Essays on the Economics of Education

Hester Burn

A thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy of Economics

Institute for Social and Economic Research

University of Essex

May 2025

Declaration

I, Hester Burn, declare that this thesis titled ‘Essays on the Economics of Education’, and the work presented in it, are my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

I declare that the work contained in this thesis was carried out in accordance with the requirements of the University’s Regulation and that no part of this thesis has been submitted for any other degree.

Chapter 1 is co-authored by myself, Professor Birgitta Rabe and Dr Laura Fumagalli. Chapter 3 is co-authored by myself, Professor Birgitta Rabe and Professor Cheti Nicoletti. Both use the National Pupil Database which is owned by the Department for Education in England, and both were produced using statistical data accessed via the ONS Secure Research Service. However the use of these data in this work does not imply the endorsement of the ONS in relation to the interpretation or analysis of the statistical data. This work also uses research datasets which may not exactly reproduce National Statistics aggregates.

Chapter 2 is exclusively my own. Unlike the other chapters, Chapter 2 makes use of anonymised data from Wales which is held in the Secure Anonymised Information Linkage (SAIL) Databank. I would like to take this opportunity to acknowledge and thank all the data providers who make anonymised data available for research.

Finally, Chapter 1 is published with the following reference information: Burn, H., Fumagalli, L., Rabe, B. (2024) Stereotyping and Ethnicity Gaps in Teacher Assigned Grades. *Labour Economics* 89. https://doi.org/10.1016/j.labeco.2024.102577. Neither Chapter 2 nor Chapter 3 are currently published in any form.

Hester Burn

July 2025

Acknowledgements

I feel immensely lucky to have a lot of quite incredible people in my life. People who listened to my fears, endured my silent presence when it was all too much, and dragged me onto the dancefloor when I needed to remember that I wasn't just a walking brain. To Martha, Connie, Charlotte, Michael, Tash, Oli, Eliza and Anne. To my family, because they're wonderful. To Vanessa, for her incredible generosity and giving me a bed whenever I needed it. And to so many more I don't have room to mention here: thank you.

I also feel incredibly grateful to have worked with some wonderful academics and supervisors over the last few years. I would particularly like to thank Birgitta Rabe, Laura Fumagalli, Angus Holford and Emilia Del Bono for helping me to feel less like an imposter, setting the bar, and believing I could reach it. I'd like to thank Ashley, Omar, Annalivia, Sonkurt and Tommaso for patiently initiating me into the world of economics. And I'd like to thank everyone at ISER, which was a really lovely community to join.

I am very grateful to Economic and Social Research Council for providing me with a scholarship to work on this thesis, and to the ESRC Research Centre on Micro-Social Change for additional funding and other opportunities that have been offered me during my PhD. This includes purchasing the data required for Chapter 2.

Lastly, I'd like to thank all the pupils I taught before starting this degree. I did it for them, really, and remembering that unique individuals with hopes and fears and struggles and accomplishments and humour and incredible resilience sat behind all the little data points that I was analysing is really what kept me going. I hope that I can contribute to making the world they enter a more equitable and supportive place.

Abstract

This thesis consists of three standalone papers which explore the extent to which education policies can change children's outcomes, whether they tend to do this for better or for worse, and what factors get in the way.

Chapter 1 explores changes in the attainment gaps between pupils from different ethnic groups when grades are assigned by teacher predictions rather than through blindly marked examinations. When grades are assigned by teachers, ethnic minority pupils in England receive higher grades in maths and lower grades in English relative to White British pupils and compared to when grades are assigned through exams. Using Gelbach decompositions, we show that observed characteristics partly explain the maths gap changes but amplify those in English. We conclude that group-specific stereotyping is a convincing explanation of the results.

Chapter 2 evaluates the medium-term effects of an extended play-based learning policy for early childhood in Wales – the Foundation Phase – on a range of school-related outcomes. I use a staggered difference-in-differences research design to compare pupils who received the Foundation Phase to pupils who instead received formal education between ages five and seven. I find no evidence of effects at ages 11 or 16.

Finally, Chapter 3 studies sibling spillover effects in temporary exclusions and unauthorised absences in England. We estimate sibling spillover effects when siblings are in the same grade and use a novel instrumental variable strategy exploiting variation in the behaviour of older siblings' peer groups within and across school cohorts. We find evidence of modest spillover effects from the older to the younger sibling in both exclusions and absences. We also find that sibling pairs are more likely to be excluded for the same stated reason, suggesting that role modelling is a plausible mechanism for our results.

Contents

List of Figures

List of Tables

Introduction

Education policy is generally designed to address one or more of the competing political goals of democratic equality, social efficiency and social mobility (Labaree, 1997). In recent years the goals of social efficiency and social mobility have increased in prominence in the UK; education is a means to secure employment outcomes, rather than an end in its own right, and emphasis is placed on breaking down barriers to individuals' successes. This focus is largely a reflection of the wider economic context. When standards of living are falling and wealth inequality is on the rise, education can be seen as a lifeline and a ticket to greater security, and schools powerful tools for change. Well-structured, evidence-informed policies, it is assumed, should be capable of improving outcomes and facilitating all children to reach their potential.

Unfortunately, empirical research consistently shows that inequalities in educational outcomes are stubbornly persistent. Factors such as socio-economic background, prior attainment and other early-life experiences explain the vast majority (around 80 to 90 percent) of variation in grades at the end of secondary schooling (Wilkinson et al., 2018). In England, the attainment gap between children from disadvantaged backgrounds and their more advantaged peers is already evident by the time children start school, and it remains wide at age 16 (Farquharson et al., 2024). Despite decades of reform, the scale and persistence of this gap suggests that many policies either do not target the root causes of inequality, or are unable to overcome them. This raises important questions about the limits of policy intervention in education and what circumstances are necessary to help it to succeed.

The three essays that constitute this thesis, each of which can be understood as a standalone academic paper, explore the extent to which education policies can change children's outcomes, whether they tend to do this for better or for worse, and what factors get in the way. Using large-scale administrative data from England and Wales, these papers examine three different domains in which education policy plays out in the real world: how different forms of assessment increase or diminish gaps in attainment, whether curriculum reform in early childhood education leads to lasting improvements, and how struggles with educational engagement spill over within families. Each paper uses quasi-experimental methods to shed light on a different mechanism through which policy may succeed or fail in shaping outcomes. Together, they offer a perspective on what education policy can achieve and where it runs up against deeper constraints.

Across these three papers a common theme emerges: education policy often falls short because it fails to account for practical, familial or structural constraints. Different forms of assessment, for example, risk reproducing inequalities due to historical patterns in attainment and unconscious biases held by assessors. Curriculum and pedagogical reforms may depend on substantial training and resources to meaningfully change what happens in classrooms. And any education policy's effects must be substantial to counteract wider factors outside of the control of schools (such as the family environment) that determine children's outcomes. Overall, each paper highlights the complexity of the educational process. Together, they underscore how policy which overlooks the real-world interplay of schools, families and broader systems is likely to fall short of its aims, however well-intentioned those aims may be.

Chapter 1 explores whether and how grades assigned using different forms of educational assessment differ systematically according to pupils' characteristics. To do this, we exploit a change in assessment methods in England which occurred during the Covid-19 pandemic. In March 2020, concerned about the possible spread of infection within schools, the government decided to cancel a series of high-stakes examinations and to instead use teachers' best predictions of the grades that pupils *would* have received in their place. We treat this change in the medium of assessment as exogeneous, and compare grade gaps in attainment between ethnic minority and White British pupils in 2020, when grades were assigned by teachers, to those in previous years, when test scores were assigned through examinations. We find that, when grades are assigned by teachers, ethnic minority pupils tend to do relatively better than White British pupils in maths and relatively worse than White British pupils in English compared to when the grades are from externally marked examinations.

As we are unable to directly relate any changes in outcomes between groups to teacher stereotyping, we then follow the literature in ruling out as many other potential channels that may explain the group-specific grade gap changes as possible. Primarily, we use an extension of the Gelbach decomposition (Gelbach, 2016) to investigate whether any changes in grade gaps can be accounted for by differences in the levels of, or returns to, observed characteristics between the 2019 and the 2020 cohorts. We find that, while changes in observed characteristics and their returns between 2019 and 2020 explain part of the higher maths grades attained in 2020 by ethnic minority pupils relative to White British pupils, they do not explain the relatively lower grades attained by ethnic minority pupils in English. In fact, our analysis suggests that accounting for these changes in observed characteristics would roughly double the magnitude of the grade gap changes in English. We also find that the grade gap changes are not driven by time trends or ceiling effects. We conclude that group- and subject-specific stereotyping is likely to drive at least part of the differences in performance between ethnic groups when grades are assigned by teachers.

Chapter 2 explores the question of what skills and capacities early childhood development policies should target, and how. To do this I evaluate the Pilot phase of an extended play-based learning policy for children between ages five and seven in Wales against a counterfactual of formal schooling. This policy – called the Foundation Phase – was piloted through a staggered roll-out in 43 schools between 2004 and 2009. My analysis therefore implements a staggered difference-in-differences research design comparing consecutive cohorts of pupils attending schools. I address recent econometric concerns with heterogeneous treatment effects in settings with staggered treatment timing by using the imputation-based difference-in-difference estimator proposed by Borusyak et al. (2024), and use carefully selected control variables to make the conditional parallel trends assumption more plausible.

I find no evidence of effects of the Foundation Phase Pilot on pupils' school-related outcomes at ages 11 or 16, including on their academic attainment, educational progression, school attendance and school exclusions. This is an interesting result as it suggests that, on average, pupils who received two fewer years of direct instruction during early Primary education still achieved similarly to their peers in later education, made similar decisions about their educational pathways, and exhibited similar school-related behaviours. It might be that there are contemporaneous effects at a younger age which have already faded out by the time pupils reach ages 11 or 16, or so-called 'sleeper' effects which only appear at older ages. In addition, should there be any time-sensitive skills and capacities which greater play-based learning afforded the pupils receiving the Pilot provision but which I am unable to measure in my data – for example, positive attitudes towards learning, improved wellbeing or active citizenship – then these results suggest that it might be possible for children to develop those skills *and* still catch up with their peers on standard measures.

Finally, Chapter 3 investigates the extent to which non-cognitive skills are transmitted between siblings. We do this by estimating the spillover effects in siblings' temporary exclusions and unauthorised absences, which are two important behavioural outcomes which we can observe for all state-educated pupils in England. The identification of sibling spillover effects is challenging because of the many traits and environments that siblings share. We attempt to overcome this primary identification issue by using an instrumental variable approach whereby the older sibling's exclusions and unauthorised absences are instrumented using idiosyncratic variation in the misbehaviour of the school peers with whom they share the most observable characteristics and thus spend, we assume, the most time. In essence, our approach is similar to a refined school fixed effect and instrumental variable estimation exploiting quasi-random variation in peer groups within and across school cohorts.

We find evidence of modest spillover effects in temporary exclusions and unauthorised absences from older siblings to younger siblings when they are in the same grade. We also find larger effects for sibling pairs who are closer in age, and that sibling pairs are more likely to be excluded for the same stated reason, making role modelling a plausible mechanism for our results. We find that siblings who are more economically disadvantaged have slightly larger spillover effects than those who are more advantaged. This suggests that such externalities may have implications for persistent inequalities. It also suggests that spillover effects should be included in estimations of the aggregate effects of targeted interventions, as otherwise policy makers and evaluators may underestimate the impact on inequalities of interventions that successfully improve the non-cognitive skills of disadvantaged pupils who are also older siblings.

The remainder of this thesis consists of each chapter presented separately, followed by some concluding remarks.

1 Stereotyping and Ethnicity Gaps in Teacher Assigned Grades

1.1 Introduction

Pupils are regularly assessed directly by teachers in educational settings. However, non-blind teacher assessments have been found to be susceptible to group-specific stereotyping, for example by ethnic group or gender (Botelho et al., 2015; Carlana, 2019). Stereotypes are beliefs that people have about groups with a particular characteristic and are often held unconsciously, including by teachers (Starck et al., 2020). When held, stereotypes may inform the 'mental shortcuts' that teachers use when they assess pupils (Burgess and Greaves, 2013), and can affect pupils in a number of important ways. They contribute to the formation of pupils' own beliefs about their ability in school subjects and are instrumental for the subjects they choose (Burgess et al., 2022) as well as for their performance in tests (Lavy and Sand, 2018). In many contexts non-blind teacher assessments also directly influence pupils' education trajectory at major branching points, for example in England where teacher grade predictions are used for university applications, in Germany where teacher assessments affect school tracking decisions, or in the US where they inform pupils' Grade Point Averages. Whether and how grades assigned by teachers differ systematically according to pupils' characteristics is therefore an important consideration for education policy.

This paper examines the contribution of stereotyping to attainment gaps between pupils from different ethnic groups when grades are assigned by teachers. We focus on attainment gaps between

White British pupils and pupils from the four largest non-White ethnic groups in England in national examinations at the end of compulsory full-time schooling, at age 16. In England, ethnic minority pupils are less likely to enrol into competitive university courses than their White peers with the same educational profile and preferences, driven in part by a mismatch between predicted and achieved grades (Del Bono et al., 2022). The contribution of stereotypes to teacher assessments may help to explain differences by ethnicity in important outcomes such as this.

We use data from the National Pupil Database, an administrative dataset of children in state-funded schools in England containing students' grades as they progress through school. We focus on two core subjects that are compulsory for all pupils at age 16: English and maths. We exploit a change in assessment methods during the 2020 Covid-19 pandemic, when examinations were cancelled, and teachers were instead asked to predict the grades the pupil would have received had the examinations taken place. We compare grade gaps in attainment between ethnic minority and White British pupils in 2019, when test scores were assigned through blindly marked examinations, and in 2020, when grades were assigned by teachers. This double difference approach captures changes in ethnicity grade gaps resulting from the change in assessment methods. Because stereotyped beliefs are unobservable in our data (as they often are for teachers themselves) it is not possible for us to directly relate any changes in outcomes between groups to teacher stereotyping. Instead, we follow the literature in ruling out as many other potential channels that may explain group-specific grade gap changes as possible, and interpret stereotyping as the most likely source of unexplained grade gap changes (Botelho et al., 2015; Burgess and Greaves, 2013).

We document that, when grades are assigned by teachers, ethnic minority pupils tend to do relatively better than White British pupils in maths and relatively worse than White British pupils in English compared to when the grades are from externally marked examinations. Grades were between 10 and 20% of a grade higher for ethnic minority compared to White British pupils in maths in 2020, when grades were assigned by teachers, compared to in previous years, whereas in English they were about 7% of a grade lower (with the exception of Black Caribbean pupils). Though small, these changes are not unimportant. Pupils who marginally fail to obtain a pass in English at age 16 are about nine percentage points less likely to enrol in an upper secondary qualification (Machin et al., 2020), and a one grade increase in maths is estimated to be associated with an increase of £14,579 in present value of lifetime earnings (Department for Education, 2021).

To assess the potential contribution of stereotyping to the ethnicity grade gap changes between 2020 and 2019, we consider other channels that are not related to teacher stereotyping but may drive the changes we observe. We start off by examining time trends in students' age 16 grades in maths and English. Next, we use an extension of the Gelbach decomposition (Gelbach, 2016) to investigate whether any changes in grade gaps can be accounted for by differences in the levels of, or returns to, observed characteristics between the 2019 and the 2020 cohorts. Third, as pupils received higher grades in 2020 on average, we complete separate decomposition analyses by prior attainment group to investigate whether any changes in grade gaps might be driven by ceiling effects.

Changes in observed characteristics and their returns between 2019 and 2020 explain part of the higher maths grades attained in 2020 by ethnic minority pupils relative to White British pupils. However, they do not explain the lower grades attained by ethnic minority pupils, relative to White British pupils, in English. In fact, our analysis suggests that accounting for these changes in observed characteristics would roughly double the relative drop in the English performance of ethnic minorities that followed the switch to teacher assessments. We also find that the grade gap changes are not driven by time trends or ceiling effects. There may remain characteristics which contribute to the ethnicity grade gap changes in 2020 but are unobservable in our data. Nonetheless, as long as no other explanations account for the ethnicity grade gaps changing in a positive direction for maths and negative direction for English, we argue that group- and subject-specific stereotyping is likely to drive at least part of the differences in performance between ethnic groups when grades are assigned by teachers.

As we will detail later, the timing of the announcement to cancel examinations in 2020 allows us to rule out that a change in teaching practices that might differentially favour pupils by ethnic background could be responsible for the differences we find between ethnicity grade gaps in 2020 and in years in which exams were externally marked. The announcement that examinations were to be cancelled in England occurred just two months before exams were due to begin, so that teaching practices remained largely constant across years. Teachers were also instructed to base their grade predictions only on pupils' work prior to the announcement rather than on any work completed after school closures.

Indeed, while it is possible that teachers' predictions were still informed by differences in pupils' home learning experiences during Covid-19, a substantial proportion of teachers actually stopped interacting with their pupils altogether after school closures (Eivers et al., 2020), and there is evidence that teachers primarily based their grades on the results of practice examinations taken a few months prior (Holmes et al., 2021).

The first contribution of this paper is to the literature on stereotyping which is concerned with explaining and documenting stereotypes in social situations. Our results diverge in two interesting ways from previous findings. First, we document a case where a specific group appears to be subject to both positive and negative stereotypes. While coexisting positive and negative stereotypes are themselves not unusual – for example, wealthy individuals might be viewed as both more assertive and more immoral (Tanjitpiyanond et al., 2022) – there is considerably less evidence showing divergent stereotypes within the same broad competency, such as academic achievement, outside the English context.[1] Both positive and negative stereotypes are important in our context, as both can be harmful if they affect pupils' choices or opportunities. Second, our results are *prima facie* inconsistent with economic approaches whereby stereotypes are held to be manifestations of statistical discrimination (Phelps, 1972) or representativeness heuristics (Bordalo et al., 2016; Esponda et al., 2023). In our case, most ethnic minority groups normally attain more highly than their White British peers in both maths and English, which is at odds with these explanations.

Second, this paper contributes to the literature in economics which attempts to identify the effects of using different forms of assessment, such as classroom-based assessments or examinations, in educational settings. The methodological difficulty that much of this literature seeks to overcome is that teachers often educate and assess the same pupils, making it impossible to distinguish whether effects are driven by assessment or teaching practices (Dee, 2005). Most papers address this difficulty by comparing "blind" (examination) and "non-blind" (teacher) assessments of the same pupil (Burgess and Greaves, 2013; Campbell, 2015; Lindahl, 2007). However this approach rests on the sometimes

[1] The vast majority of international evidence on teacher stereotyping finds evidence of bias and stereotyping only in the same direction across subjects, including a negative bias against girls (Alan et al., 2018; Lavy and Megalokonomou, 2019; Lavy and Sand, 2018), against boys (Lavy, 2008; Lindahl, 2007), against immigrant pupils (Alesina et al., 2018; De Benedetto and De Paola, 2023), Black pupils (Botelho et al., 2015), or those perceived low-caste (Hanna and Linden, 2012). Exceptions are Terrier (2020), who finds a negative bias against boys only in maths and not in French, and Black and de New (2020), who find a negative bias against overweight pupils only in maths and not in literacy. It is worth noting that neither of the latter papers find a positive bias in the non-maths subject.

strong assumption that both assessment methods measure the same skills. For example, many studies compare examinations intended to provide a "snapshot" of pupils' performance in examination settings with teacher judgements which consider pupils' written, practical, and oral classwork over an entire academic year (Gibbons and Chevalier, 2008). Exploiting an exogenous change in national assessment methods as we do has, to our knowledge, not been done before, but does offer unique benefits. The teachers providing non-blind assessments in 2020 were asked to report the grade that they predicted their pupils would achieve had they taken blindly marked examinations. Although we cannot be certain that teachers were able to entirely disregard information they had about students that was irrelevant to the prediction of their exam grade, including information gathered both before and after school closures, this approach provides a level of comparability at least in intended outcome which is arguably greater than in some of the existing studies.

Third, this paper contributes to a literature which specifically examines systematic differences in teachers' assessments of pupils from minority or non-native ethnic groups. Measured effects tend to depend on the context being studied.[2] In England, ethnic minority pupils aged seven to 14 have consistently been found to receive lower teacher assessed grades than examination grades in English, and either similar or higher teacher assessed grades than examination grades in maths (Burgess and Greaves, 2013; Campbell, 2015; Gibbons and Chevalier, 2008).[3] However, the evidence regarding pupils older than age fourteen is scarce.[4] This is notable because ethnic minority pupils in England generally attain more highly relative to White British pupils as they get older (Dustmann et al., 2010), making it likely that teachers' perceptions of their skills and knowledge will also change. This paper, by focusing on teachers' predictions of ethnic minority pupils' performance across compulsory

[2] Alesina et al. (2018), for example, find evidence of a negative grading bias for age 14 immigrant pupils in Italy, while Van Ewijk (2011) finds no evidence that pupil ethnicity directly affects the grades that teachers give age 11 pupils in the Netherlands, rather affecting teachers' expectations of those pupils.

[3] Comparing teacher assessments with examination grades at age seven, Campbell (2015) finds that all non-White groups in her sample receive lower teacher assessed grades in reading but similar teacher assessed and examination grades in maths. Comparing teacher assessments with examination grades at age 11, Burgess and Greaves (2013) find that pupils from Pakistani, Black African, and Black Caribbean backgrounds are approximately twice as likely to receive lower teacher assessment grades than White British pupils in English, whereas for maths these likelihoods are much more similar to White British pupils for nearly all groups. Comparing teacher assessments with examination grades at age fourteen, Gibbons and Chevalier (2008) find that pupils from all ethnic minority groups have significantly lower teacher assessments and higher examination grades in English, but higher teacher assessments and lower examination grades in maths.

[4] One exception is Murphy and Wyness (2020), who compare teacher predictions with examination grades that pupils receive at age 18, and find that Black and Asian pupils are more likely to be over-predicted than White British pupils. However due to data restrictions the authors are unable to differentiate between the subjects that the pupils study, which, at age 18, are entirely determined by pupils' preferences.

subjects in high-stakes examinations taken at age 16, therefore adds to a growing picture regarding the attainment and experiences of ethnic minority pupils as they progress through schooling in England. Predicted grades also have particular policy relevance in the setting studied; as teacher predictions of pupil examination performance are used by both further and higher education providers in England to compare applicants, our findings can also inform ongoing, national policy debates about current pupil assessment and university admissions procedures (Department for Education, 2022).

The paper proceeds as follows. Section 1.2 provides information on the institutional context. Section 1.3 outlines the data and sample used. Section 1.4 documents ethnic minority attainment gaps and how they changed in 2020. Section 1.5 explores the role of time trends, while Section 1.6 uses decomposition analysis to explore the extent to which differences between the cohorts may have driven the patterns in ethnicity grade gap changes. Section 1.7 investigates whether ceiling effects matter for our findings. Section 1.8 provides robustness checks. Section 1.9 concludes.

