1 marks	0.245	-0.029	-0.396	-0.154
	(0.222)	(0.140)	(0.278)	(0.159)
Cut-off: threshold +2 marks	-0.001	0.001	-0.375	-0.008
	(0.227)	(0.156)	(0.289)	(0.172)

Notes: all estimates are from separate regressions. A bandwidth of ±15 marks around the level threshold is used. Robust standard errors clustered at primary school level in parentheses. ***, ** and * denote statistical significance at the 1 per cent, 5 per cent and 10 per cent levels respectively.

Chapter Five:

Conclusion

This thesis evaluates how outcomes for pupils in England's state schools are shaped by the ways schools are organised and held to account for their performance. Each research chapter combines a rich administrative dataset – the National Pupil Database (NPD) – with transparent, justified and well-established policy evaluation methods.

Chapter two considers the effect of increasing the autonomy of primary schools – via the academies programme – on pupils' attainment during and at the end of primary school, as well as on the entry-year intake of primary schools. Chapter three investigates the impact on individual pupils of not participating in standardised tests at primary schools, while other pupils continue to sit the standardised tests, on subsequent attainment throughout their schooling. Finally, chapter four examines how the feedback each pupil receives based on their performance in the same series of standardised tests affects their subsequent schooling attainment and attendance.

5.1 Summary of Chapter Two

The first academy schools opened in September 2002 and replaced persistently under-performing secondary schools in deprived neighbourhoods. Research has shown that replacing these schools with academies led to an improvement in pupils' attainment during and beyond schooling. In 2010, the government offered all 'well-performing' schools the right to convert to academy status. Academies quickly came to dominate the secondary school sector and around one-quarter of primary schools became academies. However, the rapid rise of the academy model occurred with a limited understanding of whether the model would be effective in schools that were not failing secondary schools.

Chapter two is among the first attempts to evaluate how attainment is affected when primary schools with a track record of good performance transition to the academy model. The chapter uses a difference-in-differences strategy exploiting the availability of a panel of pupil attainment data. The strategy uses differences in the timing of primary schools' transitions to academy status to define treatment and control groups. Schools in both groups have a desire to become academies, but some schools (the treatment group) become academies early on while others (the control group) take longer to switch. Baseline covariates are not statistically different between the two groups which also share a common time trend in the pre-treatment period; this provides suggestive evidence that the identifying assumptions hold.

The results of the difference-in-differences models are consistent with each other. There is no evidence of an effect of academy conversion on attainment in either reading or maths, during or at the end of primary school for the average pupil. However, there is evidence that the attainment of pupils eligible for free school meals does slightly improve.

The academies programme is not inexpensive. As such, the fact that pupil attainment on average neither improves nor deteriorates because of a primary school's transition to the academy model does not indicate that the policy is not harmful. The policy is harmful because of the opportunity costs associated with the policy's facilitation; resources could have been otherwise spent on policy reforms which have been shown to be effective.

There remain several unanswered questions about the effectiveness of the academy programme. The impact on attainment in historically under-performing secondary schools of the transition to academy status is the focus of many papers, however, there is comparatively little academic research on the attainment impacts on historically under-performing primary schools as well as well-performing secondary schools.

Further research on these two contexts may shed light on why academy status does not affect attainment in well-performing primary schools but does affect attainment in challenged secondaries.

5.2 Summary of Chapter Three

Pupils in England's primary schools have had to sit standardised tests at the end of their primary schooling since the mid-1990s. While the principal purpose of these tests is to provide data to track primary school performance, the tests also serve a formative function. That is, the tests are intended to provide insightful feedback to parents and teachers and thus aid pupils' progress.

However, there is considerable concern that standardised testing is harmful for pupils' learning. Standardised testing incentives schools to teach to the test, ignoring untested subject matter while overly focusing on test taking skills. As standardised tests are often introduced (and discontinued) across an entire population of school children and in tandem with school accountability reforms, there is rarely an opportunity to empirically test the effect of participation in standardised tests on subsequent schooling attainment.