1.2 Institutional context

1.2.1 School assessments in England

Pupils in England take examinations for the General Certificate of Secondary Education (GCSEs) at the end of full-time compulsory schooling, the summer of the year in which they turn 16. Pupils usually take examinations in eight or nine subjects, with maths and English (and science) as compulsory.[5] The examinations are graded from one (low) to nine (high), with any scores below a one awarded a 'U' for 'ungraded'. Grade boundaries are set by a national regulatory body once the distribution of raw marks is known so that the proportion of pupils achieving each grade is roughly comparable between years. School-level averages are then reported in league tables to parents although, to disincentivise schools

[5] Although science is a compulsory subject, students can choose to take double or triple science. For triple science, students study separate GCSEs in biology, chemistry and physics, while for double science these subjects are combined and students receive a grade on a different scale. The two qualifications are therefore not comparable across students and, given this fact and the fact that we do not have prior attainment measures for science at age 11, we do not use science as a main outcome. However we do provide, in Appendix Figure A.1, a graph of raw grade gaps in 2018, 2019, and 2020 for science, treating grades in double and triple science as comparable (we cannot include grades for 2017 as reformed science GCSEs began in 2018). The figure shows that the pattern of grade gap changes for science are similar to those we will show for maths in Section 1.4. As science and maths are both broadly STEM (as opposed to 'arts') subjects, this arguably expands and strengthens the case for group- and subject-specific stereotyping as a possible explanation for at least part of the grade gap changes that we observe for maths and English.

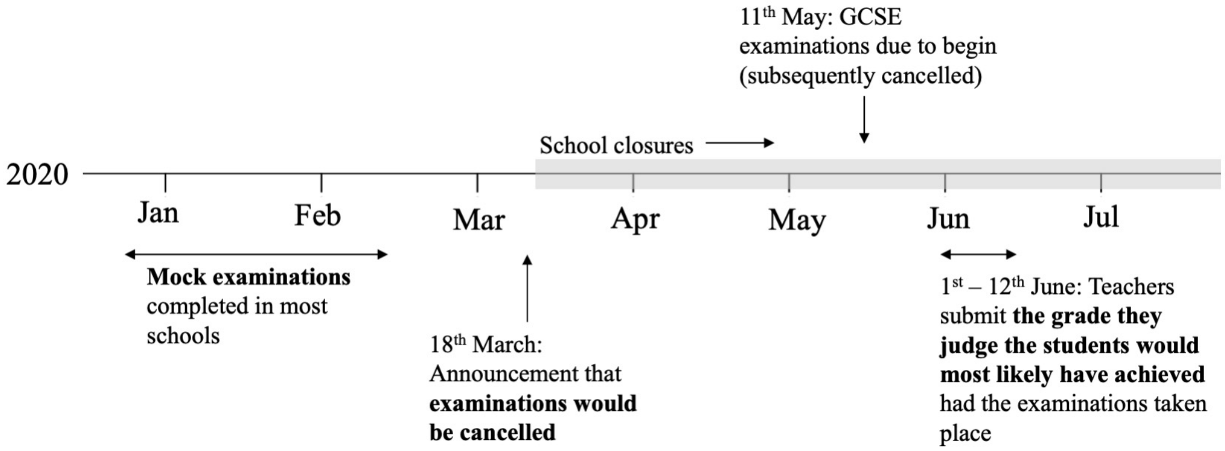
from prioritising pupils on the margin of achieving a pass grade (four), value-added measures are also given emphasis in national accountability frameworks. GCSE grades are highly determinate of the options available to pupils for post-16 education and training (Machin et al., 2020). For example most universities require applicants to have at least a grade four in English (either English Literature or English Language, whichever is higher) and maths, and any pupils who do not achieve this grade at age 16 are legally obliged to continue studying the subjects the following year.

There is little scope for grading bias in the English examination system. Grades for English Literature, English Language, and mathematics GCSE qualifications are entirely determined by pupils' performance in examinations taken at the end of a two-year course. The exam papers are externally and blindly marked by qualified teachers either from other schools or no longer in the profession, and identified by anonymous candidate numbers instead of names. Pupils' handwriting is visible to the external marker and could give away a group identity – for example if handwriting differed by gender – yet grading biases associated with handwriting have not been supported by existing evidence (Baird, 1998). It is possible for pupils to take examinations early and this could give rise to bias if it is more likely to occur for certain groups. However early entry accounts for an average of only 2.4% of entries for English Literature, English Language, and mathematics (Department for Education, 2020a), and is no more likely to occur for certain ethnic groups than for others. Schools are also able to send exam scripts to be regraded once results have been received, though this occurrence again is rare (0.05% of pupils in 2019 (Ofqual, 2020a)) and uncorrelated with pupils' demographic characteristics (Machin et al., 2020).

1.2.2 The change in assessment methods

Figure 1.1 is a timeline detailing the change in assessment methods in 2020. In England, the school year begins in September. Schools were instructed to close because of the Covid-19 pandemic on 18th March 2020. The same announcement saw the cancellation of GCSE examinations scheduled for May and June that year. On 3rd April teachers received guidance indicating that they would be required to assign grades in place of the examinations. By mid-June schools were then asked to submit, for each pupil and for each subject in which pupils were entered, the grade that they judged the pupil would most likely have received had the examinations taken place. They were also asked to submit a rank

Figure 1.1 Timeline showing details of the examination cancellations and teacher assigned grades in 2020


11th May: GCSE examinations due to begin (subsequently cancelled)
School closures
2020
Jan
Feb
Mar
Apr
May
Jun
Jul
Mock examinations completed in most schools
18th March: Announcement that **examinations would be cancelled**
1st – 12th June: Teachers submit **the grade they judge the students would most likely have achieved** had the examinations taken place

order of each pupil in each subject. The grades and rankings were accompanied by evidence, mainly comprising marks and scripts from 'mock' (practice) examinations taken by pupils prior to school closures, often in January or February (Holmes et al., 2021). Although a statistical moderation process on these grades was initially implemented by a national regulatory body, it was later discarded due to a loss of public confidence in the process. As a result, 95% of GCSE grades received by pupils in 2020 were the predicted grades assigned directly by teachers and schools, with the remaining five percent calculated through statistical modelling (Ofqual, 2020c).

Precautions were taken to ensure that the grades that pupils received in 2020 were as comparable to those of previous cohorts as possible (Ofqual, 2020d). Schools were instructed to base their judgements on existing evidence rather than any work completed by pupils after school closures, and a survey completed by teachers shortly after the grades had been submitted showed that practice ('mock') examinations – which tend to be sat in exam conditions and are marked by the teacher using official marking schemes – were indeed the primary source of evidence used to inform and support their predictions (Holmes et al., 2021).[6] Judgements made by individual teachers were also signed off by at

[6] Marks from mock examinations are kept 'on file' to use to facilitate progression to further education, training, or employment if the pupil is unable to sit the actual exam, so must be marked strictly. Teachers have little incentive to be inaccurate; though a lower mark may 'encourage' a pupil to revise more for the real exam, a lower mark would be detrimental if for any reason the mock needs to be used in place of the actual exam. In this way the grade from mock examinations should be very close to those from a real exam. It would also be reasonable to assume that any pupils who experience test anxiety are likely to be affected similarly in the mocks as in the real exams, further increasing the level of comparability between the information that teachers were using to make grade predictions and pupils' likely performance.

least two members of staff – one of which was the lead teacher of the subject within the school – and head teachers (principals) were required to submit a declaration that the grades had been generated according to the guidance. Teachers were also instructed to not discuss the grades with pupils or their families and in many cases did not continue to interact with these pupils after schools were closed and examinations cancelled, instead prioritising online learning provision for other groups (Eivers et al., 2020). We therefore assume that the 2020 grades are not influenced by differences in school closure experiences between groups, or by manipulation to appease families.

1.3 Data and sample

We use the National Pupil Database (NPD) which contains administrative data on the universe of pupils in England who attend state-funded schools. It includes demographic information about pupils, measures of their attainment at age seven, 11, 16, and 18, and information about schools and local authorities. As all state-funded schools and examination boards in England are required to return these data by law, the NPD is both accurate and highly complete. We use the cohorts of pupils in the NPD who took exams to mark the end of compulsory full-time schooling in the academic years ending in 2017, 2018, 2019 and 2020. As the change to teacher predicted grades occurred for the end of year examinations in 2020 only, we remove any grades from examinations to which pupils were entered a year early. This inevitably includes a small proportion for whom this was the only recorded grade in a given subject (about 2.5% of pupils per cohort). The resultant sample comprises 2,252,123 pupils, or about 560,000 observations for each of the four cohorts. We drop the exam results of pupils in fee-paying schools and non-mainstream schools which together comprise 6.78% of the data.[7] We then remove pupils with missing data for any of the pupil or school characteristics we consider (12.7%) or with no recorded grade for GCSE maths or either GCSE English Literature or GCSE English Language (0.800%).[8] This leaves a final sample size of 1,823,542 pupils in 3,380 schools across all four cohorts.

[7] Fee-paying schools are not required to submit Pupil Census data so we cannot observe pupil characteristics. The non-mainstream state-funded schools that we exclude are those solely for pupils with special educational needs, in hospital, who have been removed from mainstream education due to behavioural concerns, or who have been given custodial sentences. Low proportions of these students are entered for GCSE examinations.

[8] The vast majority of missing pupil or school characteristics are prior attainment scores at age 11. Missing grades are rare, and likely to occur because pupils are entered for different types of qualification (for example for those unlikely to reach GCSE standard), or pupils being entered for the GCSE but not showing up on the day.

Our outcomes of interest are GCSE (age 16) grades in English and maths. There is one maths GCSE qualification in England. However, there are two GCSE English qualifications: English Language and English Literature. The vast majority of pupils in England take both. If their grades differ, the higher of the two is used in headline performance measures for schools and, by pupils, to meet performance benchmarks (Department for Education, 2020b). We follow these national conventions and use the higher of the two grades in our main analyses. Results for English Language and English Literature separately, as well as for the mean of the two, are presented as a robustness check.[9]

The ethnic groups that we consider are Pakistani and Bangladeshi, Indian, Black African, and Black Caribbean, which are the largest ethnic minority groups in England and Wales, and we compare them to White British students (Office for National Statistics, 2022).[10] The additional pupil-level characteristics we consider in the decomposition analyses are gender, whether the pupil has been identified as having a special educational need (SEN), whether the pupil speaks English as an additional language (EAL), and pupils' attainment at age 11 (KS2 subject-specific point scores), which we standardise by year and subject.[11] As proxies for pupils' socio-economic status, we use eligibility for free school meals at any point in the last six years (FSM6) and a rank based on the proportion of all children aged zero to 15 living in income deprived households in the pupils' local area of residence (IDACI score). At the school-level, we consider a number of characteristics including the type of school, whether the school has selective admissions, and the school region. School quality is proxied by a measure of school value-added from the end of primary school to examinations at age 16 in eight prescribed subjects (called 'Progress 8').

We present descriptive statistics for our sample, separately for each year between 2017 and 2020, in Appendix Table A.1, and separately for each ethnic group across years in Appendix Table A.2.

[9] In June 2020, 567,277 and 531,626 16-year-old pupils received English Language and English Literature grades respectively, which is similar in proportion to 2019 for which the figures are 546,607 and 514,191 (Joint Council for Qualifications, 2020). The overall increase in entries is due to the 2020 cohort being approximately 3% larger than that of 2019 (Ofqual, 2020b). 5.4% of pupils were entered for English Literature a year early in 2019, which explains much of the discrepancy between the subjects (Department for Education, 2020a). As our sample is restricted to the end of year examinations in the year in which pupils turn 16, such 'early entries' are dropped. However these differences are not likely to cause any differences in results between English Literature and English Language. First, we are primarily interested in changes between 2019 and 2020, and patterns of entries appear consistent between these years (Joint Council for Qualifications, 2020). Second, early entrants in English Literature are proportionate to the full sample with regards to ethnicity.

[10] In our analyses we also include indicators for 'Multiethnic' and 'Any Other Ethnic Group' but do not report or interpret results for these groups as each combines a very diverse group of pupils.

[11] For KS2 subject-specific point scores, we use the total marks in the KS2 maths tests for maths, and the marks in the KS2 English reading test for English. All of these tests are blindly and externally marked.

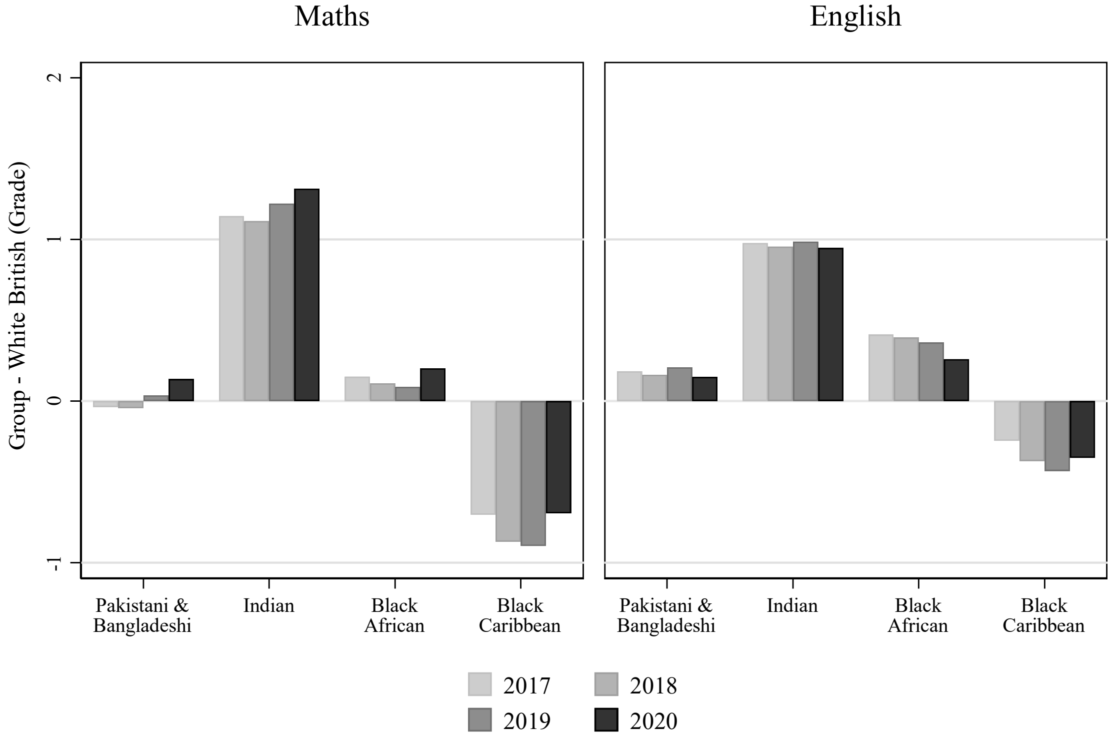
Appendix Table A.2 shows that the pupils in our sample who are members of our ethnic minority groups of focus live, on average, in more deprived neighbourhoods than White British pupils, and are more likely to speak English as an additional language. Compared to the other ethnic minority groups, Indian pupils are also more likely to attend schools with selective admissions, and Black Caribbean pupils both less likely to speak English as an additional language and more likely to be diagnosed with a special educational need.

1.4 Ethnicity grade gaps

We begin by documenting raw ethnicity grade gaps at age 16 in England and how they change over time. Figure 1.2 shows raw grade gaps for the four ethnic minority groups in 2017, 2018, 2019 and 2020. For each year, the bar shows the difference in the average grade that pupils received in relation to the reference group, White British pupils, in that year. Pakistani and Bangladeshi, Indian, and Black African pupils generally attain more highly than White British pupils in both English and maths, while Black Caribbean pupils attain lower grades in both subjects. Of particular interest are the differences in the 2020 gaps compared to the same differences in the preceding years. In maths, all ethnic minority groups received higher grades relative to White British pupils than they did in the preceding years, whereas in English all apart from Black Caribbean pupils received lower relative grades than in the preceding years. These grade gaps by year and subject are also reported in Panel A of Table 1.1. We will show in Section 6 that once we control for changes across cohorts in the levels of, and returns to, observable characteristics, grade gap changes in English for the Black Caribbean group become more aligned with those of other ethnic minority groups.

The first row of Panel B in Table 1.1 reports the difference between the grade gap of each ethnic minority group and White British pupils in 2020 and in 2019 (i.e. the difference between the fourth and third row of Panel A of Table 1.1), alongside standard errors on the difference. We call these double differences the ‘ethnicity grade gap changes’. For every subject and ethnic minority group, these changes measure how the grade gap between the ethnic minority group and White British pupils in 2020 (when grades are assigned by teachers) compares to that same gap in 2019, when grades are assigned through examinations marked by external examiners. A positive change indicates that,

Figure 1.2 Grade gaps by ethnic group, year and subject



Notes: National Pupil Database, 2017 - 2020. Sample comprises age 16 pupils in mainstream schools in England with no missing data. Exam grades are the May and June (end of year) examinations only. The English grade is the highest of English Literature and English Language if pupils received grades for both.

compared to White British pupils, an ethnic minority group receives relatively higher grades when grades as assigned by teachers than through examinations marked by external examiners; a negative change indicates that, compared to White British pupils, an ethnic minority group is given relatively lower grades when grades are assigned by teachers than through examinations marked by external examiners.

For maths, the raw ethnicity grade gap changes are positive for all ethnic minority groups. For Pakistani and Bangladeshi, Indian, and Black African pupils, who all achieve more highly relative to White British pupils in 2019, the relative grade gap changes for maths between 2019 and 2020 are 10.1%, 9.1%, and 11.5% of a grade respectively. For Black Caribbean pupils, who achieve lower than White British pupils in prior years, the relative grade gap change in 2020 is 20.1%. In other words, ethnic minority pupils received relatively higher maths grades than White British pupils in 2020 when the grades were assigned by teachers. The raw ethnicity grade gap changes for English are less consistent. For Pakistani and Bangladeshi and Black African pupils they are negative, at -6%

and -10.3% of a grade respectively, and significant at the one percent level. Pakistani and Bangladeshi and Black African pupils therefore received relatively lower grades than White British pupils in 2020 compared to 2019 in English. The point estimate for Indian pupils is negative (-4.1%) though not statistically significant. For Black Caribbean pupils it is positive (8.2%).

Table 1.1 Grade gaps by ethnic group, year and subject

	Maths				**English**			
	P&B	I	BA	BC	P&B	I	BA	BC
Panel A: Raw ethnicity grade gaps								
Group - White British:								
2017	-0.037	1.145***	0.153***	-0.703***	0.185***	0.978***	0.412***	-0.247***
	(0.035)	(0.054)	(0.032)	(0.038)	(0.033)	(0.041)	(0.032)	(0.037)
2018	-0.044	1.114***	0.110***	-0.870***	0.164***	0.956***	0.393***	-0.371***
	(0.036)	(0.052)	(0.032)	(0.039)	(0.034)	(0.044)	(0.032)	(0.037)
2019	0.036	1.222***	0.089***	-0.895***	0.210***	0.988***	0.363***	-0.432***
	(0.037)	(0.052)	(0.032)	(0.035)	(0.035)	(0.044)	(0.031)	(0.036)
2020	0.137***	1.314***	0.204***	-0.694***	0.149***	0.947***	0.260***	-0.350***
	(0.034)	(0.048)	(0.028)	(0.031)	(0.032)	(0.039)	(0.027)	(0.030)
N	1,823,542				1,823,542			
Panel B: Ethnicity grade gap changes								
Double difference:								
2020 - 2019 change	0.101***	0.091***	0.115***	0.201***	-0.060***	-0.041	-0.103***	0.082**
	(0.023)	(0.028)	(0.024)	(0.035)	(0.020)	(0.028)	(0.023)	(0.036)
N	934,590				934,590			
Panel C: Predicted ethnicity grade gaps								
Predicted 2020	0.059	1.239	0.053	-1.015	0.212	0.984	0.340	-0.534
Actual - predicted 2020	0.078**	0.075	0.151***	0.320***	-0.062**	-0.037	-0.080***	0.184***
	(0.034)	(0.048)	(0.028)	(0.031)	(0.032)	(0.039)	(0.027)	(0.030)
N	1,823,542				1,823,542			

Notes: National Pupil Database, 2017 - 2020. ‘P&B’ stands for Pakistani and Bangladeshi, ‘I’ Indian, ‘BA’ Black African, ‘BC’ Black Caribbean. Sample comprises age 16 pupils in mainstream schools in England with no missing data. Exam grades are the May and June (end of year) examinations only. The English grade is the highest of English Literature and English Language if pupils received grades for both. The reference group is White British. Standard errors in parentheses, clustered at the school-level. * Significant at 10%; ** significant at 5%; *** significant at 1%.

The pattern of higher relative grades in maths and lower relative grades in English is consistent with the findings of the extant research into ethnic differences in teacher assessment and examination grades in England for younger pupils; ethnic minority pupils tend to receive lower teacher assessed grades than examination grades in English and either similar or higher teacher assessed grades than examination grades than in maths (Burgess and Greaves, 2013; Campbell, 2015; Gibbons and Chevalier, 2008). The size of the estimated grade change is not negligible. Using the standard deviations for 2020 (see Appendix Table A.5), a 10% grade change in either subject is equivalent to an effect size of approximately 0.05, which is the same as the average effect size of school-based interventions aimed at improving academic achievement in developed countries (Fryer Jr, 2017).

1.5 The role of time trends

We want to ensure that the gap changes presented in Panel B of Table 1.1 are driven by the switch to teacher assigned grades, rather than the results of pre-existing trends. Therefore, we explore the extent to which the ethnicity grade gap changes in 2020 differed from trends in the grades received by ethnic minority pupils relative to White British pupils over time. We first regress each subject grade on an indicator for ethnicity, year, and their interaction, excluding the year 2020. This estimates separate trends in grades by subject and ethnic group for the years in which grades were assigned only through blindly-marked examinations.[12] We then use the coefficients estimated from this model to predict grades (and thereby grade gaps) in 2020. Finally, we test whether or not these predicted 2020 grade gaps are different from those observed in 2020 (fourth row of Panel A, Table 1.1).

We present the results in Panel C of Table 1.1, separately for maths and English. The first row of the panel shows the 2020 grade gap predicted based on a time trend, and the second shows the differences between the actual 2020 grade gap and the predicted grade gap. These differences are the raw ethnicity grade gaps in 2020 net of a time trend. Comparing the differences to the observed ethnicity grade gap changes in the first row of Panel B of Table 1.1 suggests no clear pattern. Accounting for a time trend in some cases exacerbates the gaps for some ethnic minority pupils, compared to White British pupils, and in other cases the gap remains unchanged or is attenuated. In the latter cases (Pakistani and Bangladeshi pupils in maths and Black African pupils in English) the trends predict only about 20% of the actual grade gap changes that we observe, indicating that time trends do not explain the ethnicity grade gap changes we observe.

However, grade gaps in 2020 could differ because returns to characteristics other than ethnicity are different in 2020 compared to earlier years, due to the change to teacher assessment. For example, a pupil characteristic such as eligibility for free school meals may have greater returns for attainment in one subject in 2020 compared to 2019. If this characteristic is more greatly represented in a specific

[12] We make Y_i, that is a pupil's grade, depend on a set of observed characteristics as follows:

$$Y_i = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 year_i + \alpha_3 [year_i \times X_{1i}] + \eta_i$$

where X_{1i} contains dummy variables indicating whether the pupil is in any of the ethnic minority groups and $year$ is a continuous variable. We then use this model to estimate the predicted grades for all ethnic groups in the year 2020. As the grades in 2020 are based on a naive prediction, there is no variance in these predictions by group and year and we do not report a standard error.

ethnic group, then this group will see a greater relative grade increase in 2020 due to factors other than ethnicity. Understanding the role of these factors is the focus of the next section.

1.6 The role of differences between cohorts

In this section we assess whether changes in pupil characteristics and their returns across years explain the grade gap changes we documented, and, if so, which characteristics mostly contribute to the results. To do so, we decompose the changes in performance gaps between 2019 and 2020 into an 'explained part', namely the part explained by changes in returns to observed characteristics or changes in their prevalence by group, and an 'unexplained' part. We use an extension of the standard Gelbach decomposition (Gelbach, 2016) that allows returns to characteristics to differ by year and, unlike the Oaxaca-Blinder decomposition (Blinder, 1973; Oaxaca, 1973), yields results that are order-invariant, that is, they do not depend on the order in which observed characteristics are added to the model. We allow both the levels of, and the returns to, observed characteristics other than ethnic group to differ across the years. We do not allow the returns to observed characteristics to differ by ethnic group as we want the 'explained' gap change to capture the part of the grade gap changes which can be accounted for by factors other than ethnicity and ethnicity-based stereotyping. The effects of ethnicity and stereotyping, including changes in the returns to observed characteristics that differ by ethnic group as well as other teacher stereotyping effects acting through variables that are not observed in our model, are then contained in the 'unexplained' part.

1.6.1 Decomposition methodology

We start by formalising the *raw* ethnicity grade gap changes as:

$$Y_i = \beta_0^{raw} + \beta_1^{raw}\mathbf{X}_{1i} + \beta_2^{raw}2020_i + \beta_3^{raw}[2020_i \times \mathbf{X}_{1i}] + \mu_i \tag{1.1}$$

where Y_i indicates pupil i's grades, $\mathbf{X}_{1i}$ is a vector of indicators of whether pupil i is in any of the ethnic minority groups, 2020_i is an indicator for whether i is in the 2020 cohort, $[2020_i \times \mathbf{X}_{1i}]$ contains the interactions between the 2020_i indicator and the variables in $\mathbf{X}_{1i}$, and μ_i is an error. Here β_0^{raw} is the mean grade achieved by White British pupils in 2019, β_1^{raw} is the vector of gaps in the mean grade of each ethnic minority group compared to White British pupils in 2019, β_2^{raw} is the partial effect of the year 2020 on the mean grade for White British pupils, and β_3^{raw} is a vector of the raw ethnicity grade gap changes in 2020 compared to 2019 (the same gap changes we show in Panel B of Table 1).

To derive the *unexplained* ethnicity grade gap changes we can estimate the following equation:

$$Y_i = \beta_0 + \beta_1 \mathbf{X}_{1i} + \beta_2 2020_i + \beta_3[2020_i \times \mathbf{X}_{1i}] + \beta_4 \mathbf{X}_{2i} + \beta_5[2020_i \times \mathbf{X}_{2i}] + \varepsilon_i \qquad (1.2)$$

where Y_i, $\mathbf{X}_{1i}$, 2020_i are as defined above, and $\mathbf{X}_{2i}$ contains observable individual and school characteristics of pupil i including gender, family and neighbourhood deprivation (SES), whether the pupil has a special educational need, whether the pupil speaks English as their first language, the pupils' subject-specific prior attainment at age 11, a number of school characteristics, and a measure of school quality, $[2020_i \times \mathbf{X}_{2i}]$ is the interaction between $\mathbf{X}_{2i}$ and the 2020_i indicator, and ε_i is an error term. β_3 is the vector of the unexplained ethnicity grade gap changes, that is what is left of the raw ethnicity grade gap changes after controlling for the observable characteristics in $\mathbf{X}_{2i}$ and their interactions with 2020_i, β_4 is the vector of returns to $\mathbf{X}_{2i}$ on Y_i in 2019, and β_5 is the vector of the changes in the returns to $\mathbf{X}_{2i}$ on Y_i in 2020 compared to 2019.

The difference between the raw and the conditional (unexplained) ethnicity grade gap changes ($\beta_3^{raw} - \beta_3$) are the *explained* grade gap changes, which are contained in the vector δ. The explained grade gap changes combine the effects of changes in the distribution of observed characteristics across ethnic groups in 2020 compared to 2019, and the effects of changes in the returns to these characteristics in 2020 compared to 2019. We can examine these combined effects and how they differ for different pupil characteristics contained in $\mathbf{X}_2$ by implementing an extension of the standard decomposition proposed by Gelbach (2016) in which we allow the returns to the observed characteristics to differ by year.[13] To do this, for each characteristic k in $\mathbf{X}_2$, we estimate the differences in the levels (means) of that characteristic for each ethnic group compared to White British pupils in 2019 ($\hat{\boldsymbol{\Gamma}}^k_{2019,eth}$) and

[13] See in particular Gelbach (2016) Sections IV and V.B.

the changes in those differences in levels in 2020 ($\hat{\mathbf{\Gamma}}^k_{2020,eth}$) using a set of auxiliary regressions.[14] We then use $\hat{\mathbf{\Gamma}}^k_{2019,eth}$ and $\hat{\mathbf{\Gamma}}^k_{2020,eth}$ as well as coefficients from equation (1.2) to derive the vector of explained gap changes, $\hat{\delta}$, as follows:[15]

$$\hat{\delta} = \hat{\beta}_3^{raw} - \hat{\beta}_3 = \sum_{k=1}^{K} [\hat{\mathbf{\Gamma}}^k_{2020,eth}\hat{\beta}_4^k + \hat{\mathbf{\Gamma}}^k_{2019,eth}\hat{\beta}_5^k + \hat{\mathbf{\Gamma}}^k_{2020,eth}\hat{\beta}_5^k] \tag{1.3}$$

where $\hat{\delta}$ is the explained part of $\hat{\beta}_3^{raw}$ made up of the sum of $\hat{\mathbf{\Gamma}}^k_{2020,eth}\hat{\beta}_4^k$ (the changes in the differences in the levels of k in 2020 multiplied by the returns to k in 2019), $\hat{\mathbf{\Gamma}}^k_{2019,eth}\hat{\beta}_5^k$ (the differences in the levels of k in 2019 multiplied by the changes in the returns to k in 2020), and $\hat{\mathbf{\Gamma}}^k_{2020,eth}\hat{\beta}_5^k$ (the changes in the differences in the levels of k in 2020 multiplied by the changes in the returns to k in 2020).

$\hat{\delta}$ can then be split into components by ethnic group and pupil characteristic. Consider for example the contributions of the gender ('male') and socio-economic status ('SES') characteristics to the change in the grade gap between Pakistani and Bangladeshi ('PB') and White British pupils in 2020 compared to 2019. They can be written as follows:

$$\text{Male component: } \hat{\delta}_{PB}^{male} = \hat{\mathbf{\Gamma}}^{male}_{2020,PB}\hat{\beta}_4^{male} + \hat{\mathbf{\Gamma}}^{male}_{2019,PB}\hat{\beta}_5^{male} + \hat{\mathbf{\Gamma}}^{male}_{2020,PB}\hat{\beta}_5^{male}$$

$$\text{SES component: } \hat{\delta}_{PB}^{SES} = \hat{\mathbf{\Gamma}}^{SES}_{2020,PB}\hat{\beta}_4^{SES} + \hat{\mathbf{\Gamma}}^{SES}_{2019,PB}\hat{\beta}_5^{SES} + \hat{\mathbf{\Gamma}}^{SES}_{2020,PB}\hat{\beta}_5^{SES}$$

Here the male (SES) component shows how changes between 2019 and 2020 in the proportion of Pakistani and Bangladeshi compared to White British males (low SES pupils) and in the returns to being male (a low SES student) contribute to the observed changes in the ethnicity grade gaps in 2020 compared to 2019 in the absence of ethnicity-based stereotyping, that is when changes in the returns to the characteristics do not differ by ethnic group.

[14] $\hat{\mathbf{\Gamma}}^k_{2019,eth}$ and $\hat{\mathbf{\Gamma}}^k_{2020,eth}$ can be estimated using a set of auxiliary models with each of the k characteristics in $\mathbf{X}_2$ acting as the dependent variable such that

$$\mathbf{X}^k_{2i} = \mathbf{\Gamma}^k_{2019} + \mathbf{\Gamma}^k_{2019,eth}\mathbf{X}_{1i} + \mathbf{\Gamma}^k_{2020}2020_i + \mathbf{\Gamma}^k_{2020,eth}[2020_i \times \mathbf{X}_{1i}] + w_i^k$$

where k represents one of the characteristics in $\mathbf{X}_2$, and w_i^k is a residual.

[15] We implement the decomposition using the 'b1x2' Stata command based on Gelbach (2016).

1.6.2 Decomposition results

Table 1.2 reports the results for the decomposition of the changes in the grade gaps between each ethnic minority group and White British pupils in 2020 compared to 2019. The first row shows the raw ethnicity grade gap changes and is identical to that reported in the second panel of Table 1.1. The second row shows estimates for $\hat{\delta}$, that is the part of the ethnicity grade gap changes which are explained by differences in the levels of, and returns to, observed characteristics between the years. In other words, it shows the ethnicity grade gap changes that we would have expected in 2020, given observed changes in the characteristics and their returns between years (and assuming no changes in the differences in returns to these characteristics by ethnic group). Below this, Table 1.2 shows the separate contributions of different characteristics or groups of characteristics to this overall explained component.[16] Finally, the bottom row shows contributions to the ethnicity grade gap changes which remain unexplained. These include changes in the returns to observed characteristics that differ by ethnic group as well as any factors unobserved in our model, including teacher stereotyping effects acting through variables that are not observed. The estimated returns can be found in Appendix Table A.4.

The decomposition results for maths are reported in the four columns on the left of Table 1.2. Changes in the levels of and returns to pupils' socio-economic status and diagnoses of special educational needs across groups and years contribute to explaining the ethnicity grade gap changes for most ethnic minority groups, but to a relatively small extent.[17] Note that the estimated contributions of subject-specific previous attainment at age 11 do not reach statistical significance (see seventh row of Table 1.2). In fact, in the case of Indian and Black African pupils, the point estimates of these contributions are negative, in contrast with the positive raw gap changes these contributions should help explain.

Taken together, for Pakistani and Bangladeshi pupils, changes in the observed characteristics and their returns explain 6.5% of a grade (see second row of Table 1.2), that is around two thirds of the 10% raw maths grade gap change, with only 3.6% of a grade left unexplained (see bottom panel of Table 1.2). For Black Caribbean pupils, the observed characteristics and their returns explain 5.6% of

[16] Each of the individual characteristic contributions separately can be found in Appendix Table A.3.

[17] Changes in school characteristics and being male contribute positively to the ethnicity grade gap changes for Black Caribbean pupils, and in school value-added for Pakistani and Bangladeshi pupils.

Table 1.2 Gelbach decomposition of the change in grade gaps by ethnic group

	Maths				English			
	P&B	I	BA	BC	P&B	I	BA	BC
Raw gap change	0.101***	0.091***	0.115***	0.201***	-0.060***	-0.041	-0.103***	0.082**
	(0.023)	(0.028)	(0.024)	(0.035)	(0.020)	(0.028)	(0.023)	(0.036)
Explained gap change	0.065***	-0.009	0.018	0.056*	0.080***	0.036*	0.075***	0.104***
	(0.022)	(0.024)	(0.021)	(0.030)	(0.019)	(0.022)	(0.020)	(0.025)
Amount explained by:								
Male	0.001	0.000	0.002	0.003*	0.003	0.004	-0.004	0.005
	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)	(0.004)	(0.004)	(0.005)
Socio-economic status	0.006	0.012***	0.015***	0.006	0.015***	0.011***	0.031***	0.020***
	(0.004)	(0.003)	(0.005)	(0.006)	(0.004)	(0.003)	(0.006)	(0.006)
Special educational needs	0.004***	0.007***	0.004***	-0.003	0.005***	0.007***	0.006***	0.000
	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	(0.004)
First language not English	0.011	0.008	0.005	0.000	0.011	0.009	0.006	0.000
	(0.009)	(0.008)	(0.007)	(0.001)	(0.010)	(0.008)	(0.007)	(0.001)
Prior attainment (age 11)	0.024	-0.026	-0.015	0.027	0.002	0.036***	0.030***	0.047***
	(0.016)	(0.019)	(0.016)	(0.026)	(0.010)	(0.014)	(0.012)	(0.016)
School characteristics	0.004	-0.008	0.014	0.018*	0.031***	-0.023***	0.018	0.029**
	(0.008)	(0.007)	(0.010)	(0.011)	(0.009)	(0.008)	(0.011)	(0.012)
School value-added (lagged)	0.015*	-0.003	-0.008	0.004	0.012	-0.008	-0.011	0.003
	(0.009)	(0.009)	(0.008)	(0.010)	(0.009)	(0.010)	(0.008)	(0.010)
Unexplained gap change	0.036*	0.100***	0.096***	0.145***	-0.140***	-0.077***	-0.179***	-0.022
	(0.019)	(0.020)	(0.019)	(0.026)	(0.020)	(0.025)	(0.021)	(0.030)
N	934,590				934,590			

Notes: National Pupil Database, 2017 - 2020. 'P&B' stands for Pakistani and Bangladeshi, 'I' Indian, 'BA' Black African, 'BC' Black Caribbean. Sample comprises age 16 pupils in mainstream schools in England with no missing data. Exam grades are the May and June (end of year) examinations only. The English grade is the highest of English Literature and English Language if pupils received grades for both. The gap is the gap in average grades between the ethnic minority group and White British pupils. Socio-economic status includes free school meals indicator and rank based on the proportion of all children aged 0 to 15 living in income deprived households in the pupil's local area of residence (IDACI score). Prior attainment (age 11) includes subject-specific age 11 attainment score, standardised by year. School characteristics are indicators for selectivity, urban, region (9 categories), and school governance type (4 categories), and continuous measures of cohort size, proportion of pupils eligible for free school meals, and average neighbourhood deprivation (IDACI) score of the school. School value-added refers to how many average grades higher or lower the pupils in that school achieve across eight qualifying GCSE subjects compared to pupils across the country who score comparatively at age 11. Standard errors in parentheses, clustered at the school level. * Significant at 10%; ** significant at 5%; *** significant at 1%.

a grade, that is less than a third of the over 20% of a grade raw maths grade gap change, leaving 14.5% of a grade unexplained. For Indian and Black African pupils, changes in the observed characteristics and their returns explain very little of the raw maths grade gap change, as shown by the small and statistically insignificant estimates of the explained gap changes (second row of Table 1.2). As a consequence, a considerable proportion of the raw grade gap changes are left unexplained, as we can see by comparing the top and the bottom panel of the second and third columns of Table 1.2.

The decomposition results for English are reported in the four columns on the right of Table 1.2. Pupils' socio-economic status, diagnoses of special educational needs, and school characteristics all contribute to explaining the ethnicity grade gap changes. However, for most ethnic groups, pupil prior attainment (row 7) appears to be the most important explanatory factor. This is particularly relevant in the case of Black Caribbean pupils, where changes in prior attainment and their returns explain more than half of the observed grade gap change.[18] The second row of Table 1.2 shows the explained

[18] The main characteristic that is different for the 2020 Black Caribbean cohort is a large increase in English prior attainment, both compared to the other cohorts and to other ethnic minority groups. We can see from Appendix Table

components of the grade gap, summarising the contribution of all the observed characteristics. These components are positive, larger than those estimated for maths and statistically significant for all ethnic groups. These results indicate that, given the changes in the groups' characteristics and the returns to those characteristics in 2020 compared to 2019, especially in prior attainment, and in absence of differential changes in returns by ethnic group, we would have expected all ethnic minority pupils to have achieved more highly in 2020 than in 2019 relative to White British pupils. We would have expected Pakistani and Bangladeshi pupils to have achieved relatively more highly by 8% of a grade, Indian pupils by 3.6% of a grade, Black African pupils by 7.5% of a grade, and Black Caribbean pupils by 10.4% of a grade. It is therefore noteworthy that instead we see negative raw grade gap changes for most groups, showing that the opposite was in fact the case – they achieved relatively lower.

The bottom panel of Table 1.2 reports the unexplained gap changes, which are the differences between the ethnicity grade gap changes in English predicted by changes in the observed characteristics and their returns and the raw ethnicity grade gap changes. These unexplained components of the English grade gap changes are negative for all ethnic groups. For Pakistani and Bangladeshi, Indian and Black African pupils, the unexplained components of the English grade gap changes are highly statistically significant, and around double in magnitude than the raw grade gap changes (compare the top and the bottom panel of Table 1.2). This suggests that in 2020, when grades were assigned by teachers, these ethnic minority groups received lower grades in English than White British pupils when compared to the previous year despite the 2020 cohort having characteristics (previous attainment in particular) that, in absence of differential changes in returns by ethnic group or unobserved factors, would have predicted an improvement in their relative performance. Unlike for these groups, the raw 2020 grade gap change in English for Black Caribbean pupils was positive (see the last column of the top row of Table 1.2). However our decomposition shows that this is entirely driven by changes in observed characteristics and their returns. Once we control for these, the point estimate of the unexplained gap change in English for the Black Caribbean group becomes negative – although not statistically significant – and thus aligns more with the other groups (see bottom panel of Table 3).

In summary, Table 1.2 shows that changes in observed characteristics and their returns explain part of why ethnic minority pupils received relatively higher grades in maths in 2020, though a substantial

A.4 that in English prior attainment is more highly rewarded in 2020 than in 2019, and this has a large effect on Black Caribbean students in 2020.

proportion of the raw grade gap changes remain unexplained. Changes in observed characteristics and their returns fully explain why Black Caribbean pupils received relatively higher grades in English. However, they do not explain why Pakistani and Bangladeshi, Indian and Black African pupils received relatively lower grades in English. In fact, our results suggest that had those changes not occurred, the relative drop in the English performance that followed the switch to teachers' assessment would have been roughly double in magnitude for these ethnic minority groups. Note that the direction of the effects of the explained contributions are roughly consistent across both maths and English, so we would not expect any factors that we are not able to observe to be causing the contrasting direction of the unexplained grade gap changes. We conclude that there is evidence suggesting that unexplained factors, differing by subject and likely including teacher stereotyping, had a role in determining the 2020 GCSE results.

1.7 The role of ceiling effects

Another way that the shift to teacher assessments might have indirectly affected the relative performance gaps across ethnic groups is through ceiling effects. Ceiling effects, a type of scale attenuation effect, are observed when there is an upper limit to the dependent variable. In our case this is the GCSE upper limit of a grade 9. Ceiling effects may in principle be relevant in our case as, as shown in Table 1.3, the grade averages for both maths and English in 2020 were considerably higher than in the preceding years. Higher grades in 2020 therefore resulted in a smaller spread of grades overall (as evidenced by slightly smaller standard deviations), likely to be driven by pupils at the top end of the grade distribution having less growth potential than those lower down. Such ceiling effects may affect changes in grade gaps between ethnic groups which are positioned differently, on average, within the overall distribution of grades. Most of our ethnic minority groups of focus generally attain more highly than their White British peers in both subjects (see Table 1.1). As the relative grades of ethnic minority pupils *improve* in maths in 2020, it seems unlikely that ceiling effects drive the observed ethnicity grade gap changes in maths. However the grades for English are on average higher than those for maths. It is therefore possible that ceiling effects are driving some of our results for English in a way which they are not for maths, despite most of the ethnic minority groups achieving more highly than White British pupils, in ‘normal’ years, in both subjects. Mean grades broken down by subject, group, and year, can be found in Appendix Table A.5.

Table 1.3 Mean grades by subject and year

	Maths	English
2017	4.71	5.23
	(2.03)	(1.90)
2018	4.75	5.22
	(2.03)	(1.90)
2019	4.77	5.22
	(2.03)	(1.90)
2020	5.06	5.40
	(1.99)	(1.82)
N	1,823,542	1,823,542

Notes: National Pupil Database, 2017 - 2020. Sample comprises age 16 pupils in mainstream schools in England with no missing data. Exam grades are the May and June (end of year) examinations only. The English grade is the highest of English Literature and English Language if pupils received grades for both. Standard deviations in parentheses.

Following Murphy and Wyness (2020), we check the contribution of ceiling effects to the observed ethnicity grade gap changes by partitioning the sample into three equally-sized groups according to pupils’ prior attainment at age 11 in the same subject. We then repeat the decomposition of the ethnicity grade gap changes for each of these three groups. The results are shown in Table 1.4. If ceiling effects were driving our results, we would expect to see only the highest attaining ethnic minority pupils receiving lower relative grades in 2020, and potentially more so in English than in maths. Contrary to this, Table 1.4 shows the raw, explained, and unexplained components of the grade gap changes for each of the three attainment groups, and reveals that the patterns presented above remain across them. This shows that there remain unexplained factors, such as stereotyping effects, which are contributing to the observed ethnicity grade gap changes throughout the distributions of grades, and suggests ceiling effects are not driving our results.

1.8 Robustness checks

We check that the decomposition results are robust to multiple alternative specifications. For simplicity, Table 1.5 shows just the raw, explained, and unexplained components from the decomposition results from each of these sensitivity analyses. In Panel A, we account for possible nonlinearities in the association between pupil and school characteristics with outcomes. The results are consistent with those presented above. In Panel B, we expand our decomposition reference group from the year 2019

Table 1.4 Gelbach decomposition by ethnic and prior attainment groups

	Maths				**English**			
	P&B	I	BA	BC	P&B	I	BA	BC
Low prior attaining								
Raw gap change	0.088***	0.123***	0.177***	0.229***	-0.036	-0.051	-0.115***	0.101**
	(0.026)	(0.040)	(0.028)	(0.037)	(0.024)	(0.038)	(0.030)	(0.041)
Explained gap change	0.065***	0.003	0.081***	0.078***	0.073***	-0.001	0.057**	0.055**
	(0.022)	(0.026)	(0.022)	(0.025)	(0.021)	(0.024)	(0.023)	(0.025)
Unexplained gap change	0.023	0.119***	0.097***	0.151***	-0.109***	-0.050	-0.171***	0.046
	(0.026)	(0.036)	(0.027)	(0.032)	(0.028)	(0.038)	(0.029)	(0.039)
N	312,268				319,447			
Medium prior attaining								
Raw gap change	0.145***	0.181***	0.185***	0.129***	-0.041	-0.091**	-0.097***	0.041
	(0.025)	(0.036)	(0.030)	(0.046)	(0.026)	(0.035)	(0.033)	(0.049)
Explained gap change	0.062***	0.040*	0.037*	0.015	0.093***	-0.010	0.061***	0.091***
	(0.021)	(0.022)	(0.020)	(0.024)	(0.023)	(0.024)	(0.023)	(0.027)
Unexplained gap change	0.083***	0.141***	0.148***	0.115***	-0.134***	-0.081**	-0.158***	-0.050
	(0.028)	(0.036)	(0.030)	(0.044)	(0.030)	(0.038)	(0.033)	(0.046)
N	313,868				318,017			
High prior attaining								
Raw gap change	0.059**	0.075***	0.049	0.124**	-0.079***	-0.087***	-0.127***	0.046
	(0.027)	(0.028)	(0.033)	(0.063)	(0.029)	(0.034)	(0.037)	(0.070)
Explained gap change	0.011	-0.019	-0.012	-0.078**	0.085***	-0.014	0.047*	0.076**
	(0.022)	(0.021)	(0.023)	(0.034)	(0.024)	(0.025)	(0.025)	(0.034)
Unexplained gap change	0.048*	0.094***	0.061*	0.203***	-0.164***	-0.073**	-0.173***	-0.031
	(0.028)	(0.027)	(0.032)	(0.056)	(0.032)	(0.033)	(0.037)	(0.063)
N	308,454				297,126			

Notes: National Pupil Database, 2017 - 2020. ‘P&B’ stands for Pakistani and Bangladeshi, ‘I’ Indian, ‘BA’ Black African, ‘BC’ Black Caribbean. Sample comprises age 16 pupils in mainstream schools in England with no missing data. Exam grades are the May and June (end of year) examinations only. The English grade is the highest of English Literature and English Language if pupils received grades for both. The gap is the gap in average grades between the ethnic minority group and White British pupils. Explanatory characteristics are the same as those in Table 1.2. Standard errors in parentheses, clustered at the school level. * Significant at 10%; ** significant at 5%; *** significant at 1%.

alone to a combined 2018 and 2019 group.[19] The results are also in line with those in Table 1.2. In fact, when 2018 is included in the reference group, the estimated negative unexplained contributions to the English grade gap changes become even larger. Moreover, a negative unexplained contribution, statistically significant at the five percent level, is also estimated for Black Caribbean pupils in English.