A 2010 boycott of KS2 SATs tests, sat by eleven-year olds throughout England, forms the basis of a natural experiment that chapter three exploits. As the boycott had partial coverage (26 per cent of schools participated in the boycott) and since data on pupil attainment is available for numerous cohorts, chapter three again employees a difference-in-differences strategy. Propensity score matching is used to identify non-boycotting primary schools that are similar in observable characteristics to boycotting primary schools. Baseline covariates are not statistically different between the schools that boycotted and the matched control group. They also share a common cohort trend across cohorts preceding the boycott affected cohort; this provides suggestive evidence that the identifying assumptions hold.

As the boycott was officially called only a matter of school days prior to the tests being due to take place, the chapter presents an argument (supported by primary survey data) that the treatment effect of the boycott reflects non-participation in the test only. In other words, the boycott did not affect how pupils prepared for the tests.

The results of the difference-in-differences models indicate that pupils affected by the boycott received inflated contemporaneous teacher assessments in English and maths. However, teacher assessment attainment three years after the boycott as well as secondary school qualification attainment are both slightly adversely impacted by boycott participation for the average pupil. Further analysis also reveals that the subject choice of pupils was affected by the boycott as pupils became less likely to choose an "academic" set of subjects in their final two years of schooling. Boycott affected pupils also missed marginally fewer secondary school lessons.

However, the chapter also presents a collection of difference-in-differences models that accommodate heterogeneity in the treatment effect. The effect of boycott participation on attainment is quite diverse depending on the characteristics of the pupil affected with some sub-groups of pupils estimated to have benefited from participation in the boycott.

The chapter concludes by providing speculation as to the nature of the mechanism. The link between boycott participation and secondary school qualification grade targets is identified. Pupils who were affected by the boycott had artificially high secondary school qualification grade targets. As the proportion of pupils meeting these grade targets was published in school league tables, secondary schools are keen to maximise their rate of success at meeting the targets. This may mean that schools are incentivised to switch attention from boycott affected pupils to non-affected pupils.

5.3 Summary of Chapter Four

As referred to in the previous section, in England's primary schools, pupils are provided with feedback on their performance in standardised tests that are sat at the end of primary school. The purpose of this feedback is to aid pupils' subsequent learning. However, the feedback is relatively coarse, taking the form of integer grades ranging from 2 to 6. The integer grades are linked to descriptors of attainment: levels 2 and 3 denote attainment below the nationally expected standard for eleven-year olds, level 4 denotes attainment at the expected standard, while level 5 denotes attainment beyond the expected level.

The allocation of grades to pupils is based exclusively on test mark thresholds. The government's advice to schools is that pupils should not be informed of their raw mark score on the test. The implication of this guidance is that, in the immediate region of the thresholds, two pupils who score almost identical marks can be given substantially different feedback and remain unaware of how close they came to be receiving the others' feedback.

The result is that idiosyncratic circumstances such as minor arithmetic errors, misreading questions or other misfortune can determine whether a pupil is told that their entire attainment in English or maths throughout primary school is below, at or above a level deemed to be satisfactory for their age. Chapter four evaluates whether "just" crossing the raw mark thresholds to achieve the level 4 and level 5 grades influences secondary school qualification attainment five years later.

The assignment of grades to pupils based on a test score means that the treatment effect of grades is well suited to estimation using sharp regression discontinuity models. The results of the regression discontinuity models indicate that grades on age eleven standardised tests have a polarising effect on attainment in important secondary school qualifications. That is, a pupil who narrowly crosses the threshold required to achieve the level 4 or level 5 grade will on average perform marginally better in their qualifications than if they had just fallen short of passing the threshold.

Heterogeneous treatment effect analysis reveals that, particularly in the case of the English standardised test, that the effect of grades on subsequent attainment are more substantial for pupils with a history of free school meal eligibility than pupils with no history of eligibility. Therefore, chapter four indicates one small avenue by which socioeconomic disadvantage translates into reduced attainment at school. The chapter also considers whether grades affect attendance in the first year of secondary school; no evidence of an effect is uncovered by the regression discontinuity models.

5.4 Limitations of the Thesis

Each research chapter uses data from the NPD. As an administrative dataset covering the entire population of English state funded schools and their pupils, researchers have little reason to be concerned about the accuracy of the NPD data or whether it is a representative sample of the population.