In Panel C we explore how our results change when we use a dichotomous pass or fail outcome (a pass being a grade 4 and above), rather than a continuous grade, to take into account that grades are ordinal measures of attainment. The grade gap changes and unexplained contributions are smaller when we use this binary outcome, but interestingly the positive and negative unexplained components for maths and English remain. An exception is the ethnicity grade gap for Indian pupils, which becomes statistically indistinguishable from zero. This is most likely because Indian pupils are a particularly high-attaining group and few are observed at this threshold. Finally, in Panel D we restrict our sample to a balanced panel of schools across 2019 and 2020 to explore whether changes in the inclusion of entire schools across years are driving our results. We find that the results are robust to this further restriction.

[19] It is not possible to include 2017 in the reference group due to the need to include lagged school value-added measures in the decompositions. This is not possible for 2017 given changes in national assessments implemented that year.

Table 1.5 Alternative specifications

	Maths				English			
	P&B	I	BA	BC	P&B	I	BA	BC
Panel A: Non-linear specification								
Raw gap change	0.101***	0.091***	0.115***	0.201***	-0.060***	-0.041	-0.103***	0.082**
	(0.023)	(0.028)	(0.024)	(0.035)	(0.020)	(0.028)	(0.023)	(0.036)
Explained gap change	0.062***	-0.011	0.006	0.054*	0.082***	0.055**	0.067***	0.092***
	(0.022)	(0.025)	(0.021)	(0.030)	(0.019)	(0.022)	(0.020)	(0.024)
Unexplained gap change	0.039**	0.102***	0.109***	0.147***	-0.142***	-0.096***	-0.170***	-0.010
	(0.019)	(0.020)	(0.019)	(0.026)	(0.020)	(0.025)	(0.021)	(0.030)
N	934,590				934,590			
Panel B: Combined 2018 and 2019 comparison year								
Raw gap change	0.140***	0.144***	0.105***	0.189***	-0.038**	-0.025	-0.118***	0.052
	(0.022)	(0.025)	(0.021)	(0.030)	(0.019)	(0.024)	(0.021)	(0.031)
Explained gap change	0.102***	0.034	0.020	0.060**	0.103***	0.068***	0.084***	0.111***
	(0.020)	(0.022)	(0.019)	(0.026)	(0.018)	(0.020)	(0.018)	(0.022)
Unexplained gap change	0.039**	0.110***	0.084***	0.128***	-0.140***	-0.093***	-0.202***	-0.059**
	(0.017)	(0.018)	(0.017)	(0.023)	(0.018)	(0.021)	(0.018)	(0.027)
N	1,373,869				1,373,869			
Panel C: Binary outcome (grade 4 and above)								
Raw gap change	0.021***	-0.007	0.029***	0.061***	-0.010**	-0.022***	-0.013***	0.029***
	(0.004)	(0.004)	(0.005)	(0.009)	(0.004)	(0.004)	(0.004)	(0.008)
Explained gap change	0.013***	-0.013***	0.008**	0.024***	0.012***	-0.005	0.013***	0.024***
	(0.004)	(0.004)	(0.004)	(0.005)	(0.003)	(0.003)	(0.003)	(0.004)
Unexplained gap change	0.008*	0.006	0.021***	0.037***	-0.022***	-0.017***	-0.026***	0.005
	(0.004)	(0.004)	(0.005)	(0.008)	(0.004)	(0.004)	(0.004)	(0.007)
N	934,590				934,590			
Panel D: Balanced panel of schools								
Raw gap change	0.089***	0.092***	0.107***	0.194***	-0.075***	-0.037	-0.110***	0.077**
	(0.020)	(0.028)	(0.023)	(0.035)	(0.019)	(0.027)	(0.022)	(0.036)
Explained gap change	0.048**	-0.010	0.011	0.044	0.061***	0.039*	0.068***	0.091***
	(0.020)	(0.023)	(0.021)	(0.030)	(0.018)	(0.021)	(0.019)	(0.024)
Unexplained gap change	0.040***	0.101***	0.096***	0.150***	-0.137***	-0.076***	-0.178***	-0.014
	(0.019)	(0.021)	(0.019)	(0.027)	(0.020)	(0.025)	(0.021)	(0.031)
N	891,899				891,899			

Notes: National Pupil Database, 2017 - 2020. 'P&B' stands for Pakistani and Bangladeshi, 'I' Indian, 'BA' Black African, 'BC' Black Caribbean. Sample comprises age 16 pupils in mainstream schools in England with no missing data. Exam grades are the May and June (end of year) examinations only. The English grade is the highest of English Literature and English Language if pupils received grades for both. The reference group is White British. Explanatory characteristics are the same as those in Table 1.2. Standard errors in parentheses, clustered at the school-level. * Significant at 10%; ** significant at 5%; *** significant at 1%.

We also examine the extent to which our results are sensitive to the choice of the English outcome used. So far we have used the higher grade that pupils achieved in English Literature and English Language, as this is the outcome used for performance benchmarks for both schools and pupils (Department for Education, 2020b). Instead, Appendix Table A.6 reports separate results for English Literature and English Language. As in Table 1.2, for both outcomes the point estimates for the explained gap changes are positive and consistently significant (apart from for Indian pupils), indicating that we would expect to see positive grade gap changes for all ethnic minority groups in both English outcomes, given the changes in the groups' characteristics and the returns to those characteristics in 2020 compared to 2019. The unexplained contributions are generally negative for both outcomes, indicating that there remain unexplained factors, such as stereotyping effects, which are leading to the observed ethnicity grade gap changes by reverting the positive impact of the observed characteristics. As the raw gap changes are generally null in the case of English Language and negative in the case of

English Literature, these unexplained negative components are considerably larger (about twice as large) in the case of English Literature.

To explore why potential stereotyping effects might be greater for English Literature than English Language, we analyse the correlations between English Literature and English Language grades across the ethnic minority groups. We find the correlation coefficient for White British pupils (0.843) is larger compared to the ethnic minority groups (0.825 for Pakistani and Bangladeshi pupils, 0.815 for Indian pupils, 0.809 for Black African pupils, and 0.816 for Black Caribbean pupils). As shown in Appendix Table A.5, this difference appears to be the result of ethnic minority pupils tending to achieve higher, relative to White British pupils, in English Literature than in English Language.[20] We therefore report, in Table 1.6, the raw, explained, and unexplained components of a decomposition using the mean English grade as the outcome, rather than the higher of English Literature and English Language. Though the estimates are smaller than those previously reported, the pattern of the results is unchanged.

Table 1.6 Decomposition using mean English grade

	P&B	I	BA	BC
Raw gap change	-0.051***	-0.018	-0.078***	0.085**
	(0.019)	(0.028)	(0.022)	(0.034)
Explained gap change	0.079***	0.054**	0.078***	0.098***
	(0.019)	(0.021)	(0.019)	(0.024)
Unexplained gap change	-0.130***	-0.073***	-0.156***	-0.013
	(0.019)	(0.023)	(0.020)	(0.028)
N	934,590			

Notes: National Pupil Database, 2017 - 2020. ‘P&B’ stands for Pakistani and Bangladeshi, ‘I’ Indian, ‘BA’ Black African, ‘BC’ Black Caribbean. Sample comprises age 16 pupils in mainstream schools in England with no missing data. Exam grades are the May and June (end of year) examinations only. The English grade is the highest of English Literature and English Language if pupils received grades for both. The reference group is White British. Explanatory characteristics are the same as those in Table 1.2. Standard errors in parentheses, clustered at the school-level. * Significant at 10%; ** significant at 5%; *** significant at 1%.

1.9 Discussion and conclusion

An important question in education policy is whether pupils with different characteristics receive systematically different grades in teacher assessments compared to blindly-marked examinations. We

[20] Establishing why there are differences in ethnic minority pupils’ relative grades for English Literature and English Language is beyond the scope of this paper. However, English Language is a more skills-based exam, requiring pupils to demonstrate proficiency in comprehension, analysis, and free writing. English Literature is more knowledge-based and requires pupils to write essays about selected texts. Neither has a coursework element during our years of focus. It may be, then, that the skills assessed in English Literature are different to those assessed in English Language in a way that matters for ethnic minority pupils’ performance. For example the English Language syllabi may be more culturally biased than those for English Literature (Ofqual, 2022).

contribute new evidence about this question by exploiting a change in assessment methods to examine teachers' predictions of pupils' examination performance. In doing so we use a different comparison to that used in much existing research on teacher assessment, which relies on comparing "blind" (examination) assessments to "non-blind" (teacher) assessments which are intended to consider pupils' written, practical, and oral classwork over an entire academic year. Our approach provides a level of comparability in outcome – in our case examination performance – which is arguably greater than in some of these existing studies.

We find that, when grades are assigned by teachers rather than through externally marked examinations, ethnic minority pupils tend to do relatively better than White British pupils in maths and relatively worse than White British pupils in English. These ethnicity grade gap changes do not appear to be driven by differences across groups and cohorts, time trends, ceiling effects, or by changes across years in the observed characteristics of pupils and schools, or their returns. In fact, while for maths the changes in the levels of and returns to observed characteristics do explain some – but not all – of the observed ethnicity grade gap changes, in the case of English our results suggest that had those changes not occurred, the relative drop in the English performance of ethnic minority pupils that followed the switch to teachers' assessment would have been even larger – roughly double – in magnitude.

There are other potential explanations for the ethnicity grade gap changes that we document. The timing of the announcement to cancel examinations in 2020 allows us to rule out that changes in teaching practices are driving our results, as this announcement came just two months before the examinations were due to begin. Similarly, as teachers were asked to predict precisely how they believed their students would perform had the examinations gone ahead, primarily based on mock examinations completed prior to school closure, we do not believe that our findings are driven by systematic differences in how pupils perform in class-based work and examinations, or by differences in their experiences during school closure.

One limitation of our analysis is that we cannot fully check whether the 2019 and the 2020 cohorts differ in unobserved characteristics. As most time-invariant unobservable characteristics should be captured by the measures of past attainment (at age 11) that we use in our decomposition, any effect of time-varying unobserved characteristics will be included in the unexplained part. However, in order to explain the residual grade gap changes we observe, the effect of unobservables would have to be very large in maths, and very large and working in the opposite direction to the effect of observable characteristics in English. Based on this, we argue that group- and subject-specific teacher stereotyping contributes to at least some part of the ethnicity grade gap changes we document.

Unlike much of the literature on stereotyping, we find both positive and negative unexplained components for a gap in outcomes for the same group within the same broad competency – academic achievement. Our results suggest that teachers hold positive stereotypes about the performance of ethnic minority pupils in maths in high-stakes examinations taken at age 16, and negative stereotypes about the performance of ethnic minority pupils in English. These findings are in line with, and add to, existing research about teachers' assessments of pupils aged seven to 14 in England (Campbell, 2015; Gibbons and Chevalier, 2008). However they do not conform to common economic approaches whereby stereotypes are held to be manifestations of statistical discrimination (Phelps, 1972) or representativeness heuristics (Bordalo et al., 2016; Esponda et al., 2023), as most ethnic minority groups tend to perform better than White British pupils in both subjects in 'normal' years. It may be that teachers' stereotypes are based on information about pupils at a younger age; for example at school entry, when ethnic minority pupils tend to perform relatively worse in literacy and English than in maths compared to White British pupils (Dustmann et al., 2010). Alternatively they may be informed by the educational choices of older pupils; conditional on prior attainment and family educational background, most ethnic minority pupils are more likely to choose to study STEM subjects at A Level (ages 16 to 18) than White British pupils (Codiroli Mcmaster, 2017).

Our finding that teacher assigned grades are likely affected at least in part by stereotypes – both positive and negative – is relevant to important policy considerations. Most countries use non-blind assessments at some stages of children's educational trajectories, and in England, teacher assessments are still considered the primary contingency plan for any future examination disruption (Ofqual, 2023). There is also an ongoing policy debate in England about whether teacher assigned grades should replace high-stakes national examinations because they may be better for pupil well-being, curriculum breadth, and assessing pupils according to competencies that examinations cannot measure (Council of Skills Advisors, 2022). Our results suggest that the benefits of teacher assessments must be weighed against the risk that favourable or unfavourable non-blind assessments reflecting stereotypes might affect pupils' opportunities, future performance, and education preferences (Burgess et al., 2022;

De Benedetto and De Paola, 2023). If teachers' assessments and expectations differ systematically according to pupils' characteristics then this could affect gaps in educational outcomes and returns both during schooling and in higher education. With better knowledge of the mechanisms that lead teachers to adopt stereotypes, interventions such as information campaigns could be devised to lessen their effects and make teacher assessments more consistent.

2 The Medium-Term Effects of Extended Play-Based Learning in Early Childhood

2.1 Introduction

There is a broad consensus that the early years of children's development – from birth to around age eight – are crucially important for children's later life outcomes (Black et al., 2017). The environments to which young children are exposed have been found to impact the developing architecture of the brain, affecting children's cognitive and noncognitive development (Noble et al., 2015). Returns to investments in early childhood may therefore be higher than those in later childhood or adolescence, as they set children up for more successful learning and development in later years (Cunha and Heckman, 2007). At least in part in response to this consensus, governments around the world are investing heavily in early childhood care and education (Engel et al., 2015): in the US, the Biden administration published an Executive Order to increase access to high-quality childcare; in England, the Department for Education announced a universal childcare offer which will cost billions; in Finland, the government is trialling an intervention which would lower the starting age of mandatory preschool from age six to age five (Harjunen et al., 2022). Crucial to the success of these policies is that they target skills and capacities which are time-sensitive, meaning those that are important for upcoming developmental transitions – for example, the transition to formal schooling – and which would not develop eventually anyway, had an intervention not occurred (Duncan et al., 2023).

Yet, despite a consensus about the importance of early childhood development, it is still not clear what skills and capacities early childhood development policies *should* target, or how they are best implemented. One reason for the lack of clarity is that the existing economics literature has tended to evaluate early childhood interventions through the effects of additional time in childcare, with what actually happens within settings remaining largely a black box. Many analyses in European countries, for example, exploit variation in universal preschool access across space and time against a counterfactual of parental or informal care (Blanden et al., 2016; Cornelissen et al., 2018; Havnes and Mogstad, 2015), while analyses of different pedagogical approaches introduced into contexts with pre-existing universal provision are rare and generally not yet old enough to measure long-run effects (Rege et al., 2024). A second reason for the lack of clarity is that short-term effects of early childhood development policies tend to fade out or differ from long-run effects, which by their nature take considerable time to evaluate. A meta-analysis of targeted preschool programs in the US, for example, found that cognitive effect sizes drop roughly in half in one to two years (Li et al., 2020), while rigorous analyses have found positive noncognitive effects for the famous Perry Preschool Project which persisted into adulthood (García et al., 2021; Heckman et al., 2013).

In this paper I evaluate the short- and medium-term effects of an extended play-based learning policy for children between ages five and seven in Wales – the ‘Foundation Phase’ – against a counterfactual of formal schooling. The Foundation Phase policy constituted a "radical change" in the curricula and pedagogy of children aged five to seven in Wales (Taylor et al., 2015b). Previously, children at these ages received their education in a formal setting with lessons in several academic subjects. Instead, drawing inspiration from early years programmes in Scandinavia, New Zealand, and Reggio Emilia in Italy, the Foundation Phase allowed children to experience a more interactive school environment where they could learn through guided-play rather than direct instruction, with curriculum focus given to personal and social development and wellbeing. The policy design was informed by pedagogical theories suggesting that a prolonged exposure to play-based learning would allow children to encounter formal schooling at a more developmentally-appropriate age and lead to long-term cognitive and noncognitive gains (Hirsh-Pasek et al., 2009; Maynard et al., 2013). The policy constituted three main components: a new curriculum, financial resources intended to support schools to create appropriately interactive learning environments, and a non-mandatory reduction in class sizes.

The Foundation Phase was piloted through a staggered roll-out in 43 schools between 2004 and 2009, and then implemented nation-wide. My analysis therefore relies on a staggered difference-in-differences research design comparing consecutive cohorts of pupils attending schools. Under a parallel trends assumption, the variation generated by the fact that only some schools were included in the Foundation Phase Pilot allows me to obtain causal estimates of the effects of the Foundation Phase on pupils' medium-term outcomes. In particular, I estimate the effects of Foundation Phase Pilot provision on pupils' academic attainment and school attendance at age 11, and academic attainment, school exclusions and educational destinations at age 16. My empirical strategy allows me to rule out various confounding factors: first, school-specific differences fixed in time (e.g., pupils at Pilot schools may already have better or worse test scores); second, differences across time that affect all pupils in a similar way (e.g., certain macroeconomic fluctuations or social trends); third, differences across time that may affect Pilot school pupils differently (e.g., other educational policies which disproportionately affect Pilot schools). I address recent econometric concerns with heterogeneous treatment effects in settings with staggered treatment timing by using the imputation-based difference-in-differences estimator proposed by Borusyak et al. (2024). Nonetheless there are limits to the external validity of my results. As differences-in-differences methodologies estimate an 'average treatment effect on the treated' (ATT) effect, these cannot be assumed to be similar to those that we would expect of all schools. I do find, for example, that the characteristics of pupils attending Pilot schools are systematically different at baseline compared to the averages of all other schools in Wales, which might mean that the Pilot schools have higher (or lower) gains from implementing the Foundation Phase policy.

I find no statistically significant effects of the Foundation Phase Pilot on pupils' school-related outcomes at ages 11 or 16. This is an interesting result as it suggests that, on average, pupils who received two fewer years of direct instruction during early Primary education still achieved similarly to their peers in later education, made similar decisions about their educational pathways, and exhibited similar school-related behaviours. These findings are consistent with other studies evaluating the long-term effects of play-based compared to more formal pedagogical approaches (Biroli et al., 2018).

Following a number of short-term early childhood intervention studies (Duncan et al., 2023; Fidjeland et al., 2023), I also examine different treatment effects by the gender and the socio-economic status of the pupil. I find little evidence of statistically significant heterogeneous effects of the Foundation Phase Pilot, though some indication of more positive results for girls and for more deprived pupils.

There are a number possible reasons for these results. First, initial gains or losses may have already ‘faded out’ by the time that the pupils receiving Foundation Phase Pilot provision reached 11 and 16 years old. Relatedly, it also is possible that the Foundation Phase pupils made additional gains in other skill domains which are harder to measure in administrative data. Indeed, though the measures that I use are include those that the Welsh Government cited as some of the intended future success criteria for the policy (National Assembly for Wales, 2003), the outcomes available to me in the Welsh administrative datasets are likely poor proxies for the full range of outcomes that policy designers and educational practitioners envisioned for pupils who received the Foundation Phase. I am unable, for example, to directly measure children’s creativity or attitudes towards learning, their socioemotional skills or wellbeing, or the extent to which they have or will become active citizens in their communities. Measurable effects of the Foundation Phase Pilot may also have been diminished by factors related to the policy’s implementation. For example, the Foundation Phase Pilot provision may not have differed as much as intended from the counterfactual provision, which in this case was business-as-usual formal schooling. Alternatively, effects of the change in curriculum and pedagogy may have been obscured by the additional funding and reduction in class sizes, or by parents changing their investments at home. Unfortunately, strong evidence supportive of any of these hypotheses is beyond the scope of the data made available for the present research.

This paper is not the first to evaluate the Foundation Phase Pilot. Quantitative and qualitative evaluations of the Foundation Phase Pilot’s contemporaneous and short-term effects were completed for the Welsh government shortly after its implementation, and this work is summarised in Taylor et al. (2015b). Using data from a cohort study the researchers find tentative evidence of negative contemporaneous effects (at age seven) on both pupil attainment and wellbeing at the end of the program (Taylor et al., 2015a). Using administrative data, they then find evidence to suggest an improvement in test scores at age 11 (Davies et al., 2015). Taken together, these results lend tentative support for the theoretical rationale of the Foundation Phase: that extended play-based learning may lead to lower outcomes in the short-term but improved outcomes in the long-term, when compared to the business-as-usual, direct instruction approach. However, due to the relatively short amount of time between implementation of the Pilot and these evaluations, both studies used relatively small sample sizes and, for age 11 outcomes, only half of the Pilot schools as a treatment group, particularly restricting the conclusions that can currently be drawn about the policy's medium-term effects.[1] I build on these existing evaluations by using all Pilot schools to evaluate age 11 outcomes, and by extending the evaluation to age 16.

As formal childcare provision continues to expand globally, this paper also contributes much-needed causal evidence about the conditions under which early childhood education is effective. Much of the existing economics literature examines the short- and long-term effects of either targeted or universal pre-school programs against a counterfactual of informal care (Dietrichson et al., 2020; Duncan et al., 2023).[2] However these studies do little to aid decisions about what kind of provision young children should receive when that provision is universal. Worldwide, existing approaches to early childhood education usually tends towards either a social pedagogical tradition – which holds that more time engaging in free- and guided-play develops children's curiosity, motivation and wellbeing – or a school readiness tradition, where formal schooling using direct instruction is believed to help to boost and track children's academic achievement and reduce inequalities. Because different approaches are seldom implemented in comparable contexts, comparisons are rare. One exception is Biroli et al. (2018), who compare the Reggio Emilia (social pedagogical) approach to other preschool pedagogies in nearby cities in Italy and find very few significant effects across a large number of short- and long-term measures spanning IQ, educational attainment and health. Another is Rege et al. (2024), who find positive short-term effects on test scores when five-year-olds in Norway receive a structured rather than an unstructured curriculum, but are not able, yet, to examine any medium- or long-run effects. I therefore contribute one of the few studies which estimates the medium-term effects of one

[1] Specifically, Taylor et al. (2015a) have to rely primarily on a cohort study containing a sample of just 91 children in their analyses of contemporaneous effects at age seven. In addition, due to the relatively early timing of the short-term effects' analyses at age 11, Davies et al. (2015) were only able to use about half of all Pilot schools for the test score analyses, again a relatively small sample size.

[2] In general, this literature finds that interventions emphasising basic maths and literacy skills can have impressive short-term effects which often fade out during elementary school due to control group catch up (Bailey et al., 2020; Li et al., 2020), while persistent effects are consistently found for long-run outcomes such as adequate primary and secondary school progression, years of schooling and earnings (Dietrichson et al., 2020; Gray-Lobe et al., 2023).

pedagogical approach in early childhood compared to another, namely a guided-play approach for children between ages five and seven against a counterfactual of formal schooling.