However, the NPD records a limited collection of pupil characteristics compared to survey datasets. For example, the NPD does not contain any information on the characteristics of parents, or the circumstances of a pupil's household. These characteristics are important determinants of educational attainment, and therefore it would be advantageous to include them as control variables in the analysis. At the very least including a richer set of control variables will reduce the variance of the error term in the regression models which would subsequently reduce the standard errors of the coefficient estimates, thus narrowing confidence intervals. More importantly, controlling for a wider range of determinants reduces the potential for omitted variable bias.

A second issue with the NPD dataset is the precision of the measure of attainment at the end of secondary school; the principal outcome of interest in chapters three and four. The NPD records letter grades achieved in GCSE qualifications. This is a relatively broad measure of attainment and limits the amount of variation in attainment that can be observed. A more granular measure of attainment will mean finer variation in attainment can be detected. A similar issue is variation in the way that attainment in GCSEs and equivalent qualifications are recorded in the dataset. Depending on the cohort and subject in question, the NPD either records a pupil's best GCSE grade for that subject or the grade of their first exam entry for that subject. However, the inclusion of cohort fixed effects absorbs this inconsistency.

Secondary school qualifications are often a means to an end. Many pupils complete some form of further education, and therefore only rely on such qualifications to gain access to further education courses. For this reason, there is considerable value in investigating post-secondary school attainment outcomes as part of chapters three and four. The NPD does not contain complete further education data, but it can be found in the Individualised Learner Record (ILR) dataset. The ILR and NPD datasets can be matched at the pupil level, however, it was not possible to access ILR data as part of this programme of research. However, this remains an obvious avenue of future research possibilities.

References

Abdulkadiroğlu, A., Angrist, J. D., Dynarski, S. M., Kane, T. J., & Pathak, P. A. (2011). Accountability and Flexibility in Public Schools: Evidence from Boston's Charters And Pilots. *The Quarterly Journal of Economics*, *126*(2), 699–748. https://doi.org/10.1093/qje/qjr017

Abdulkadiroğlu, A., Angrist, J. D., Hull, P. D., & Pathak, P. A. (2016). Charters without Lotteries: Testing Takeovers in New Orleans and Boston. *American Economic Review*, *106*(7), 1878–1920. https://doi.org/10.1257/aer.20150479

Abdulkadiroğlu, A., Angrist, J. D., Narita, Y., & Pathak, P. A. (2015). *Research Design Meets Market Design: Using Centralized Assignment for Impact Evaluation*. National Bureau of Economic Research (NBER) Working Paper (21705).

Alcott, B. (2017). Might progress assessments hinder equitable progress? Evidence from England. *Educational Assessment, Evaluation and Accountability*, *29*(3), 269–296. https://doi.org/10.1007/s11092-017-9264-2

Anders, J. (2012). The Link between Household Income, University Applications and University Attendance. *Fiscal Studies*, *33*(2), 185–210. https://doi.org/10.1111/j.1475-5890.2012.00158.x

Andersen, S. C., & Nielsen, H. S. (2016). *The Positive Effects of Nationwide Testing on Student Achievement in a Low-Stakes System*.

Angrist, J. D., Cohodes, S. R., Dynarski, S. M., Pathak, P. A., & Walters, C. R. (2016). Stand and Deliver: Effects of Boston's Charter High Schools on College Preparation, Entry, and Choice. *Journal of Labor Economics*, *34*(2), 275–318. https://doi.org/10.1086/683665

Angrist, J. D., Dynarski, S. M., Kane, T. J., Pathak, P. A., & Walters, C. R. (2010). Inputs and Impacts in Charter Schools: KIPP Lynn. *American Economic Review*, *100*(2), 239–243. https://doi.org/10.1257/aer.100.2.239

Angrist, J. D., & Keueger, A. B. (1991). Does Compulsory School Attendance Affect Schooling and Earnings? *The Quarterly Journal of Economics*, *106*(4), 979–1014. https://doi.org/10.2307/2937954

Angrist, J. D., Pathak, P. A., & Walters, C. R. (2011). Explaining Charter School Effectiveness. *American Economic Journal: Applied Economics*, *5*(4), 1–27.

https://doi.org/10.3386/w17332

Autor, D. H. (2003). Outsourcing at Will: The Contribution of Unjust Dismissal Doctrine to the Growth of Employment Outsourcing. *Journal of Labor Economics*, *21*(1), 1–42. https://doi.org/10.1086/344122