Finally, this paper also contributes to a literature on school starting ages. Most prior contributions to this literature exploit birth date school entry rules to compare pupils who are similar in age but enter formal schooling around one year apart. They generally find positive and persistent effects for children who enter school at an older age than their classmates, for example for their cognitive skills (Dhuey et al., 2019), non-cognitive skills (Lubotsky and Kaestner, 2016), earnings (Fredriksson and Öckert, 2014), avoidance of crime (Landersø et al., 2017) and mental health (Broughton et al., 2023). However the two likely mechanisms of these effects – relative age effects (i.e. pupils being older rather than younger than their peers) and school readiness effects (i.e. formal schooling being more developmentally appropriate for older children) – are difficult to disentangle.[3] I contribute to this literature by examining a context in which children of comparable ages receive play-based (rather than formal) educational provision earlier or later than their peers, therein providing greater clarity as to the likely mechanisms of positive school starting age effects. Specifically, my results imply that relative age effects are a more compelling explanation than school readiness effects, as children who begin formal schooling at a later age appear to experience no difference in medium-term outcomes.

The rest of this paper proceeds as follows. Section 2.2 provides an overview of education in Wales, the Foundation Phase policy, and the Pilot implementation. Section 2.3 then discusses the data and sample used. Section 2.4 covers the empirical strategy. Section 2.5 contains the baseline results, a series of robustness and sensitivity checks, and an exploration of heterogeneous effects. Section 2.6 concludes.

2.2 Education in Wales and the Foundation Phase policy

2.2.1 The Welsh education system

In Wales, four-year-old children are offered free, full-time education in Reception (kindergarten) classes. Compulsory full-time education then begins at age five. The compulsory schooling years

[3] One exception is Dee and Sievertsen (2017), who find that a one-year delay in kindergarten entry improves children's self-regulation considerably more than other noncognitive measures, arguing that children who experience extended exposure to playful preschool environments may therefore have improved school readiness.

are split into four Key Stages, with teacher assessments completed at the end of the first three Key Stages and externally-marked examinations (GCSEs) at the end of the fourth.[4] Children transition from Primary to Secondary schooling at the end of the second Key Stage, at age 11, and are allowed to leave school at the end of the fourth, at age 16. Academic and vocational options are available to those who wish to continue in schooling from age 16 to 18, after which some may enter university.[5]

Prior to age four children may receive preschool provision of some kind. Since 2008, all three- and four-year-olds have been offered a minimum of 10 hours per week of preschool provision (Glyn et al., 2019). However, at the time of the Foundation Phase Pilot (2004 to 2009), the extent of this free provision varied by geographical area (Hanney, 2000). The children considered in this evaluation, therefore, could have had different levels of education between the ages of three and five, with some children potentially receiving none. However the year before the Pilot began, 75% of three-year-olds and 80.6% of three- and four-year-olds attended some form of educational setting, with the great majority of four-year-olds attending Reception (kindergarten) classes at a school (Siraj-Blatchford et al., 2005; Wilton and Davies, 2017).

Schooling across Wales is quite homogeneous. Structural constraints such as curricula, examinations, teacher qualification requirements and teacher remuneration are all governed by national policy, with implementation of these national policies directed by the 22 Local Authorities (administrative districts). The vast majority of schools in Wales are state-funded; in 2004, when the Foundation Phase Pilot began, only 1.36% of Primary school children attended privately-funded schools (StatsWales, 2023a). Unlike grammar schools in England, there are no state-funded schools in Wales which can admit pupils based on academic ability. Schooling in Wales is also bilingual. At the time of the Foundation Phase Pilot, about 30% of Primary schools in Wales used Welsh as the sole or main medium of instruction (StatsWales, 2023b).

[4] Standard Assessment Tests (SATs) were phased out in Wales between 2000 and 2005.

[5] In Wales, the school leaving age is 16.

2.2.2 The Foundation Phase

Policy design and counterfactual provision

Prior to the introduction of the Foundation Phase, children aged three to seven in Wales followed two pre-existing curricula. For three- to five-year-olds (preschool and Reception), several areas of learning developed children's school readiness through a free-play approach (National Assembly for Wales, 2000a). Five- to seven-year-olds (Key Stage 1) then followed a National Curriculum including content in eleven statutory curriculum subjects – for example English, maths, science, history, geography, and art – delivered through direct instruction in a formal setting (National Assembly for Wales, 2000b). At age seven children were assessed, both by teachers and through specified tasks or tests, in English, Welsh if it is a child's first language, maths, and science. These pre-existing curricula and assessments continued to be used in settings which were not part of the Foundation Phase Pilot.

The Foundation Phase policy constituted a "radical change" from these pre-existing curricula and assessments (Taylor et al., 2015b). Inspired by the early years programmes in Scandinavia, New Zealand, and Reggio Emilia in Italy, the Foundation Phase introduced a play-based pedagogy for three- to seven-year-olds which comprised seven Areas of Learning: Personal and Social Development, Well-Being and Cultural Diversity; Language, Literacy and Communication Skills; Mathematical Development; Welsh Language Development (in English-speaking schools); Knowledge and Understanding of the World; Physical Development; and Creative Development (National Assembly for Wales, 2008). The intended academic outcomes of the Foundation Phase were similar to those of the previous curricula, but the accompanying curriculum and assessment frameworks made it clear that learning was to occur in a more play-based, developmentally-appropriate, and cross-curricula way (Maynard et al., 2013). For example, the End of Foundation Phase Assessments covered each of the seven Areas of Learning and assessed children during snack time, circle time, reading together, or playing a game. The curriculum also stated that children should learn through "first-hand experiential activities" and "play", and that the Foundation Phase had "the development of children's self-image and feelings of self-worth and self-esteem" at its core (National Assembly for Wales, 2008).

It is helpful to define the Foundation Phase treatment, most precisely, as constituting guided-play pedagogy for children aged five to seven against a counterfactual of direct instruction. This is because the changes brought about by the Foundation Phase were considerably more stark for children aged five to seven than for children aged three to five. Exemplar vignettes of Foundation Phase lessons show that the intended pedagogical model was free-play for children aged three to five, as in the counterfactual provision, but guided-play for children aged five to seven, which was very different from the previous, direct instruction approach (Taylor et al., 2015b). In addition, only 9.3 and 15.6 percent of teachers leading the implementation of the Foundation Phase during the subsequent national roll-out felt that it was "considerably different" to what they were already implementing in preschool and Reception classes (ages three to five), while 56.8 and 64.9 percent felt this way regarding the two older age groups (ages five to seven) (Taylor et al., 2015b).

Finally, two further inputs, apart from the new curriculum, were part of the Foundation Phase policy: increased financial resources, and an intended (though non-statutory) reduction in class sizes. The financial resources were partly intended to support schools to turn classrooms with tables and chairs into environments which were more appropriate for play-based learning, including outdoor space. However a considerable proportion of the resources were used to reduce pupil-teacher ratios, for example from 1:20 to an intended 1:15 for five- to seven-year-olds (Davies et al., 2015; Siraj-Blatchford et al., 2007). It is not clear whether the qualifications of additionally-employed adults differed from those already employed. However, for the Pilot phase, all classes had at least one qualified teacher to comply with national minimum requirements, with other staff qualifications varying from no qualifications to an advanced vocational qualification (Taylor et al., 2015b).

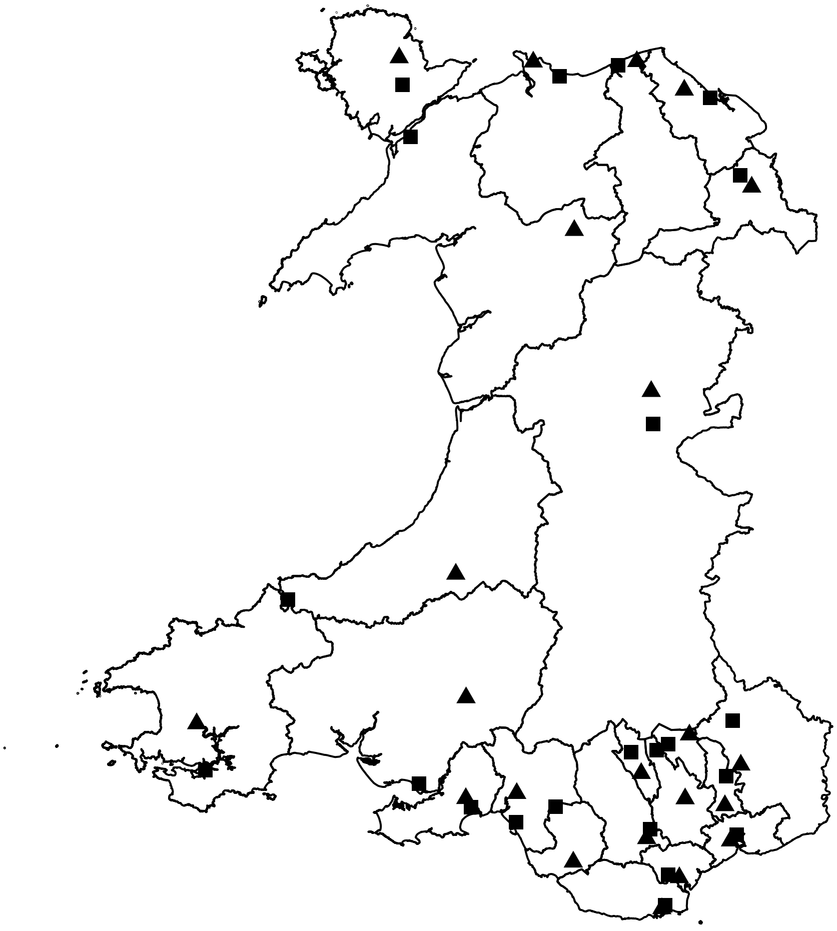
Pilot implementation

The Foundation Phase was piloted in 43 schools in two implementation groups prior to the national roll-out in 2008/09.[6] These Pilot schools were not selected at random. Instead we know that, for the first group of 21 Pilot schools (‘Group 1’), each Local Authority (LA) suggested one or two schools to the Welsh government for inclusion, of which one from each LA was then selected according to Welsh Government-prescribed criteria (Siraj-Blatchford et al., 2007; Taylor et al., 2015a). For the second group of 22 Pilot schools (‘Group 2’), we know that the schools selected were again one from each LA

[6] In the Pilot phases, the roll-out began for children aged three to five (i.e. nursery and reception) and this was intended in the national roll-out. However the Welsh Government amended these plans to allow schools longer to implement the necessary changes, meaning that the Foundation Phase was implemented in nurseries from 2008/09 and reception classes from 2009/10, despite the ‘headline date’ being 2008/09.

but also that they were serving particularly low-income communities (Taylor et al., 2015b).[7] A map indicating the locations of the Pilot schools can be seen in Figure 2.1, showing their even geographical dispersal across the nation. I only have access to the names of the selected schools, not of all those suggested, and I examine systematic differences between the Pilot schools and all schools in Wales in Section 2.3.2.

Figure 2.1 Map showing the Foundation Phase Pilot schools



Notes: Map showing Wales and its Local Authorities. Schools in Pilot Group 1 are shown by square icons and schools in Pilot Group 2 by triangles. Each Local Authority contains exactly two pilot schools, one in Group 1 and one in Group 2.

It is also important to note that significant geographical and temporal variation in formal guidance and training appears to have been available when the Pilot settings first introduced the Foundation

[7] The second group of pilot schools were in areas which were also in receipt of a service called Flying Start, which targeted children up to age four in specific areas of high deprivation. The late introduction of Flying Start means that it does not greatly threaten the identification strategy used here. Nonetheless I explore the possible effects of this contemporaneous policy in the robustness checks.

Phase (Maynard et al., 2013; Siraj-Blatchford et al., 2005). For example only 50% of surveyed Pilot Group 1 school practitioners agreed that the support and training they received was appropriate (comparable survey data is not available for Group 2), and a Foundation Phase Pilot DVD – which highlighted good and effective practice – was made available in early 2006, more than one year after Pilot Group 1 schools had begun implementation (Siraj-Blatchford et al., 2005). It is therefore appropriate to interpret treatment effects of all Pilot schools with caution, as their provision may have differed somewhat from that of schools in the national roll-out. As I will show later, the results also tentatively suggest more positive effects for Pilot Group 2, which may be related to the later provision of additional training materials.

2.3 Data and descriptive statistics

2.3.1 Data and construction of the treatment indicator

I use administrative data from Wales which is held by the SAIL Databank. Many of the datasets I use differ in the years for which the data are available. An overview of the datasets used and the years available can be found in Appendix Section B.1.1. Appendix Section B.1.2 then provides an in-depth account of data cleaning and variable construction. In the interest of space I provide only a summary here.

The Education Wales dataset contains individual-level administrative data relating to the education system in Wales including the Pupil Level Annual School Census (PLASC), academic attainment, exclusions, absenteeism, and school-level data. As all state-funded schools in Wales are required to return these data by law, the Education Wales data are both accurate and highly complete. The earliest data available are for the 2003/04 academic year, though some subsets of the data are available only in later years.

For further demographic data and health outcomes, I make use of the unique provisions of the SAIL Databank to anonymously link pupils to a number of healthcare datasets. The Welsh Demographic Service dataset is a register containing demographic information about individuals registered at primary care (GP) establishments in Wales, whether or not that establishment submits additional health data to the SAIL Databank. This dataset records pupils' week of birth as well as the neighbourhood deprivation of the pupil's home address at every age using deciles of the Welsh Index of Multiple Deprivation (WIMD).[8] I also use three additional healthcare datasets to identify health outcomes, covering primary, secondary and emergency healthcare.[9] The healthcare datasets are anonymously linked using an encrypted National Health Service number. The education and healthcare data are then linked using an encrypted individual identifier derived from individuals' name, gender, postcode of residence, and date of birth.

Finally, for treatment identification I supplied publicly available data on the unique identifying codes of the schools which took part in each stage of the Foundation Phase Pilot. These were then flagged anonymously by the Welsh Government prior to releasing the data extract. However it should be noted that this method of treatment identification is different to that used in earlier evaluations of the Foundation Phase Pilot (Taylor et al., 2015b). In particular, the earlier evaluations had access data on Foundation Phase assessments completed by seven-year-old pupils in the Pilot schools, which are no longer available. Both methods of treatment identification may contain measurement error in the treatment indicator, and it is possible that the groups of children that they identify are not exactly the same. In the following section I therefore outline some additional details which need to be taken into account when understanding the identification of the Pilots schools and their pupils using the names of Pilot schools, as in the present study.

Construction of the treatment indicator

The Foundation Phase Pilot schools were directed to introduce the Foundation Phase to one cohort at a time, starting with children in Reception (kindergarten) classes, at age four (Davies et al., 2013).[10] 20 schools were in the first Pilot group ('Pilot Group 1'), beginning in 2004/05 (henceforth 2005), and

[8] I use the Welsh Index of Multiple Deprivation 2014 index, which combines data on eight domains of income, employment, health, education, access to services, community safety, physical environment, and housing.

[9] The Welsh Longitudinal General Practice dataset contains attendance and clinical information for all General Practice (GP) interactions – including symptoms, diagnoses and prescriptions – at GP practices which submit data to SAIL. This includes around 80% of the Welsh population, and the available data are representative of the entire Welsh population with respect to age, sex, and level of deprivation. The two hospital datasets that I use in addition to these primary care records are the Patient Episode Dataset for Wales, which contains attendance and clinical information for all hospital admissions in Wales, and the Outpatient Dataset, which contains attendance information for all hospital outpatient appointments.

[10] This means that, when the first cohort then graduated to the next curriculum year, the Foundation Phase would be delivered to the existing cohort *and* the new Reception (kindergarten) class, and only in the following year would it be delivered to all children up to age seven (the end of Key Stage 1) for the first time.

another 22 schools were in the second ('Pilot Group 2'), beginning in 2007/08 (henceforth 2008).[11] However I identify the pupils that took part in the Foundation Phase Pilot, and the periods of their participation, taking into account some additional contextual and data-related challenges.

First, I split Pilot Group 1 into two groups. This is because some Primary schools in Wales use mixed-age classes, especially in smaller schools. In their prior analyses Davies et al. (2013) found that about half of the schools in Pilot Group 1 submitted Foundation Phase assessment results – rather than business-as-usual, age seven assessment results – a year earlier than expected, and interpreted this as an indication that the previous cohort in these schools also received Foundation Phase provision, most likely because their classes were mixed (Davies et al., 2013). Although I do not have access to the Foundation Phase assessment results, I am able to identify 13 Pilot Group 1 schools who appear to be missing the counterfactual assessments in the expected year. I therefore treat these schools as a subsection of Pilot Group 1 ('Pilot Group 1a') and as though they implemented the Foundation Phase a year earlier. As the older cohort of pupils will have been ages five to six (rather than four to five) when the Foundation Phase was first rolled out in their school they can still be understood as having received the Foundation Phase treatment as defined here. In Section 2.5.2, I show that performing this separation does not meaningfully change my results.

Second, I identify as their school the schools that pupils attend when they are ages six to seven, which aligns with the final year of Foundation Phase provision. This is because the Education Wales dataset begins in the 2003/04 academic year, meaning that I now observe just one cohort of pupils at ages five to six prior to the Foundation Phase Pilot implementation beginning for Pilot Group 1a. This is a problem because having just one untreated cohort restricts my ability to assess the plausibility of the parallel trends assumption. However, because pupils reach ages six to seven one academic year later than they reach ages five to six, identifying pupils at ages six to seven allows me to identify comparable school cohorts consistently for at least two years prior to the first implementation of the Foundation Phase Pilot for Pilot Group 1a.[12] Figure 2.2 illustrates the data range and Pilot roll-out with

[11] One setting from Pilot Group 1 was a nursery and so not included here. One is also a special school, which I drop (see section 2.3.2 for more information). In addition, two Primary schools (one a Pilot school) amalgamated in the final year of the Pilot so I drop the original Pilot school for this final cohort only. One Primary school closed in the final year of the Pilot and I am therefore unable to observe any pupils in that school for the final cohort only.

[12] The implication is that I am able to observe the very first (fully untreated) cohort *only* at ages six to seven and not at ages five to six, as this is when the data begins.

school attended identified at age six, including the separation of Pilot Group 1 into two implementation groups.

Figure 2.2 Graphical representation of the Pilot roll-out

Year Cohort Began School

	2002	2003	2004	2005	2006	2007	2008	2009	2010
Pilot group 1a	□	□	▨	▨	▨	▨	▨	▨	▨
Pilot group 1b	□	□	□	▨	▨	▨	▨	▨	▨
Pilot group 2	□	□	□	□	□	□	▨	▨	▨
All schools	□	□	□	□	□	□	□	□	▨

Notes: Years represent the end of each academic year, e.g. ‘2004’ represents the academic year 2003/04. Pupils begin school at ages four to five (Reception). School attended is identified at ages six to seven (Grade 2). Shaded cells represent observable cohorts who received the Foundation Phase. Pilot Group 1 is split into Pilot Group 1a and Pilot Group 1b to account for mixed-age classes.

Identifying pupils as attending school when they are ages six to seven rather than ages five to six risks a lack of precision caused by pupils potentially moving schools between these years.[13] To explore the extent of this problem, I estimate the rate of school mobility between ages five to six and ages six to seven for the cohorts that I can observe at both of these ages. As shown in the first two columns of Appendix Table B.3, I find that the mobility rate at this age is 4.53%, meaning that 4.53% of pupils are in different schools at ages six to seven to the schools that they are in at ages five to six. I find that this mobility rate is not statistically different for pupils attending Pilot schools or for pupils attending Pilot schools in the post-implementation period. However, as this proportion is still relatively high, in Section 2.5.2 I also restrict my treated sample to those pupils who definitely attended pilot schools at age five and six, and find no difference in the results.

Finally, I identify schools in a flexible way that accounts for a gradual amalgamation of Primary schooling that occurred in Wales during the evaluation period. Specifically, the Foundation Phase Pilot was introduced in a mixture of Infant schools (ages four to seven) and Primary schools (ages four to eleven). To ensure that I can observe schools across time, I assume that any Infant and Junior (ages

[13] In particular, the identified treatment group may not have all been fully treated, and there may be some spillovers into the control group.

seven to eleven) schools that amalgamated to create Primary schools during the observation period served the same pupils.[14] For example, if Red Infant School amalgamated with Red Junior School to make Red Primary School at some point between the year 2003 and the year 2020, I assume them to essentially function as one school for the duration. I check this assumption by estimating, in the third column of Appendix Table B.3, the likelihood of pupils changing school between age six to seven (for mergers, the last year of Infant school) and seven to eight (for mergers, the first year of Junior school), by whether the school merged during the observation period. I find that the rate of school mobility is slightly higher for merging schools (5.98%) than for non-merging schools (3.55%). However in Section 2.5.2 I find that removing merger schools does not meaningfully change my results.

2.3.2 Sample and descriptive statistics

The main sample is made by retaining pupils born across eight cohorts who began school (in Reception) between 2002 and 2009. I drop 11.5% of pupils because they are not observed at ages six to seven (henceforth six), the age of treatment identification. I then drop a further 1.26% who are missing Welsh Index of Multiple Deprivation data due to imprecise linkages between the education and healthcare datasets, a further 2.70% who are missing other demographic characteristics, and a further 0.81% who attend a special school at some point during their education.[15] This leaves a final sample of 245,669 pupils, of whom 11,920 (4.85%) attend Foundation Phase Pilot schools.

I present descriptive statistics for the characteristics of the pupils attending Pilot schools and the rest of Wales, including separately by Pilot group, in Table 2.1. The Pilot schools are more ethnically diverse and more deprived than the schools in the rest of Wales. For example 47.0% of pupils attending Pilot Group 2 schools, which were selected to be from more-deprived neighbourhoods, live in the 10% most deprived neighbourhoods in Wales, compared with just 11.4% of pupils attending non-Pilot schools. It is also noteworthy that the Pilot schools – both in aggregate and separately – are less likely to achieve the End-of-Primary ‘Core Subject Indicator’ threshold at age 11. Though age 11 attainment is likely to more greatly reflect pupil demographics than school quality (Wilkinson et al., 2018), these differences suggest that the Pilot schools were not selected because of above-average performance on standard measures.

[14] Precisely five Pilot Infant schools merged with Junior schools during the Pilot period.

[15] Special schools provide education for children with severe Special Educational Needs or disabilities.