Avery, C., Gurantz, O., Hurwitz, M., & Smith, J. (2017). Shifting College Majors in Response to Advanced Placement Exam Scores. *Journal of Human Resources*. https://doi.org/10.3368/jhr.53.4.1016-8293R

Banks, J., & Mazzonna, F. (2012). The Effect of Education on Old Age Cognitive Abilities: Evidence from a Regression Discontinuity Design. *The Economic Journal*, *122*(560), 418–448. https://doi.org/10.1111/j.1468-0297.2012.02499.x

Battistin, E., & Rettore, E. (2008). Ineligibles and eligible non-participants as a double comparison group in regression-discontinuity designs. *Journal of Econometrics*, *142*(2), 715–730. https://doi.org/10.1016/J.JECONOM.2007.05.006

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How Much Should We Trust Differences-In-Differences Estimates? *The Quarterly Journal of Economics*, *119*(1), 249–275. https://doi.org/10.1162/003355304772839588

Björklund, A., Lindahl, M., & Plug, E. (2006). The Origins of Intergenerational Associations: Lessons from Swedish Adoption Data. *Quarterly Journal of Economics*, *121*(3), 999–1028. https://doi.org/10.1162/qjec.121.3.999

Black, S. E. (1999). Do Better Schools Matter? Parental Valuation of Elementary Education. *The Quarterly Journal of Economics*, *114*(2), 577–599. https://doi.org/10.1162/003355399556070

Boyd, D., Lankford, H., Loeb, S., & Wyckoff, J. (2008). The Impact of Assessment and Accountability on Teacher Recruitment and Retention. *Public Finance Review*, *36*(1), 88–111. https://doi.org/10.1177/1091142106293446

Bradley, S., & Lenton, P. (2007). Dropping out of post-compulsory education in the UK: an analysis of determinants and outcomes. *Journal of Population Economics*, *20*(2), 299–328. https://doi.org/10.1007/s00148-006-0110-y

Burgess, S., Propper, C., Slater, H., & Wilson, D. (2005). *Who wins and who loses from school accountability? The distribution of educational gans in English secondary schools* (Centre for Market and Public Organisation No. 05/128).

https://doi.org/10.1016/S1043-2760(97)84344-5

Burgess, S., Wilson, D., & Worth, J. (2013). A natural experiment in school accountability: The impact of school performance information on pupil progress. *Journal of Public Economics*, *106*, 57–67. https://doi.org/10.1016/j.jpubeco.2013.06.005

Cameron, A. C., & Miller, D. L. (2015). A Practitioner's Guide to Cluster-Robust Inference. *Journal of Human Resources*, *50*(2), 317–372. https://doi.org/10.3368/jhr.50.2.317

Canaan, S., & Mouganie, P. (2018). Returns to Education Quality for Low-Skilled Students: Evidence from a Discontinuity. *Journal of Labor Economics*, *36*(2), 395–436. https://doi.org/10.1086/694468

Chowdry, H., Crawford, C., Dearden, L., Goodman, A., & Vignoles, A. (2013). Widening participation in higher education: analysis using linked administrative data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *176*(2), 431–457. https://doi.org/10.1111/j.1467-985X.2012.01043.x

Clark, D. (2009). The Performance and Competitive Effects of School Autonomy. *Journal of Political Economy*, *117*(4), 745–783. https://doi.org/10.1086/605604

Connor, M. J. (2001). Pupil stress and standard assessment tasks. *Emotional and Behavioural Difficulties*, *6*(2), 103–111. https://doi.org/10.1080/13632750100507660

Connor, M. J. (2003). Pupil stress and standard assessment tasks (SATs) An update. *Emotional and Behavioural Difficulties*, *8*(2), 101–107. https://doi.org/10.1080/13632750300507010

CREDO. (2013). *National Charter School Study*. Stanford, CA: Center for Research on Education Outcomes.

Day, C., & Smethem, L. (2009). The effects of reform: Have teachers really lost their sense of professionalism? *Journal of Educational Change*, *10*(2–3), 141–157. https://doi.org/10.1007/s10833-009-9110-5

DCSF. (2009). *National Curriculum Assessments at Key Stage 2 in England, 2008 (Revised)*. London.

Dee, T. S. (2004). Are there civic returns to education? *Journal of Public Economics, 88*(9–10), 1697–1720. https://doi.org/10.1016/j.jpubeco.2003.11.002

Dee, T. S., & Jacob, B. A. (2011). The impact of no Child Left Behind on student achievement. *Journal of Policy Analysis and Management*, *30*(3), 418–446. https://doi.org/10.1002/pam.20586

Deming, D. J., Cohodes, S. R., Jennings, J., & Jencks, C. (2016). School Accountability, Postsecondary Attainment, and Earnings. *Review of Economics and Statistics*, *98*(5), 848–862. https://doi.org/10.1162/REST_a_00598

Department for Education. (2010). *The Importance of Teaching, the Schools White Paper 2010*. London: Her Majesty's Stationery Office.

Department for Education. (2014). *The Education Services Grant*. London: Department for Education.

Department for Education. (2016). *Educational Excellence Everywhere*. London: Her Majesty's Stationery Office.

Department for Education. (2018). *Schools, pupils and their characteristics: January 2018*. London.

Dobbie, W., & Fryer, R. G. (2011a). Are high-quality schools enough to increase achievement among the poor? Evidence from the Harlem Children's Zone. *American Economic Journal: Applied Economics*, *3*(3), 158–187. https://doi.org/10.1257/app.3.3.158

Dobbie, W., & Fryer, R. G. (2011b). Getting beneath the Veil of Effective Schools: Evidence from New York City. *American Economic Journal: Applied Economics, 5*(4), 28–60. https://doi.org/10.3386/w17632

Dobbie, W., & Fryer, R. G. (2015). The Medium-Term Impacts of High-Achieving Charter Schools. *Journal of Political Economy*, *123*(5), 985–1037. https://doi.org/10.1086/682718

Eyles, A., Hupkau, C., & Machin, S. (2016a). Academies, charter and free schools: do new school types deliver better outcomes? *Economic Policy*, *31*(87), 453–501. https://doi.org/10.1093/epolic/eiw006

Eyles, A., Hupkau, C., & Machin, S. (2016b). School reforms and pupil performance. *Labour Economics*, *41*(C), 9–19. https://doi.org/10.1016/j.labeco.2016.05.004

Eyles, A., & Machin, S. (2015). *The Introduction of Academy Schools to England's Education*. Institute for the Study of Labor (IZA) Discussion Paper (9276).

Eyles, A., Machin, S., & Mcnally, S. (2016). *Unexpected School Reform: Academisation of*

Primary Schools in England. Centre for Economic Performance (LSE) Discussion Paper (1455).

Eyles, A., Machin, S., & Silva, O. (2015). *Academies 2: The New Batch*. Centre for the Economics of Education (LSE) Discussion Paper (1370).

Fan, J. (1992). Design-adaptive nonparametric regression. *Journal of the American Statistical Association*, *87*(420), 998–1004. https://doi.org/10.1080/01621459.1992.10476255

Figlio, D. N. (2006). Testing, crime and punishment. *Journal of Public Economics*, *90*(4–5), 837–851. https://doi.org/10.1016/j.jpubeco.2005.01.003

Figlio, D. N., & Kenny, L. W. (2009). Public sector performance measurement and stakeholder support. *Journal of Public Economics*, *93*(9–10), 1069–1077. https://doi.org/10.1016/j.jpubeco.2009.07.003

Figlio, D. N., & Loeb, S. (2011). School Accountability. In E. Hanushek, S. Machin, & L. Woessmann (Eds.), *Handbook of the Economics of Education* (1st ed., Vol. 3, pp. 383–421). Elsevier B.V. https://doi.org/10.1016/B978-0-444-53429-3.00008-9

Figlio, D. N., & Lucas, M. E. (2004). What's in a Grade? School Report Cards and the Housing Market. *American Economic Review*, *94*(3), 591–604. https://doi.org/10.1257/0002828041464489

Figlio, D. N., & Winicki, J. (2005). Food for thought: the effects of school accountability plans on school nutrition. *Journal of Public Economics*, *89*(2–3), 381–394. https://doi.org/10.1016/j.jpubeco.2003.10.007