Table 2.1 Characteristics of pupils in Pilot and control schools

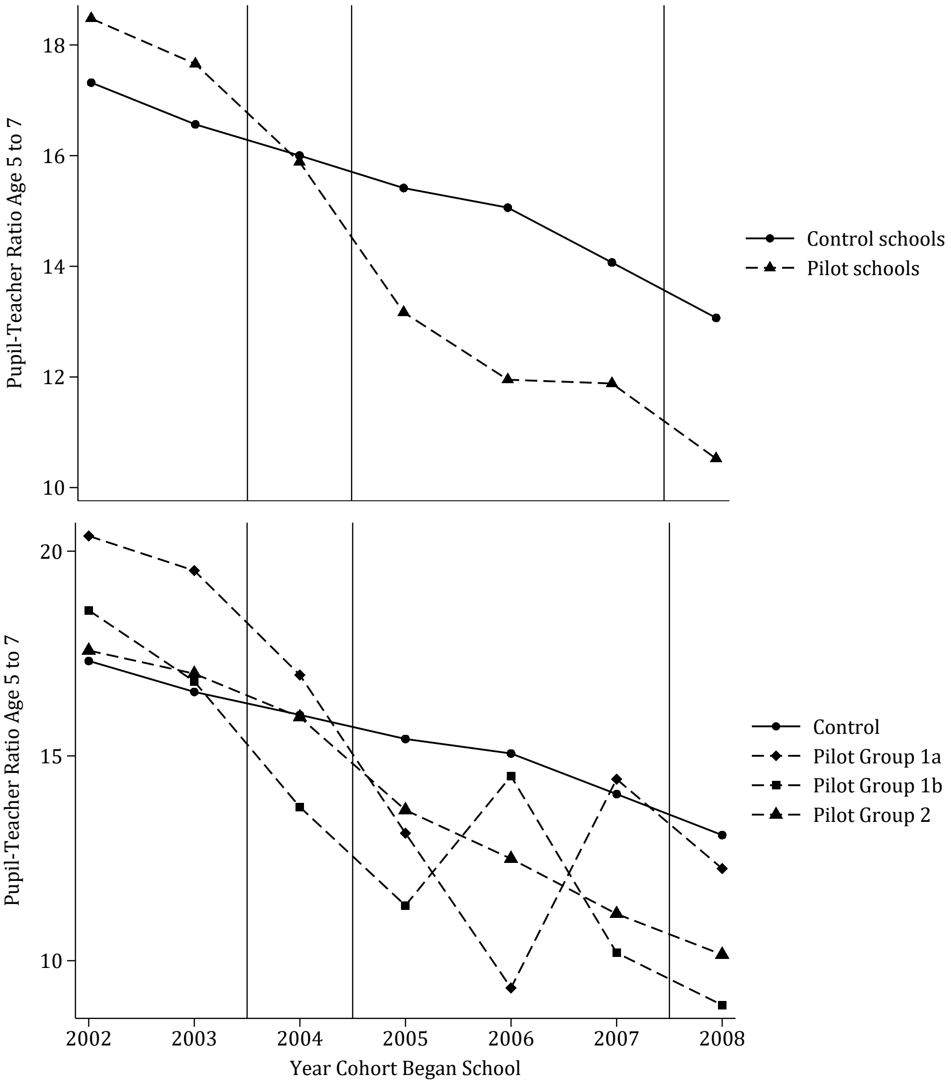
	All Pilot Schools	Pilot Group 1a	Pilot Group 1b	Pilot Group 2	Rest of Wales
Male	0.513	0.519	0.506	0.514	0.511
	(0.500)	(0.500)	(0.500)	(0.500)	(0.500)
Ethnicity					
White British	0.902	0.966	0.793	0.915	0.938
	(0.241)	(0.241)	(0.241)	(0.241)	(0.298)
Asian and Mixed White and Asian	0.051	0.013	0.108	0.045	0.024
	(0.154)	(0.154)	(0.154)	(0.154)	(0.219)
Black and Mixed White and Black	0.016	0.005	0.032	0.014	0.012
	(0.110)	(0.110)	(0.110)	(0.110)	(0.125)
Other Ethnic Groups	0.032	0.016	0.068	0.026	0.026
	(0.158)	(0.158)	(0.158)	(0.158)	(0.175)
First language					
English	0.850	0.867	0.794	0.861	0.848
	(0.359)	(0.359)	(0.359)	(0.359)	(0.357)
Welsh	0.071	0.117	0.021	0.072	0.108
	(0.310)	(0.310)	(0.310)	(0.310)	(0.257)
Non-Native	0.079	0.016	0.184	0.067	0.044
	(0.206)	(0.206)	(0.206)	(0.206)	(0.270)
Free School Meals at age six	0.324	0.195	0.248	0.386	0.189
	(0.391)	(0.391)	(0.391)	(0.391)	(0.468)
WIMD decile at age six					
(Least deprived) 1	0.014	0.051	0.015	0.002	0.098
	(0.298)	(0.298)	(0.298)	(0.298)	(0.116)
2	0.045	0.071	0.066	0.030	0.094
	(0.292)	(0.292)	(0.292)	(0.292)	(0.207)
3	0.048	0.078	0.050	0.038	0.093
	(0.290)	(0.290)	(0.290)	(0.290)	(0.214)
4	0.048	0.075	0.064	0.035	0.091
	(0.288)	(0.288)	(0.288)	(0.288)	(0.214)
5	0.063	0.050	0.132	0.047	0.100
	(0.300)	(0.300)	(0.300)	(0.300)	(0.243)
6	0.064	0.166	0.035	0.042	0.100
	(0.299)	(0.299)	(0.299)	(0.299)	(0.245)
7	0.098	0.150	0.133	0.073	0.100
	(0.300)	(0.300)	(0.300)	(0.300)	(0.298)
8	0.118	0.156	0.086	0.116	0.103
	(0.304)	(0.304)	(0.304)	(0.304)	(0.323)
9	0.160	0.145	0.224	0.146	0.107
	(0.309)	(0.309)	(0.309)	(0.309)	(0.367)
(Most deprived) 10	0.342	0.060	0.196	0.470	0.114
	(0.318)	(0.318)	(0.318)	(0.318)	(0.475)
End-of-Primary Core Subject Indicator	0.767	0.802	0.790	0.750	0.834
	(0.372)	(0.372)	(0.372)	(0.372)	(0.423)
N	11920	2241	2204	7475	233749
N of which treated	4872	1664	1363	1845	0
Schools	42	12	8	22	1346

Notes: SAIL Databank data. Sample comprises pupils who can be observed from age six, attend mainstream schools in Wales and have no missing data. Data are for both ‘pre’ and ‘post’ Foundation Phase years. Pilot Schools are further split into Groups 1a, 1b and 2, which each implement the Foundation Phase in different years. WIMD stands for the Welsh Index of Multiple Deprivation of the pupil’s home address, 2014 index, which combines data on eight domains of income, employment, health, education, access to services, community safety, physical environment, and housing. The End-of-Primary Core Subject Indicator is a binary variable indicating that a pupil is working at the age-expected level, based on teacher assessments, in English or Welsh (first language), mathematics, and science, in combination. N is the number of pupil-level observations. Standard deviations in parentheses.

As a descriptive indication of change occurring in the Pilot schools following the introduction of the Foundation Phase, I also construct a measure of pupil-teacher ratios over the years studied.[16]

[16] I urge some caution in interpreting this outcome. The pupil-teacher ratio measure is a self-created measure made by dividing the number of pupils in Grades 1 and 2 by the number of teachers assigned to different groups of children in Key Stage 1 (Grades 1, 2, or ‘M’ (mixed)), by school and year. The final cohort (2009) is dropped as measures are affected by a change in variable definition. Schools with any resultant ratios above the legal infant class size limit of 30 pupils are

Figure 2.3 Pupil-teacher ratios at ages five to seven



Notes: SAIL Databank data. Sample comprises pupils who can be observed from age six, attend mainstream schools in Wales and have no missing data. Years represent the end of each academic year, e.g. '2004' represents the academic year 2003/04. Pilot Group 1a implemented in 2004, 1b in 2005, and 2 in 2008. The pupil-teacher ratio measure is a self-created measure made by dividing the number of pupils in Grades 1 and 2 by the number of teachers assigned to different groups of children in Key Stage 1 (Grades 1, 2, or 'M' (mixed)), by school and year. The final cohort (2009) is dropped as measures are affected by a change in variable definition. Schools with any resultant ratios above the legal infant class size limit of 30 pupils are dropped (48%). The resultant sample contains 679 Primary schools, including 26 Pilot schools: 7 in Pilot Group 1a, 4 in Pilot Group 1b, 15 in Pilot Group 2. See Appendix Section B.1.2 for more details.

Recall that a reduction of pupil-teacher ratios from 1:20 to 1:15 was an intended, if non-statutory, aspect of the Foundation Phase Pilot implementation. Figure 2.3 shows the change in pupil-teacher ratios over time in Pilot and non-Pilot schools, both together and then disaggregated by Pilot group,

dropped (48%). The resultant sample contains 679 Primary schools, including 26 Pilot schools: 7 in Pilot Group 1a, 4 in Pilot Group 1b, 15 in Pilot Group 2. See Appendix Section B.1.2 for more details.

with the vertical lines representing the years in which more of the Pilot schools began delivering the Foundation Phase. The intended 1:15 ratio appears to have actually been surpassed by both Pilot and control schools on average, but by a considerably greater margin in the Pilot schools, and there is clear evidence of this substantial change to school environments as the Foundation Phase Pilot is rolled out. However, once disaggregated, the changes appear somewhat less-precisely related to the implementation year of each group, likely indicative of how variable such a measure can be; in a school with one class per cohort, creating two classes would effectively halve the pupil-teacher ratio.[17]

Outcome measures

Finally, I construct two sets of educational outcomes distinguished by the age-related timing of measurement. It is not possible to use age seven measures, which would correspond with the end of the treatment, as there are no administrative data available.[18] However I am able to use measures at ages 11 and 16, which mark the ends of distinct Key Stages of schooling in Wales and are ages at which national attainment data are collected.

The existing literature on the long run effects of early years interventions – albeit the vast majority considering preschool attendance against a counterfactual of informal or parental childcare – tends to find consistently (positive) effects on age-adequate Primary and Secondary school progression and mixed effects on test scores and school grades (Dietrichson et al., 2020). I select measures which can be similarly understood as an indication that a pupil is academically 'on track', and which it may therefore be appropriate to interpret as valid primarily for pupils close to these 'on track' thresholds (likely towards the middle of the distribution of academic attainment).[19] For age 11 this measure is the Core Subject Indicator (CSI), a binary variable indicating that a pupil is working at the age-expected level, based on teacher assessments, in English or Welsh (first language), mathematics, and science, in combination. For age 16 it is a binary variable indicating that a pupil has achieved Level 2 in their age 16 examinations, meaning that that they have achieved a secure pass in at least five GCSE or equivalent

[17] It is worth noting that evidence on the effects of class size reductions on student achievement have been mixed (Hanushek, 2020). Nonetheless, one might expect a positive effect from reductions in class size alone, and could therefore assume any measured effects to already include positive class size effects.

[18] Specifically, attendance data only begins in 2007/08, and pupils who did receive the treatment completed different age seven academic assessments to their counterparts in non-treated schools, the data from which is unavailable.

[19] Key Stage 'Levels' are available at age seven but most are one of three options, and are designed to be discrete measures rather than a continuous scale. They are also only available for pupils who did not receive Foundation Phase provision. Grades are not available in the current version of the GCSE data available to me.

qualifications, a common threshold for pupil progression into further education or training.[20] I also measure, at age 16, whether pupils enrol in the academic track for post-16 education (sixth form).[21] However it should be noted that pupils have another route of staying in education at age 16: attendance at a Further Education College, which tend to offer more vocational qualifications.[22]

For non-cognitive outcomes, I follow the recent literature in economics by using measures of behavioural outcomes at each age as proxies for non-cognitive skills (Bertrand and Pan, 2013; Heckman and Kautz, 2012; Kautz and Zanoni, 2024). At age 11, I use school attendance, defined as the proportion of possible school days attended.[23] At age 16, I measure whether a pupil has been temporarily excluded (suspended) from school during their last year of compulsory schooling.[24] These outcomes are of interest because there is evidence that changes in such skills predict larger impacts on long-run outcomes than test scores (Jackson, 2018). In addition, they create a more holistic picture of the impacts of the Foundation Phase, which aligns with the some of the Welsh Government's original intentions for the policy (National Assembly for Wales, 2003).

Not all outcomes are available or reliable for every observed year. To avoid contamination by the 2020 Covid-19 pandemic, I exclude 2020 data (the last cohort) at age 16 for exclusions and sixth form attendance, while for GCSE examinations data from 2020 is not available. It is also worth noting that GCSE examinations data are available from 2014 only, meaning that I can observe one fewer pre-treatment year than for my other outcomes.[25] To retain the largest sample that I can, I create separate samples for pupils who are observed with non-missing covariates and outcomes at age 11, at age 16, and at 16 in the first academic year of elective schooling (henceforth 'age 16 with sixth form'). As discussed further in Section 2.5.2, I find evidence of slightly higher sample attrition from the Pilot schools compared to the control schools but no significant compositional change effects from

[20] Another, closely-related measure at age 16 examinations would be 'Level 2 Plus', meaning that pupils have achieved a secure pass in at least five GCSE or equivalent qualifications including English or Welsh (first language) and maths. However unfortunately this benchmark is not available in the existing data until 2017, and thereby unusable in this analysis.

[21] I drop pupils who are no longer present in the Welsh Demographic Service dataset as I would otherwise be unable to rule out the possibility that pupils have moved out of Wales.

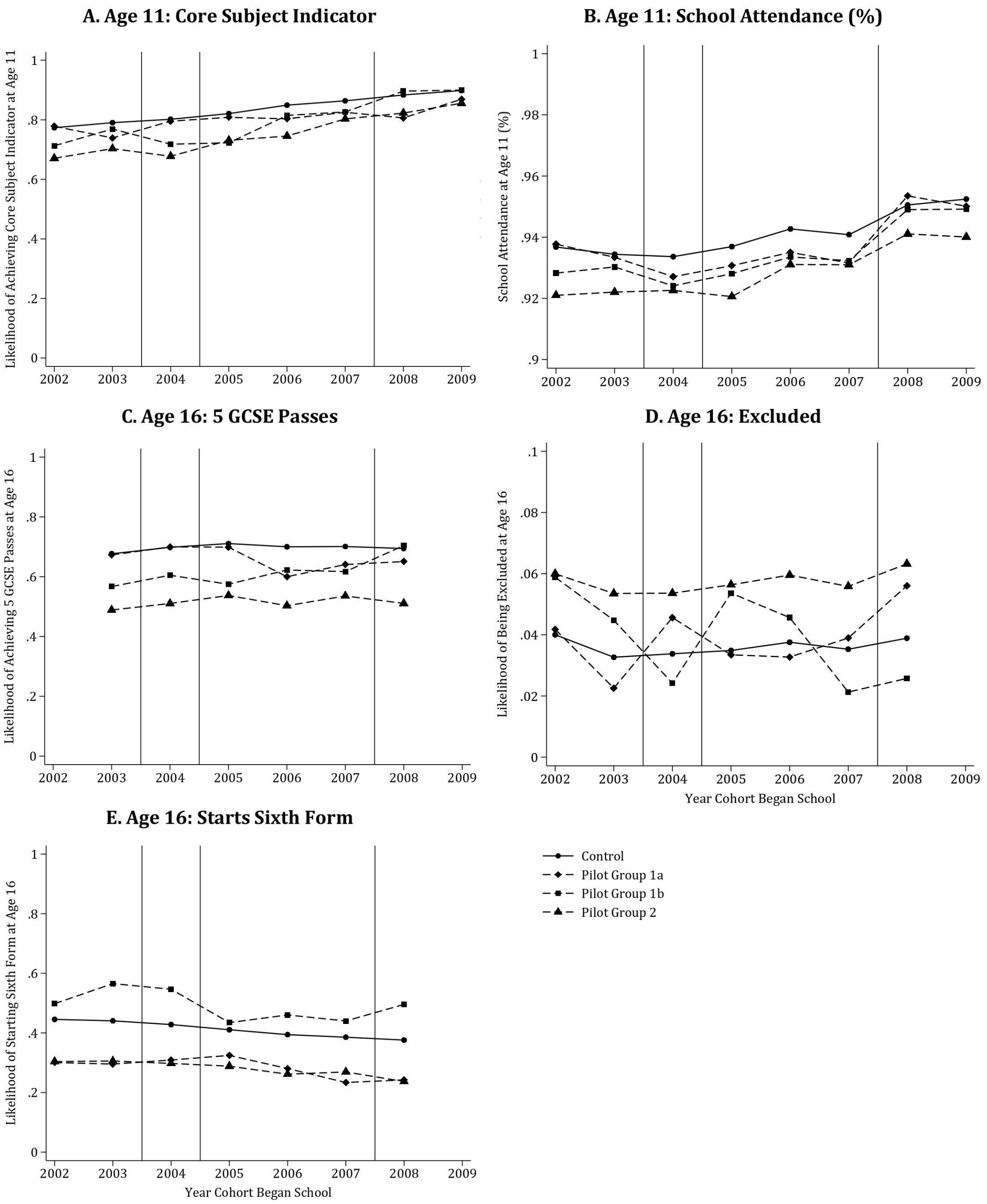
[22] Data on Further Education enrollments appear to be available and, once this is confirmed, will be sought for a later version of this paper.

[23] I treat both authorised and unauthorised absences as non-attendance. Attendance at age 16 is difficult to interpret due to the use of 'study leave' from school, prior to examinations, and so not included here.

[24] I use only the last year as the exclusions data begins late, in 2012/13, meaning that any earlier age would require less robust pre-trends testing.

[25] A different version of the GCSE examinations data has been requested.

Figure 2.4 Unadjusted trends in outcomes over time



A. Age 11: Core Subject Indicator

B. Age 11: School Attendance (%)

C. Age 16: 5 GCSE Passes

D. Age 16: Excluded

E. Age 16: Starts Sixth Form

Notes: SAIL Databank data. Sample comprises pupils who can be observed from age six, attend mainstream schools in Wales and have no missing data. Years represent the end of each academic year, e.g. '2004' represents the academic year 2003/04. Pilot Group 1a implemented in 2004, 1b in 2005, and 2 in 2008. Not all outcomes are available for all years due to: (i) lack of coverage in the available data tables, or (ii) contamination with 'Covid' years (see Appendix Section B.1.1). 'Core Subject Indicator' is a binary variable indicating that a pupil is working at the age-expected level, based on teacher assessments, in English or Welsh (first language), mathematics, and science, in combination. '5 GCSE Passes' a binary variable indicating that a pupil has achieved Level 2 in their age 16 examinations, a common threshold for pupil progression into further education or training. Exclusions are fixed-term exclusions only. Sixth form is the non-compulsory academic track for advanced schooling at age 16.

the pre-Pilot to post-Pilot periods across any demographic. Trends in the outcome measures over time, disaggregated by Pilot group, are shown in Figure 2.4.

2.4 Empirical strategy

2.4.1 Difference-in-differences model

The goal of this paper is to identify the causal impact of the Foundation Phase Policy on later educational outcomes. To do this, I exploit the staggered roll-out of the Foundation Phase Pilot across schools in Wales between 2004 and 2008, and leverage in particular the fact that two groups of schools implemented the Foundation Phase before the national roll-out as part of a Pilot program. Under a set of assumptions described below, the quasi-experimental variation generated by the staggered roll-out of the Foundation Phase Pilot allows me to estimate the causal impact of the Foundation Phase on later educational outcomes using a difference-in-differences strategy. Comparing consecutive cohorts of pupils attending schools, the strategy compares the before-after differences in outcomes between pupils in schools which implemented the Foundation Phase Pilot and pupils in schools that did not implement the Foundation Phase Pilot. In other words, I estimate the differences in outcomes for pupils in the cohorts before the Foundation Phase was rolled out in their school to pupils in cohorts after, net the average changes in outcomes that occur for all schools in the rest of Wales.

I use a staggered difference-in-differences design to account for the fact that the Foundation Phase Pilot was implemented to different groups of schools at different times. The basic model which I estimate can be written as follows:

$$Y_{isgt} = \alpha + \beta FP_{gt-b} + \gamma \mathbf{X}_i + \delta FSM_{it} + \eta (FSM_{it} \times PDG_t) + \theta_s + \lambda_t + \varepsilon_{isgt} \tag{2.1}$$

where Y_{isgt} represents an outcome for individual i in year t who attended Primary school s that belongs to Pilot Group g; FP_{gt-b} is an indicator for whether, in year $t-b$ (b={6,10,11}), the Foundation Phase was being implemented at schools in Pilot Group g; θ_s indicates school fixed effects; λ_t indicates year fixed effects; and ε_{isgt} is an error term. Standard errors are clustered at the school level.

$\mathbf{X}_i$ is a vector of time-invariant pupil demographics. Though the difference-in-difference design does not require treatment and control groups to match on observable characteristics or outcomes prior to the implementation of the treatment, it does require an assumption that the *trends* in the outcomes of the two groups would have been parallel in the absence of the treatment. It is therefore necessary to control for any characteristics that differ between the two groups and which might affect their trends in outcomes over time. This instead ascribes a *conditional* parallel trends assumption. The time-invariant pupil demographic controls that I include are gender (male or female), ethnicity (White British, Asian or Mixed White and Asian, Black or Mixed White and Black, or Other Ethnic Group), and first language (English, Welsh, or non-native). I also control for decile of Welsh Index of Multiple Deprivation, which measures the neighbourhood deprivation of the pupil's home address, at age six, and for a pupil's eligibility for free school meals (FSM) contemporaneously to the outcome measured (FSM_{it}).[26] Any observable and unobservable time-invariant differences between schools are accounted for by the school fixed effects.

The difference-in-differences design also requires consideration of any coincident policy changes which might affect trends in outcomes over time. The Pupil Deprivation Grant (PDG) was introduced in 2013 and provides extra funding for schools for each FSM-eligible pupil on roll.[27] The implementation of the PDG was after the Foundation Phase Pilot roll-out, when the Pilot school pupils were aged eight or above. Despite the PDG occurring in all schools, which would mean that it any resultant improvements in the attainment of eligible pupils might be assumed to be a time trend which is 'differenced out' by the difference-in-differences design, its introduction might bias my estimates upwards as PDG funding is both correlated with the treatment (the treatment settings are more deprived than the national average) and with the outcomes of interest. I therefore create a variable – PDG_t –

[26] I do not control for WIMD at later ages to account for the possibility that a pupil's home address may have changed in response to being in receipt of the Foundation Phase pilot. Similarly, I have no measure of cognitive or non-cognitive skills at an age prior to when pupils would have received the Foundation Phase, and do not include any measures after this age due to the risk that any such measures would themselves be affected by the treatment. However I do control for a pupil's eligibility for free school meals (FSM) contemporaneously to the outcome measured (i.e. at later ages), as free school meals eligibility is determined by parents' receipt of income-determined benefits which is very unlikely to be affected by their child being in receipt of the Foundation Phase Pilot.

[27] PDG funding was extended to looked after children in 2013/14. However unfortunately I am not able to observe whether children are classified as looked after in the data due to data sensitivity precautions.

which is equal to the amount of per-pupil additional funding that schools received, and interact this with the contemporaneous pupil-level FSM-indicator.[28]

As highlighted by the recent literature on difference-in-differences, for equation (2.1) to recover an un-biased average treatment effect on the treated (ATT) we must also assume treatment effect homogeneity both across treated groups and across time. This is because two-way fixed effects (TWFE) models, such as equation (2.1), implicitly use already-treated comparison groups (so called 'forbidden comparisons') and assign larger weights to larger groups and groups treated closer to the middle of the panel (Baker et al., 2022; De Chaisemartin and d'Haultfoeuille, 2023; Roth et al., 2023). In this evaluation approximately 95% of schools are never treated during the evaluation period, which somewhat lessens concerns around forbidden comparisons as any already-treated schools will be a very small proportion of the entire control group. Nonetheless, to address the potential biases of the TWFE estimator I implement all difference-in-differences analyses by estimating equation (2.1) using the estimator suggested by Borusyak et al. (2024). This estimator imputes potential outcomes for the treated group in the absence of treatment using only data on untreated units. The ATT is then estimated as the average difference between the predicted and observed outcomes across all treated pupils, with standard errors clustered at the school level. A key benefit of using the Borusyak et al. (2024) estimator in this setting is its flexibility, especially in adjusting for time-varying covariates such as those relating to free school meal eligibility and the Pupil Deprivation Grant. The Borusyak et al. (2024) estimator is also generally more efficient compared to other recent estimators, such as those proposed by Callaway and Sant'Anna (2021) and Sun and Abraham (2021).