Fryer, R. G. (2014). Injecting Charter School Best Practices into Traditional Public Schools: Evidence from Field Experiments. *The Quarterly Journal of Economics*, *129*(3), 1355–1407. https://doi.org/10.1093/qje/qju011

Gibbons, S., & Machin, S. (2003). Valuing English primary schools. *Journal of Urban Economics*, *53*(2), 197–219. https://doi.org/10.1016/S0094-1190(02)00516-8

Gleason, P., Clark, M., Tuttle, C., & Dwoyer, E. (2010). *The Evaluation of Charter School Impacts: Final Report*. Washington, DC: US Department for Education National Center for Education Evaluation and Regional Assistance.

Hahn, J., Todd, P. E., & Klaauw, W. (2001). Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design. *Econometrica*, *69*(1), 201–209. https://doi.org/10.1111/1468-0262.00183

Hanushek, E. A. (2011). The economic value of higher teacher quality. *Economics of Education Review*, *30*(3), 466–479. https://doi.org/10.1016/j.econedurev.2010.12.006

Hanushek, E. A., & Raymond, M. E. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management*, *24*(2), 297–327. https://doi.org/10.1002/pam.20091

Harmon, C., & Walker, I. (1995). Estimates of the Economic Return to Schooling for the United Kingdom. *The American Economic Review*, *85*(5), 1278–1286.

Henderson, M., Sullivan, A., Anders, J., & Moulton, V. (2018). Social Class, Gender and Ethnic Differences in Subjects Taken at Age 14. *The Curriculum Journal*, *29*(3), 298–318. https://doi.org/10.1080/09585176.2017.1406810

HM Treasury. (2017). *Public Expenditure Statistical Analyses 2017*. London: Her Majesty's Stationery Office.

Holmlund, H., Lindahl, M., & Plug, E. (2011). The Causal Effect of Parents' Schooling on Children's Schooling: A Comparison of Estimation Methods. *Journal of Economic Literature*, *49*(3), 615–651. https://doi.org/10.1257/jel.49.3.615

Hoxby, C., & Murarka, S. (2009). *Charter Schools in New York City: Who Enrolls and How They Affect Their Students' Achievement*. National Bureau of Economic Research (NBER) Working Paper (14852).

Hutchings, M., Francis, B., & De Vries, R. (2014). *Chain Effects: The Impact of Academy Chains on Low Income Students*. London: The Sutton Trust.

Hutchings, M., Francis, B., & Kirby, P. (2015). *Chain Effects 2015: The Impact of Academy Chains on Low-Income Students*. London: The Sutton Trust.

Jacob, B. A. (2005). Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics*, *89*(5–6), 761–796. https://doi.org/10.1016/j.jpubeco.2004.08.004

Jacob, B. A., & Levitt, S. D. (2003). Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating. *The Quarterly Journal of Economics*, *118*(3), 843–877. https://doi.org/10.1162/00335530360698441

Klein, S. P., Hamilton, L., McCaffrey, D. F., & Stecher, B. (2000). What Do Test Scores in Texas Tell Us? *Education Policy Analysis Archives*, *8*(49). https://doi.org/10.14507/epaa.v8n49.2000

Kolesár, M., & Rothe, C. (2018). Inference in Regression Discontinuity Designs with a Discrete Running Variable. *American Economic Review*, *108*(8), 2277–2304. https://doi.org/10.1257/aer.20160945

Koretz, D. (2002). Limitations in the Use of Achievement Tests as Measures of Educators' Productivity. *The Journal of Human Resources*, *37*(4), 752. https://doi.org/10.2307/3069616

Koretz, D., & Barron, S. (1998). *The Validity of Gains in Scores on the Kentucky Instructional Results Information System (KIRIS)*. Santa Monica, CA: RAND Corporation.