2.4.2 Parallel trends

I test for conditional parallel trends prior to the introduction of the Foundation Phase as a way of providing evidence in support of the assumption that the Pilot and control schools would have trended similarly, conditional on controls, in the absence of the introduction of the Foundation Phase Pilot. I implement the pre-trends test suggested by Borusyak et al. (2024), which avoids the bias of heterogeneous treatment effects incurred when deviations of parallel pre-trends are jointly estimated with post-treatment effects (Roth, 2022). I first estimate an event-study version of equation (2.1) with

[28] The Pupil Deprivation Grant funding that schools received per free school meals pupil was £450 from 2013/14, £918 from 2014/15, and £1150 from 2015/16.

indicators for distance to the introduction of the Foundation Phase, but using *only* never-treated and pre-treated observations. Specifically, I estimate the following specification:

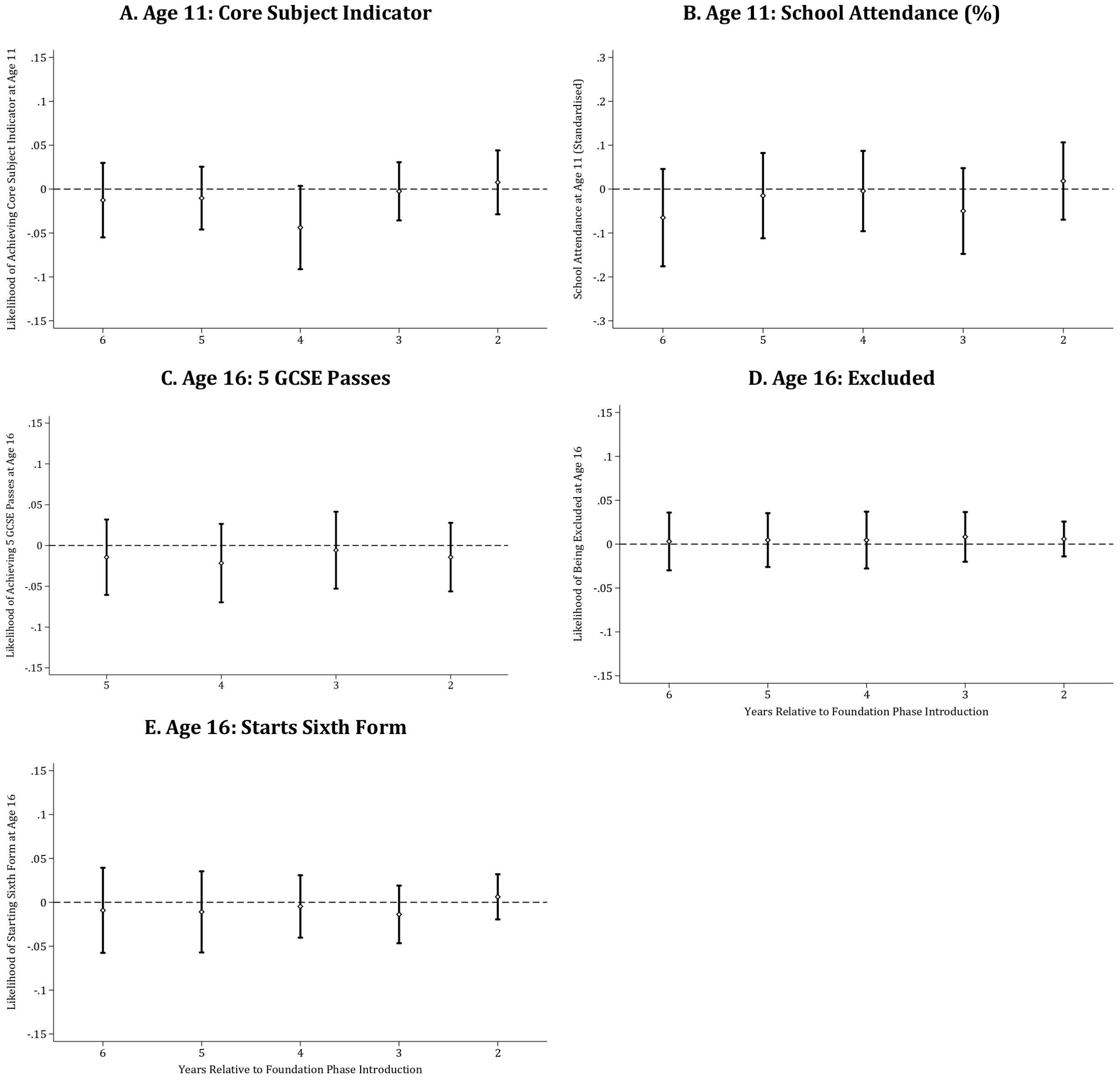
$$Y_{isgt} = \alpha + \beta_k^{pre} \sum_{k=-6}^{-1} D_{k(gt-b)} + \gamma \mathbf{X}_i + \delta FSM_{it} + \eta(FSM_{it} \times PDG_t) + \theta_s + \lambda_t + \varepsilon_{isgt} \qquad (2.2)$$

where $D_{k(gt-b)}$ is a set of indicator variables that take value one if, for group g in year $t-b$ (b={6,10,11}), the introduction of the Foundation Phase was k years away. I treat pupils in the cohort one year before the Foundation Phase was rolled out to their group of Pilot schools as the omitted category and compare them to pupils in the other cohorts.[29] The β_k^{pre} coefficients then estimate whether the difference in the conditional outcomes for the Pilot and control groups k years before the Foundation Phase Pilot was introduced are statistically significantly different from that in the year directly preceding its introduction. If no β_k^{pre} estimates are statistically significantly different from zero then the pre-trends can be said to be conditionally parallel. Figure 2.2 provides a visual test of each of the β_k^{pre} estimates individually by displaying conventional event-study plots for the *pre*-Pilot years only, for all outcomes.

I also formally test for parallel pre-trends by testing the joint significance of different β_k^{pre} coefficients using an F-test. The results are displayed in Table 2.2. As an example, the row labelled '3-year horizon' presents the test that the estimates for β_2^{pre} and β_3^{pre} are jointly equal to zero, and the row labelled '4-year horizon' the test that the estimates for β_2^{pre}, β_3^{pre} and β_4^{pre} are jointly equal to zero. As $p < 0.1$ for no horizon and outcome combination in Table 2.2, I never reject parallel trends. It should be noted that only the 2-year horizon test is balanced (contains data from all Pilot groups) with regard to pre-treatment Pilot schools; longer horizons are unbalanced and rely more heavily on data from Pilot Groups 1b and 2. In addition, as I am currently unable to observe the first year of data for pupils' academic attainment at age 16, we only observe one pre-treatment year for this outcome

[29] I deviate marginally from the method outlined by Borusyak et al. (2024) by excluding as the reference year the year directly preceding the introduction of the Foundation Phase, rather than the earliest observed period. I do this because my panel is unbalanced such that the estimates just one and two years away from the introduction of the Foundation Phase include all Pilot schools, while those a greater number of years away include only Pilot Group 1b or Pilot Group 2.

Figure 2.5 Pre-trends graphs



Notes: SAIL Databank data. Sample comprises pupils who can be observed from age six, attend mainstream schools in Wales and have no missing data. Estimates from equation (2.2) on never-treated and pre-treated observations of pupils in treated schools only. Sample sizes are as follows, by figure: (A) 232087; (B) 232087; (C) 164159; (D) 195180; (E) 193921. Additional controls in underlying regression: gender, ethnicity, language, Welsh Index of Multiple Deprivation at age six, free school meal eligibility at the age of measurement, free school meal eligibility at the age of measurement interacted with Pupil Deprivation Grant funding, Primary school fixed effects, year fixed effects. X-axis indicates moving forward in time, e.g. ‘5’ means 5 years before the Pilot implementation.

for Pilot Group 1a, which is the omitted year, meaning that Pilot Group 1a does not contribute to β_2^{pre} in column (3).[30]

[30] A different version of the GCSE examinations data has been requested.

Finally, it is possible that the different Pilot groups separately violate the conditional parallel trends assumption and yet offset one another, resulting in no violations in aggregate. This would be concerning as the offsetting could have occurred purely by chance, undermining the validity of the aggregate tests. I check for such offsetting by completing the parallel pre-trends tests outlined above but separately for each Pilot group. To allow us to identify offsetting trends it is useful to see the direction of any pre-period coefficient estimates. I therefore present separate pre-period event study figures differentiated by Pilot group and outcome in Appendix Figure B.1, and formal pre-trends tests in Appendix Table B.4. Overall the results indicate that we should not be overly concerned about parallel trends assumption violations in the separate Pilot groups. The one exception is for the Age 11 Core Subject Indicator results, where both Pilot Group 1b and Pilot Group 2 display a single β_k^{pre} which is statistically different from zero. It is also worth noting that, by thus restricting the sample of Pilot schools to fewer schools and thereby fewer clusters, I am greatly reducing the power of the statistical tests and therefore results should be interpreted with caution.

Table 2.2 Tests for parallel pre-trends

	(1) Age 11: Core Subject Indicator	(2) Age 11: School Attendance (Standardised)	(3) Age 16: 5 GCSE Passes	(4) Age 16: Excluded	(5) Age 16: Starts Sixth Form
2 year horizon					
F	0.174	0.169	0.443	0.335	0.228
P(F)	0.677	0.681	0.506	0.563	0.633
3 year horizon					
F	0.149	1.056	0.222	0.202	0.813
P(F)	0.861	0.348	0.801	0.817	0.444
4 year horizon					
F	1.504	0.708	0.463	0.147	0.565
P(F)	0.212	0.547	0.708	0.932	0.638
5 year horizon					
F	1.140	0.577	0.363	0.118	0.451
P(F)	0.336	0.680	0.835	0.976	0.772
6 year horizon					
F	0.957	0.609		0.106	0.377
P(F)	0.443	0.693		0.991	0.865
N	232087	232087	164159	195180	193921

Notes: SAIL Databank data. Sample comprises pupils who can be observed from age six, attend mainstream schools in Wales and have no missing data. Estimates from equation (2.2) on never-treated and pre-treated observations of pupils in treated schools only. Additional controls in underlying regression: gender, ethnicity, language, Welsh Index of Multiple Deprivation at age six, free school meal eligibility at the age of measurement, free school meal eligibility at the age of measurement interacted with Pupil Deprivation Grant funding, Primary school fixed effects, year fixed effects. Standard errors clustered at the school level.

2.5 Results

2.5.1 Main results

Table 2.3 presents estimates of β in equation (2.1) on five medium-term outcomes of the Foundation Phase policy, two measured at age 11 and three measured at age 16. Each result shows the estimated effect of the policy for pupils who received Foundation Phase Pilot provision compared to pupils who instead received formal schooling between ages five and seven.

The results indicate that the Foundation Phase Pilot had no measurable effects on pupils' school-related outcomes and behaviours at ages 11 or 16. Specifically, as none of the estimates reach conventional levels of statistical significance, I have insufficient evidence to reject the null hypothesis that any of the effects are zero in the population. These results are in line with the null effects found by Biroli et al. (2018) in their evaluation of the Reggio Emilia approach in Italy, which is to the best of my knowledge the only other study which has evaluated the medium- to long-term effects of a play-based and child-centered curriculum and pedagogy compared to other types of early childhood education. In addition, these null effects are important and interesting in and of themselves. They tell us that, on average, children who received two fewer years of direct instruction during early Primary education still achieved similarly to their peers in later education, made similar school-related decisions, and exhibited similar school-related behaviours.

There are also limits to the conclusions that we can draw from these results. For example, the results are unable to tell us whether there were any contemporaneous effects at age seven which I am unable to measure in these data, but which fade out by the time children reach age 11 and age 16, or any so-called 'sleeper' effects which only appear at ages older than those assessed (García et al., 2021; Gray-Lobe et al., 2023; Heckman et al., 2013). In addition, as I have previously stated, I am limited in the outcomes that I can measure in the present study. Should there be any time-sensitive skills and capacities which greater play-based learning afforded the pupils receiving the Pilot provision but which I am unable to measure – for example positive attitudes towards learning, improved wellbeing

Table 2.3 Main results

	(1) Age 11: Core Subject Indicator	(2) Age 11: School Attendance (Standardised)	(3) Age 16: 5 GCSE Passes	(4) Age 16: Excluded	(5) Age 16: Starts Sixth Form
Treatment Effect (ATT)	0.014	-0.019	-0.014	0.002	-0.017
	(0.012)	(0.031)	(0.015)	(0.006)	(0.018)
School fixed effects	✓	✓	✓	✓	✓
Year fixed effects	✓	✓	✓	✓	✓
Controls	✓	✓	✓	✓	✓
Mean	0.831	0.020	0.690	0.037	0.408
Proportion treated	0.048	0.048	0.048	0.048	0.048
N	236706	236706	167190	198211	196934

Notes: SAIL Databank data. Sample comprises pupils who can be observed from age six, attend mainstream schools in Wales and have no missing data. Estimates from equation (2.1) using the estimator proposed by Borusyak et al. (2024). Controls: gender, ethnicity, language, Welsh Index of Multiple Deprivation at age six, free school meal eligibility at the age of measurement, free school meal eligibility at the age of measurement interacted with Pupil Deprivation Grant funding, Primary school fixed effects, year fixed effects. Standard errors in parentheses, clustered at the school level. * Significant at 10%; ** significant at 5%; *** significant at 1%.

or active citizenship – then these results suggest that it might be possible for children to develop those skills *and* still 'catch up' with their peers on standard measures.[31]

While acknowledging the lack of statistical significance, it is interesting to briefly note the direction and magnitude of the estimates in Table 2.3. The point estimate of 0.014 in column 1 indicates that pupils who received Foundation Phase Pilot provision may have been around 1.4 percentage points more likely to achieve the Core Subject Indicator at age 11 compared to their peers who instead received formal education between ages five and seven. This result aligns with the overall positive effect for the age 11 teacher assessments found in the previous Foundation Phase Pilot evaluation (Davies et al., 2015). However, the point estimates for the other outcomes show a less favourable picture: they are negative for standardised attendance rates at age 11, negative for the likelihood of achieving five passes in school leaving examinations (GCSEs) at age 16, positive for the likelihood of being excluded from school at age 16, and negative for the likelihood of enrolling in the academic track for post-16 education (sixth form). Taken together, these patterns raise important questions about the longer-term implications of the Foundation Phase Pilot, particularly in relation to academic attainment and engagement beyond Primary school.

Finally, as my control group is very large, the results in Table 2.3 which use the Borusyak et al. (2024) estimator should be very similar to those estimated using a TWFE estimator. As shown in Appendix Table B.5, this is indeed the case: in general, the TWFE results indicate a slight positive

[31] Though not directly comparable to the present evaluation, non-cognitive skills have generally been found to be the primary mechanisms of positive long-run effects of preschool access, while effects on test-scores fade out (Havnes and Mogstad, 2015; Heckman et al., 2013).

bias compared to the Borusyak et al. (2024) results, but in every case the significance and direction of the estimates is the same and the magnitudes similar.

2.5.2 Threats to identification and robustness checks

The key assumption of my difference-in-differences research design is the conditional parallel trends assumption, evidence in support of which is provided previously in Section 2.4.2. In this section I discuss further threats to the validity of my identification strategy and probe the stability of my estimates to robustness checks and alternative specifications.

Compositional change and attrition

First, different types of families may have been more or less likely to enroll their children in Pilot schools because of the change in pedagogy for children between ages five and seven, threatening the assumption of parallel trends. In Wales, parents and carers apply for school places for their children, and different types of schools have different admissions procedures. The Local Authority (LA) is responsible for making admissions decisions for most schools apart from some religious schools for whom admissions decisions are made by a governing body admissions panel, and indeed 95% of the Pilot schools are those for which the LA made admissions decisions. This gives us some assurance that the schools themselves generally had little discretion over the pupils that they could admit.[32] Nonetheless it could still be that different parents apply to those schools. Although it is not possible to check whether there are any *unobservable* changes in the types of children attending Pilot schools over time, the first column of Appendix Table B.6 presents checks for the balance of their observable characteristics. It shows that, along the demographic characteristics in the data available, there are no meaningful compositional changes following the introduction of the Foundation Phase.[33]

Second, there may be different follow-up rates at ages 11 and 16 for Pilot school pupils compared to control group pupils. Specifically, despite there being no evidence of considerable compositional change in the pupils who attended Pilot schools at age six before and after treatment, nonrandom

[32] Broadly, LA admissions procedures must give some attempt to comply with parents' preferences while being "procedurally fair and are also equitable for all groups of children" (Welsh Government, 2009).

[33] 0.9% fewer Black and Mixed White and Black pupils attend Pilot schools in the period following the Foundation Phase introduction, though as this group only make up around 1.2% of all pupils in Wales this is not likely to meaningfully impact the results. The other changes are primarily in the deciles of the Welsh Index of Multiple Deprivation measure, where, for example, consecutive measures tend to increase and decrease in balance.

attrition occurring simultaneously to treatment may generate selection bias in the longer-run outcomes. Overall attrition in the sample is quite low, with 96.4% of pupils in the original sample being retained in the age 11 sample, 91.9% in the age 16 sample, and 91.3% in the age 16 with sixth form sample. As shown in Appendix Table B.7, attrition is slightly higher in Pilot schools, by 0.4 percentage points at age 11, 1.6 percentage points at age 16, and 1.5 percentage points at age 16 with sixth form. Nonetheless, these differing rates of attrition would only introduce bias into my results if the observable (and unobservable) characteristics of pupils leaving the sample change over time and concurrently with the introduction of the Foundation Phase pilot. I check this by repeating the compositional change analysis but for the age 11 and age 16 samples. The results, in the second, third, and fourth columns of Appendix Table B.6, show almost identical patterns in compositional change as in the age six sample. This provides reassurance that non-random attrition does not bias the results.

Contemporaneous policies

Any policies which occurred during the observed period and may differentially impact Pilot schools are a threat to identification because they could contaminate the measured effects of the Foundation Phase Pilot. Equation (2.1) already conditions explicitly for the Pupil Deprivation Grant, which provides extra funding for schools according to the number of FSM-eligible pupils they have on roll. However another such contemporaneous policy is Flying Start, a multi-faceted support service provided for the families of children up to the age of four and living in specific areas of high deprivation.[34] Pilot Group 2 schools were specifically selected to be in Flying Start areas, which means that I run the risk of picking up a ‘Flying Start effect’ in the Foundation Phase treatment effect estimations. However the timing of Flying Start – which was launched in 2007 – is helpful in this respect. This is because the first cohort of reception children who could have potentially received any Flying Start provision are those born in 2003/04, namely the last cohort of Pilot Group 2 pupils. In addition, although these pupils may have received a single year of Flying Start provision, this would not have constituted anything like ‘full’ treatment by Flying Start.[35] Nonetheless I explore the possible effects of this

[34] Flying Start support included free part-time childcare, enhanced health visiting, parenting guidance, and interventions to address communication difficulties. Children stopped receiving any provisions once they started in school.

[35] A Welsh Government evaluation report considers children to have ‘received’ Flying Start only if they could have potentially received the provision for two or more years because so much of the provision (specifically childcare and health visits) was only available to children aged two to three (Wilton and Davies, 2017).

contemporaneous policy in Appendix Table B.8 by excluding the last cohort of pupils, who may have received some Flying Start provision. I find, in column 1, a slightly lower point estimate for the likelihood of achieving the Core Subject Indicator at age 11, but overall no changes to the direction or significance of the results.[36]

Sample selection and treatment identification

I complete three robustness checks to ensure that my main results are robust to changes in sample selection and treatment identification. First, as discussed in Section 2.3.1, due to data unavailability I have manually identified 13 schools from Pilot Group 1 (named Pilot Group 1a) who are likely to have mixed-age classes and therefore effectively implement the Foundation Phase a year (cohort) early. To check whether performing this separation meaningfully changes my results, Appendix Table B.9 re-combines Pilot Group 1a and Pilot Group 1b into a single group: Pilot Group 1. As such, one balanced pre-intervention year is gained, but it is possible that anticipation effects will contaminate the results. However I find no changes to the direction or significance of the results apart from age 16 exclusions, which remains insignificant but the point estimate changes sign.

Second, to maximise the pre-treatment years available in my data, I previously identified pupils attending Pilot schools at ages six to seven – in the final (second) year of Foundation Phase implementation – rather than when they are ages five to six. However I also find evidence that around five percent of pupils are in different schools at ages six to seven to the schools that they were in at ages five to six (see Appendix Table B.3). Although, as previously discussed, this mobility rate is not statistically different for pupils attending Pilot schools or for pupils attending Pilot schools in the post-implementation period, in Appendix Table B.10 I restrict the sample to those pupils who I can observe from ages five to six rather than just ages six to seven. As such, one pre-intervention year is lost, but I am able to re-assign treatment status to pupils who definitely attend Foundation Phase pilot schools from age five to age seven. I find no changes to the direction or significance of the results.

Third, recall that the Foundation Phase Pilot was introduced in a mixture of Infant schools (ages four to seven) and Primary schools (ages four to eleven). To ensure that I can observe schools across time, I previously assumed that any Infant and Junior (ages seven to eleven) schools that amalgamated

[36] It is worth noting that this cohort is excluded from all age 16 measures in the main results due to either a lack of data availability or to avoid the potential contamination of a ’Covid’ effect.

to create Primary schools during the observation period served the same pupils. However I also find that the rate of school mobility is slightly higher for merging schools than for non-merging schools (see again Appendix Table B.3). Appendix Tables B.11 and B.12 therefore restrict the sample of schools to those which we observe for every cohort and those which never experience schools mergers, respectively. I find no changes to the direction or significance of the results in either case.

Timing of diagnoses of Special Educational Needs and ADHD

Changes in curriculum and pedagogy, as well as a reduction in pupil-teacher ratios, may mean that teachers delivering Foundation Phase provision identify different types and levels of skill and competency in their pupils compared to teachers delivering more formal schooling. Specifically, Foundation Phase Pilot pupils may be more or less likely to be identified as having a Special Educational Need (SEN) or Attention Deficit Hyperactivity Disorder (ADHD) between ages five and seven, leading to earlier or later provision of support.[37] If SEN and ADHD diagnoses are more or less likely to occur in Pilot schools relative to non-Pilot schools because of the policy, this would constitute an unintended part of the treatment effect. Moreover, such diagnoses have inbuilt persistence, meaning that early intervention could forestall or reduce later negative effects.