Lazear, E. P. (2006). Speeding, Terrorism, and Teaching to the Test. *Quarterly Journal of Economics*, *121*(3), 1029–1061. https://doi.org/10.1162/qjec.121.3.1029

Leckie, G., & Goldstein, H. (2017). The evolution of school league tables in England 1992–2016: 'Contextual value-added', 'expected progress' and 'progress 8.' *British Educational Research Journal*, *43*(2), 193–212. https://doi.org/10.1002/berj.3264

Lee, D. S. (2008). Randomized experiments from non-random selection in U.S. House elections. *Journal of Econometrics*, *142*(2), 675–697. https://doi.org/10.1016/j.jeconom.2007.05.004

Lee, D. S., & Card, D. (2008). Regression discontinuity inference with specification error. *Journal of Econometrics*, *142*(2), 655–674. https://doi.org/10.1016/j.jeconom.2007.05.003

Lee, D. S., & Lemieux, T. (2010). Regression Discontinuity Designs in Economics. *Journal of Economic Literature*, *48*(2), 281–355. https://doi.org/10.1257/jel.48.2.281

Lleras-Muney, A. (2005). The Relationship Between Education and Adult Mortality in the United States. *Review of Economic Studies*, *72*(1), 189–221. https://doi.org/10.1111/0034-6527.00329

Lochner, L., & Moretti, E. (2004). The Effect of Education on Crime: Evidence from Prison Inmates, Arrests, and Self-Reports. *American Economic Review*, *94*(1), 155–189. https://doi.org/10.1257/000282804322970751

Machin, S., Marie, O., & Vujić, S. (2011). The Crime Reducing Effect of Education. *The Economic Journal*, *121*(552), 463–484. https://doi.org/10.1111/j.1468-0297.2011.02430.x

Machin, S., Mcnally, S., & Ruiz-valenzuela, J. (2018). *Entry Through the Narrow Door : The*

Costs of Just Failing High Stakes Exams. CVER Discussion Paper Series (014).

Machin, S., & Vernoit, J. (2011). *Changing School Autonomy: Academy Schools and their Introduction to England's Education*. Centre for the Economics of Education (LSE) Disucssion Paper (123).

Machin, S., & Wilson, J. (2009). Public and Private Schooling Initatives in England. In R. Chakrabarti & P. E. Peterson (Eds.), *School Choice International: Exploring Public-Private Partnerships* (pp. 219–241). Cambridge, MA: MIT Press.

Martorell, F. (2004). *Do High School Graduation Exams Matter? A Regression Discontinuity Approach.*

Milligan, K., Moretti, E., & Oreopoulos, P. (2004). Does education improve citizenship? Evidence from the United States and the United Kingdom. *Journal of Public Economics*, *88*(9–10), 1667–1695. https://doi.org/10.1016/J.JPUBECO.2003.10.005

Moretti, E. (2004a). Estimating the social return to higher education: evidence from longitudinal and repeated cross-sectional data. *Journal of Econometrics*, *121*(1–2), 175–212. https://doi.org/10.1016/J.JECONOM.2003.10.015

Moretti, E. (2004b). Workers' Education, Spillovers, and Productivity: Evidence from Plant-Level Production Functions. *American Economic Review*, *94*(3), 656–690. https://doi.org/10.1257/0002828041464623

Moulton, V., Sullivan, A., Henderson, M., & Anders, J. (2018). Does what you study at age 14–16 matter for educational transitions post-16? *Oxford Review of Education*, *44*(1), 94–117. https://doi.org/10.1080/03054985.2018.1409975

National Audit Office. (2010). *The Academies Programme*. London: The Stationery Office.

National Audit Office. (2012). *Managing the Expansion of the Academies Programme*. London: The Stationery Office.

National Audit Office. (2016). *Financial sustainability of schools*. London: National Audit Office.

Neal, D., & Schanzenbach, D. W. (2010). Left Behind by Design: Proficiency Counts and Test-Based Accountability. *Review of Economics and Statistics*, *92*(2), 263–283. https://doi.org/10.1162/rest.2010.12318

Oreopoulos, P. (2007). Do dropouts drop out too soon? Wealth, health and happiness from compulsory schooling. *Journal of Public Economics*, *91*(11–12), 2213–2229. https://doi.org/10.1016/j.jpubeco.2007.02.002

Ou, D. (2010). To leave or not to leave? A regression discontinuity analysis of the impact of failing the high school exit exam. *Economics of Education Review*, *29*(2), 171–186. https://doi.org/10.1016/J.ECONEDUREV.2009.06.002

Papay, J. P., Murnane, R. J., & Willett, J. B. (2010). The Consequences of High School Exit Examinations for Low-Performing Urban Students: Evidence From Massachusetts. *Educational Evaluation and Policy Analysis*, *32*(1), 5–23. https://doi.org/10.3102/0162373709352530

Papay, J. P., Murnane, R. J., & Willett, J. B. (2014). High-School Exit Examinations and the Schooling Decisions of Teenagers: Evidence From Regression-Discontinuity Approaches. *Journal of Research on Educational Effectiveness*, *7*(1), 1–27. https://doi.org/10.1080/19345747.2013.819398

Papay, J. P., Murnane, R. J., & Willett, J. B. (2016). The Impact of Test Score Labels on Human-Capital Investment Decisions. *Journal of Human Resources*, *51*(2), 357–388. https://doi.org/10.3368/jhr.51.2.0713-5837R

Powdthavee, N. (2010). Does Education Reduce the Risk of Hypertension? Estimating the Biomarker Effect of Compulsory Schooling in England. *Journal of Human Capital*, *4*(2), 173–202. https://doi.org/10.1086/657020

PriceWaterhouseCoopers. (2008). *Academies Evaluation Fifth Annual Report*. Annesley: DCSF Publications.

Putwain, D. W., Connors, L., Woods, K., & Nicholson, L. J. (2012). Stress and anxiety surrounding forthcoming Standard Assessment Tests in English schoolchildren. *Pastoral Care in Education*, *30*(4), 289–302. https://doi.org/10.1080/02643944.2012.688063

Qualifications and Curriculum Development Agency. (2009). *2009 National curriculum tests review outcomes (provisional)*.

Rauch, J. E. (1993). Productivity Gains from Geographic Concentration of Human Capital: Evidence from the Cities. *Journal of Urban Economics*, *34*(3), 380–400. https://doi.org/10.1006/juec.1993.1042

Reardon, S. F., Arshan, N., Atteberry, A., & Kurlaender, M. (2010). Effects of Failing a High School Exit Exam on Course Taking, Achievement, Persistence, and Graduation.

Educational Evaluation and Policy Analysis, *32*(4), 498–520. https://doi.org/10.3102/0162373710382655

Reback, R. (2008). Teaching to the rating: School accountability and the distribution of student achievement. *Journal of Public Economics*, *92*(5–6), 1394–1415. https://doi.org/10.1016/j.jpubeco.2007.05.003

Rouse, C. E., Hannaway, J., Goldhaber, D., & Figlio, D. N. (2013). Feeling the Florida Heat? How Low-Performing Schools Respond to Voucher and Accountability Pressure. *American Economic Journal: Economic Policy*, *5*(2), 251–281. https://doi.org/10.1257/pol.5.2.251

Sartarelli, M. (2011). *Do Performance Targets Affect Behaviour? Evidence from Discontinuities in Test Scores in England*. IOE Department of Quantitative Social Science Working Paper 11-02.

Silles, M. A. (2009). The causal effect of education on health: Evidence from the United Kingdom. *Economics of Education Review*, *28*(1), 122–128. https://doi.org/10.1016/j.econedurev.2008.02.003

Smith, J., Hurwitz, M., & Avery, C. (2016). Giving College Credit Where It Is Due: Advanced Placement Exam Scores and College Outcomes. *Journal of Labor Economics*, *35*(1), 67–147. https://doi.org/10.1086/687568

Todd, P. E., & Wolpin, K. I. (2003). On the Specification and Estimation of the Production Function for Cognitive Achievement. *The Economic Journal*, *113*(485), F3–F33. https://doi.org/10.1111/1468-0297.00097

West, A., & Pennell, H. (2000). Publishing School Examination Results in England: Incentives and consequences. *Educational Studies*, *26*(4), 423–436. https://doi.org/10.1080/03055690020003629

Whetton, C. (2009). A brief history of a testing time: national curriculum assessment in England 1989–2008. *Educational Research*, *51*(2), 137–159. https://doi.org/10.1080/00131880902891222

Wiggins, A., & Tymms, P. (2002). Dysfunctional Effects of League Tables: A Comparison Between English and Scottish Primary Schools. *Public Money and Management*, *22*(1), 43–48. https://doi.org/10.1111/1467-9302.00295

Worth, J. (2015). *Analysis of Academy School Performance in 2015*. Slough: National Foundation for Educational Research.