I investigate the effect of the Foundation Phase Pilot on the likelihood of pupils being diagnosed as SEN or ADHD by age seven. For SEN, I identify the first record of any Special Educational Need between ages six and seven, as my education dataset only begins consistently at age six. For ADHD, I identify the first record of ADHD in primary or secondary healthcare between ages five and seven, additionally controlling for any diagnoses before age five.[38] The first three columns of Appendix Table B.13 shows the pre-trends tests and ATT estimates when I use SEN and ADHD diagnoses as outcome measures. In each case the outcome measure is the likelihood of receiving a diagnosis, by

[37] For example, Nicodemo et al. (2024) find evidence of higher ADHD rates among early school starters caused by peer-comparison bias and differences in relative age among classmates.

[38] To be precise, I construct an indicator for individuals who were only registered with contributing GP practices between ages zero and seven, and restrict my ADHD outcome sample to these individuals. Following John et al. (2022), I use validated Read v2 codes to identify children with primary care (GP) records for ADHD (including diagnoses, symptoms, and prescriptions) in the Welsh Longitudinal General Practice dataset, and ICD-10 diagnostic codes to identify children with hospital admissions or outpatient appointments for anxiety or depression in the Patient Episode Dataset for Wales or the Outpatient Dataset, respectively. Unfortunately, the proportion of records in the Outpatient Database for Wales that have been clinically coded is not sufficient to reliably identify diagnosis. I therefore report a sensitivity check excluding the Outpatient Database for Wales, to ensure that their exclusion does not considerably change the main conclusions. Appendix Section B.1.3 indexes the individual codes used.

age seven, as either SEN or ADHD. The results show null effects across all outcomes. In other words, there is no difference in the likelihood of pupils attending Foundation Phase Pilot schools receiving a diagnosis of a Special Educational Need or ADHD by age seven compared to their peers who did not attend Foundation Phase pilot schools.

Secondary school quality

Finally, children who attended Foundation Phase Pilot schools may have been more or less likely to attend higher or lower quality Secondary schools that their peers from the same Primary school but before the introduction of the Foundation Phase, and compared to their peers in the rest of Wales. This is a threat to identification as any measured effects at age 16 could be driven by the quality of the Secondary schools attended, rather than the Foundation Phase itself. To investigate this possibility, I create a basic school value-added measure by regressing whether a pupil has achieved five GCSE passes at age 16 on whether they achieved the Core Subject Indicator at age 11 and adjusting for gender, ethnicity, language, Welsh Index of Multiple Deprivation at age six, free school meal eligibility at age 16, the Pupil Deprivation grant funding, and indicators for each Secondary school. Importantly, I run these regressions separately by year and *excluding pupils who attended Foundation Phase Pilot schools*. For each year, I then extract the coefficients for each Secondary school and create standardised outcome, and use this measure as an outcome variable for equations (2.1) and (2.2). The β coefficient reported in the fourth column of Appendix Table B.13 therefore measures any changes in the quality of the Secondary schools attended by pupils in receipt of the Foundation Phase Pilot. This shows a positive estimate but a null effect, and as such provides reassurance that any measured effects in other outcomes at age 16 are unlikely to be driven (or counteracted) by Pilot school pupils attending relatively higher-quality Secondary schools.

2.5.3 Heterogeneity

Reduced differential attainment across pupil groups was an explicit goal of the Foundation Phase policy for the Welsh Government (Maynard et al., 2013). In this final section I therefore perform subgroup analysis to explore whether there are different effects of the Foundation Phase Pilot for pupils with different characteristics (gender and socioeconomic status). I also explore whether the estimated effects appear to differ across the different Pilot groups.

Main results by pupil characteristics

I begin by exploring how the effects of the Foundation Phase pilot differ according to the gender of the pupil. Fidjeland et al. (2023) find that, in Norway, boys benefit considerably more than girls from the introduction of formal schooling against a counterfactual free-play approach. One possible mechanism of these effects is that girls develop faster than boys in domains like vocabulary and socioemotional skills, and are thus likely to have an initial skill advantage and to spend more time engaging in activities that promote school readiness and skill development, compared to boys (Dietrichson et al., 2020). Indeed Power et al. (2019) find, in a qualitative evaluation of the full roll-out of the Foundation Phase, that girls and boys differed in the types of learning they were exploring; girls were more likely to be involved in self-directed learning, to be engaged in activities at workstations and desks, and to interact with their peers and adults, than boys. Existing quantitative evaluations of the Foundation Phase Pilot also find inequalities by gender favouring girls for age 11 teacher assessments, however this analysis was only completed for schools in Pilot Group 1 (Davies et al., 2013).

Panel A in Table 2.4 shows the main results for male and female pupils separately. For these estimates, potential outcomes are imputed for the treated group in the absence of treatment using data on untreated units in the full sample, as in Table 2.3, to retain the statistical power and reliability of the estimates. The ATT is then estimated as the average difference between the predicted and observed outcomes across all treated pupils of the given gender, with standard errors clustered at the school level.

The results remain insignificant for both groups, meaning that we have insufficient evidence to conclude that any of the effects are different from zero in the population. However, considering the point estimates alone, the effects do appear to be in line with the previous evaluations and existing literature as they are tentatively more positive for girls than for boys. This is primarily apparent for outcomes at age 16, where the point estimates in columns 3 and 4 suggest that male pupils who attended Foundation Phase Pilot schools at are on average *less* likely to achieve five passes in their school leaving examinations (GCSEs) and *more* likely to be excluded from school than their male

Table 2.4 Heterogeneous results

	(1) Age 11: Core Subject Indicator	(2) Age 11: School Attendance (Standardised)	(3) Age 16: 5 GCSE Passes	(4) Age 16: Excluded	(5) Age 16: Starts Sixth Form
Panel A:					
Male					
Treatment Effect (ATT)	0.012	-0.018	-0.017	0.007	-0.014
	(0.013)	(0.036)	(0.015)	(0.008)	(0.023)
Mean	0.800	-0.000	0.633	0.052	0.371
Proportion treated	0.048	0.048	0.048	0.048	0.048
N	120871	120871	85006	100850	100081
Female					
Treatment Effect (ATT)	0.016	-0.020	0.011	-0.004	-0.021
	(0.012)	(0.032)	(0.020)	(0.006)	(0.016)
Mean	0.863	0.041	0.749	0.022	0.448
Proportion treated	0.048	0.048	0.048	0.048	0.048
N	115835	115835	82184	97361	96853
Panel B:					
Free School Meals					
Treatment Effect (ATT)	0.034*	-0.054	0.022	0.003	0.001
	(0.020)	(0.056)	(0.026)	(0.013)	(0.026)
Mean	0.677	-0.433	0.407	0.076	0.221
Proportion treated	0.081	0.081	0.078	0.077	0.078
N	43201	43201	24816	29679	29501
Non-Free School Meals					
Treatment Effect (ATT)	0.006	-0.006	-0.023	0.002	-0.021
	(0.010)	(0.031)	(0.016)	(0.006)	(0.019)
Mean	0.865	0.121	0.739	0.030	0.442
Proportion treated	0.041	0.041	0.043	0.043	0.043
N	193505	193505	142374	168532	167433
Panel C:					
Pilot Group 1a					
Treatment Effect (ATT)	-0.012	-0.082*	-0.050**	0.007	0.004
	(0.021)	(0.042)	(0.022)	(0.010)	(0.036)
Mean	0.834	0.027	0.696	0.036	0.412
Proportion treated	0.010	0.010	0.010	0.010	0.010
N	227432	227432	160732	190547	189312
Pilot Group 1b					
Treatment Effect (ATT)	0.014	0.020	0.022	-0.006	-0.047*
	(0.023)	(0.072)	(0.030)	(0.013)	(0.022)
Mean	0.834	0.027	0.696	0.036	0.414
Proportion treated	0.009	0.009	0.009	0.009	0.009
N	227335	227335	160682	190482	189247
Pilot Group 2					
Treatment Effect (ATT)	0.037**	0.012	-0.000	0.002	-0.014
	(0.017)	(0.047)	(0.020)	(0.010)	(0.020)
Mean	0.831	0.021	0.691	0.037	0.409
Proportion treated	0.031	0.031	0.030	0.030	0.030
N	232465	232465	164168	194630	193373

Notes: SAIL Databank data. Sample comprises pupils who can be observed from age six, attend mainstream schools in Wales and have no missing data. Free school meal (FSM) eligibility observed at the year of measurement. Estimates from equation (2.1) using the estimator proposed by Borusyak et al. (2024), using the full sample. Controls (focal characteristic excluded): gender, ethnicity, language, Welsh Index of Multiple Deprivation at age six, free school meal eligibility at the age of measurement, free school meal eligibility at the age of measurement interacted with Pupil Deprivation Grant funding, Primary school fixed effects, year fixed effects. Standard errors in parentheses, clustered at the school level. * Significant at 10%; ** significant at 5%; *** significant at 1%.

peers who instead received formal schooling between ages five and seven, while the point estimates for the same outcomes for girls are positive and negative respectively.

I next explore how the effects of the Foundation Phase Pilot differ according to the socioeconomic status of the pupil. Many existing evaluations of early childhood education and care provision find more positive effects of preschool attendance between ages three to six for more economically disadvantaged families, most likely due to income-based differences in the nature of the counterfactual mode of care (Duncan et al., 2023). However there remains a lack of evidence about whether, in cases of universal provision, a formal or a play-based pedagogy in the early years is likely to benefit disadvantaged pupils more. On the one hand, it is possible that the Foundation Phase's child-centered approach enables disadvantaged pupils to strengthen foundational skills such as self-regulation and oracy, which in turn enhances their capacity to benefit from more formal instruction in later years. On the other hand, disadvantaged pupils may respond positively to the structured emphasis on literacy and numeracy within the (counterfactual) formal curriculum, potentially helping to mitigate skill gaps already present at the start of schooling.

Similarly to previous evaluations, in this research I use eligibility for free school meals (FSM) as a proxy for socioeconomic status. Panel B in Table 2.4 shows separate results for FSM and non-FSM pupils at the age of the outcome measurement, calculated using the full sample. Again, the results are largely insignificant for both groups. One exception, though only marginally significant at the 10% level, is in column 1. This indicates that pupils attending Pilot schools at age six who were eligible for free school meals at age 11 were 3.4 percentage points more likely to achieve the Core Subject Indicator at age 11 compared to their FSM-eligible peers who instead received formal education between ages five and seven. In addition, this more positive effect in academic attainment for FSM-eligible pupils is largely retained, though not statistically significantly so, in the age 16 outcomes; here, the point estimates in columns 3 and 5 suggest that FSM-eligible pupils who attended Foundation Phase Pilot schools at are on average *more* likely to achieve five passes in their school leaving examinations (GCSEs) and *more* likely to enrol in the academic track for post-16 education (sixth form) than their FSM-eligible peers who instead received formal schooling between ages five and seven, while the point estimates for the same outcomes for non-FSM-eligible pupils are negative. As previous evaluations of the Foundation Phase found no effect on academic attainment at age 11 by FSM eligibility (Davies et al., 2013), it could be that the broadly favourable effect(s) found here are driven by the Pilot Group 2 schools which were not included in the previous evaluation due to the early timing of its completion.[39] Separate analyses by Pilot group are the focus of the next section.

Main results by Pilot group

Finally I present, in Panel C of Table 2.4, separate results by Pilot group. As previously stated, care should be taken when interpreting these results as the number of clusters of schools included for each estimate is considerably reduced meaning that the estimations suffer from limited statistical power. Overall, in no case is a statistically significant positive (negative) effect off-set by a statistically significant negative (positive) effect across the Pilot groups. However the results do indicate some variation in effects. Pupils attending schools in Pilot Group 1a are found to have 0.082 standard deviations lower attendance at age 11 compared to their peers who received formal schooling between ages five and seven, and 5 percentage points lower likelihood of achieving five GCSE passes, which is a very large effect. Meanwhile pupils attending schools in Pilot Group 2 are found to have a 3.7 percentage point increased likelihood of achieving the Core Subject Indicator at age 11, and those in Pilot Group 1b a 4.7 percentage point lower likelihood of attending sixth form. Collating information from both significant and insignificant point estimates, the results provide suggestive evidence that Pilot Group 1a had more negative effects across most outcomes and Pilot Groups 1b and 2 more positive effects, especially at age 11. It is possible that this is due to the Pilot Groups 1b and 2 benefitting from greater and improved training compared to Pilot Group 1.[40]

2.6 Discussion and conclusion

This paper examines the Foundation Phase Pilot in Wales to explore the effect of extended play-based learning for children aged five to seven on children's later cognitive and noncognitive outcomes.

[39] In their previous evaluation, Davies et al. (2015) also found evidence of negative effects of the Foundation Phase Pilot on persistent absenteeism for more deprived groups, when using all Pilot groups in the evaluation. This finding is echoed in column 2 of Panel B. While again statistically insignificant, here the point estimate suggests that pupils attending Pilot schools at age six who were eligible for free school meals (FSM) at age 11 have lower standardised attendance rates at age 11 than their FSM-eligible peers who instead received more formal education.

[40] While Davies et al. (2013) do find tentative positive effects on age 11 attainment among Group 1, this is likely to be as a result of combining what in this study are referred to as Group 1a and Group 1b.

In doing so, this paper responds to a growing need for evidence about how different curricula and pedagogies in the early years impact children's development in the long-term. It also responds to calls to evaluate the long-run effects of this particular policy in Wales, which previously was only evaluated up to age 11 and which was later rolled out nation-wide (Power et al., 2019). It is worth repeating that this paper is only able to explore outcomes available in existing administrative datasets, which may not represent the full set of outcomes that educational practitioners and policy-makers are interested in when implementing policies of this kind. In addition, as the Foundation Phase schools were not randomly selected, the results in this paper may not be the same as the results that we might expect for the universe of schools in Wales. Moreover, as the policy which I evaluate was only the Pilot stage during the years evaluated, the provision may not have been similar to, or of the same quality as, when the national roll-out commenced.

Based on education and health administrative data for the universe of pupils in Wales born between 1998 and 2004, I find that the Foundation Phase Pilot had no measurable effects on pupils' school-related outcomes and behaviours at ages 11 and 16. This finding is in line with the null effects found by other studies evaluating the long-term effects of a play-based and child-centered curriculum and pedagogy for early childhood education when compared to other (more formal) pedagogical approaches (Biroli et al., 2018). The null effects also constitute a valuable contribution as they indicate that, on average, pupils who received two fewer years of direct instruction between ages five and seven still had similar outcomes to their peers in later academic attainment and progression, as well as in behavioural outcomes such as school attendance and exclusions.

There are a number of possible reasons that I find no measurable effects of the Foundation Phase Pilot at ages 11 or 16. First, it is possible that measurable short-term impacts did occur but had already 'faded out' by the time the children reached age 11. According to the existing evidence, this seems particularly likely if those measurable impacts are cognitive; Li et al. (2020), for example, find that cognitive effect sizes of targeted preschool programs in the US drop roughly in half in one to two years. As an earlier evaluation of the Foundation Phase found tentative evidence of negative contemporaneous effects on pupil attainment (Taylor et al., 2015a) – albeit using a very small sample of pupils from Pilot Group 1 – it is possible that Foundation Phase pupils 'caught up' by the time they reached age 11 and had experienced four additional years of formal schooling. If so, we would conclude that these cognitive skills are not time-sensitive at this age (Dietrichson et al., 2020), and that children may be best served by developing other skills during these crucial years. Relatedly, it is possible, then, that the Foundation Phase Pilot pupils developed other, time-sensitive skills during these years which are not possible for me to evaluate using the administrative data currently available, such as positive attitudes towards learning, socioemotional skills or creativity. Indeed, in general, effects of early education on noncognitive skills such as these are found to be more persistent than effects on test scores (Chetty et al., 2011; Cornelissen and Dustmann, 2019; Heckman et al., 2010).

It is also possible that measurable effects of the Foundation Phase Pilot were obscured or counteracted by factors related to its implementation or behavioural responses by parents. One way that this could have happened is if the Foundation Phase *would* have had positive or negative effects if implemented accurately, but these impacts were biased towards zero because the actual provision tended towards the pre-exiting norm, in this case formal schooling. As highlighted by Conti and Gupta (2024), rushed implementation of early childhood education provision can result in negative effects of an otherwise successful policy due to the low quality of facilities or staff training. The results presented here may tentatively support a similar narrative, as effects are broadly more positive for Pilot Group 2 who had training materials not available for Pilot Group 1 (Siraj-Blatchford et al., 2005).[41] Another way that the effects of the Foundation Phase Pilot may have been obscured or counteracted by factors related to its implementation is if the policy actually had negative effects, but these are compensated by the additional resources and increased pupil-teacher ratios that accompanied the change in curriculum and pedagogy. However it is worth noting that evidence on the effects of class size reductions on student achievement is mixed (Hanushek, 2020). Finally, parental investments during early childhood have been found to impact children's cognitive (Macmillan and Tominey, 2023) and non-cognitive (Moroni et al., 2019) skill accumulation, and parents have been found to change their investments in response to the perceived quality of schooling provision (Gelber and Isen, 2013; Gonzalez, 2020; Greaves et al., 2023). It is therefore possible that the parents of Pilot school pupils may have provided supplementary academic instruction at home as compensation for any perceived deficiencies arising from a less formal curriculum, and therein reduced or bolstered the size of any (test score) effects.

[41] Indeed, evaluating the full roll-out of the Foundation Phase, Waldron et al. (2014) observe variation in the degree of implementation of the Foundation Phase in schools and classrooms, but not differ according to any observable characteristics of the schools.

As more countries look to expand high-quality childcare and invest in early years education, it is important to identify the skills and capacities that are time-sensitive in the early years of life. Further research is needed regarding what these skills are, the pedagogical approaches in early childhood through which they are best developed, and in what contexts these pedagogical approaches are most effective. Regarding the latter two questions, the evidence in this paper provides no strong case for *or against* either play-based or formal education for children between ages five and seven. If policy-makers are willing to assume that a more play-based pedagogy better develops time-sensitive skills that will serve children in their later lives compared to more formal schooling, then the evidence provided here suggests that pupils may develop both these skills and, by age 11, reach the same level of measurable, school-related outcomes that they would had they received more formal schooling. Nonetheless, the additional resources and training required to ensure that a more play-based pedagogy remains of the highest quality should also be bourne in mind when considering such policy decisions.

3 Sibling Spillover Effects in Non-Cognitive Skills: Evidence from Conduct and Attendance at School

3.1 Introduction

Many studies highlight the importance of non-cognitive skills – such as personality traits and emotional regulation, for example – both for the development of cognitive skills and as a direct influence on later life outcomes including marriage status, earnings and criminal activity (Heckman and Kautz, 2012; Heckman et al., 2006). Nonetheless, our understanding of how factors such as genetic background, family inputs and school environments combine to shape non-cognitive skills over the lifecourse remains limited (Deming, 2022). Existing research indicates that the family environment plays a particularly significant role in the development of non-cognitive skills (Borghans et al., 2008; Cunha and Heckman, 2008), and that these skills retain their malleability for longer than cognitive skills and likely into adolescence (Almlund et al., 2011; Cunha and Heckman, 2007; Kautz and Zanoni, 2024). Therefore, as children age and gain more independence from parental influence, we may expect sibling relationships to be an increasingly important determinant of non-cognitive outcomes. This might occur, for example, if an older sibling serves as role model for particular behaviours, shares information, or directly teaches the younger sibling (Nicoletti and Rabe, 2019). Yet, causally identifying the role of sibling relationships for young people's non-cognitive skill development is challenging and evidence consequently scarce.

This paper explores sibling spillover effects in non-cognitive skills. In particular, we examine whether an older sibling's school misbehaviour in grades 8-10 (at ages 12-15) causes an increase in the younger sibling's likelihood to misbehave in the same grade. We measure misbehaviour using temporary exclusions (suspensions) and unauthorised absences as recorded in administrative data covering all state school pupils in England between 2007 and 2019. In England, absences and exclusions are recorded by law, and we are able to distinguish across different types of absences (authorised and unauthorised) and reasons for exclusions, allowing us to investigate the types of behavioural issues that are transmitted between siblings.

There is considerable diversity in how non-cognitive skills are defined and measured in the literature. Non-cognitive skills are understood to include factors such as personality traits, interpersonal skills, motivations, preferences and emotional regulation (Almlund et al., 2011; Heckman and Kautz, 2012; Papageorge et al., 2019), all of which are commonly measured using survey data. Shortcomings of such survey-based measures include reference bias from the respondent,[1] use of taxonomies of task performance without standardising for incentives or adjusting for levels in other domains (Borghans et al., 2008; Heckman and Kautz, 2012), and small and/or selected samples of the population. We avoid these problems inherent in survey-based measures of non-cognitive skills by following a recent literature which instead uses observable behaviours as skill measures (Jackson, 2018; Kautz and Zanoni, 2024; Petek and Pope, 2023). Observable behaviours such as school attendance and on-time graduation have been shown to be good predictors of earnings, college completion and job stability in later life (Heckman and Rubinstein, 2001). We use two observable behaviours available in administrative data: temporary exclusions (suspensions) and unauthorised absences from school.

To identify a *causal* sibling spillover effect in our observable behaviours, we must address three main identification issues emphasized in the literature on peer effects (Manski, 1993; Moffitt, 2001): reflection (simultaneity), endogenous peer group membership, and correlated omitted variables. Reflection is less of a concern in our setting because we focus on same-grade spillovers where the older sibling's behaviour is observed prior to the younger sibling's, ruling out reverse causality. Likewise, endogenous peer group membership is minimal since siblings cannot choose one another. The more pressing threat to identification are correlated omitted variables, whereby unobserved family

[1] This can be through self-reporting (Heckman and Rubinstein, 2001), reporting on peers (Heckman and Kautz, 2012), or reporting on one's child (Del Bono et al., 2024).

or environmental factors could influence both siblings and thus generate spurious correlations between their behaviours.

We clean the same-grade sibling associations in temporary exclusions and unauthorised absences of correlated factors by controlling for a number of individual and sibship characteristics, as well as year and grade fixed effects. These controls help to account for observable differences across individuals and families, and unobservable cohort differences. However, any remaining association between siblings' behaviours may still reflect shared unobserved characteristics – such as genetic endowments, parenting styles, or neighbourhood effects – rather than a causal effect.

To further account for the potential endogeneity of correlated omitted variables, we then instrument the older sibling's misbehaviour using the average misbehaviour of their school peers. We focus on peers who share the same school, grade, year, gender and prior attainment tercile as the older sibling. This focuses the definition of the peer group on pupils with whom the sibling shares the greatest number of observable characteristics (homogeneous peers), which captures those with whom they are most likely to interact with regularly and be grouped with in England's typically attainment-based classes. However, the validity of the instrument may not hold if there are school-level unobservable characteristics that are common between both siblings going to the same school. We therefore combine the instrumental variable estimation with an estimation akin to a fixed effects strategy using the misbehaviour of similar pupils (the homogenous peers) in the three closest consecutive cohorts. In effect, this washes out the average misbehaviour in the relevant comparison group and exploits, for identification, only idiosyncratic variation in the misbehaviour of the older siblings' own homogenous school peers compared to the homogeneous peers in nearby years. Exploiting quasi-random variation in peer groups within and across school cohorts is similar to the identification strategies used by Lavy et al. (2012) and Nicoletti and Rabe (2019).

Based on administrative data covering all state school pupils in England between 2007 and 2019, we find evidence of spillover effects in exclusions and absences between siblings when they are in the same grade, including in the reason for exclusion. The results from our main empirical specification show that one additional half-day exclusion in an older sibling increases exclusions in the younger sibling by around 0.1 half-days, equivalent to an effect size of 0.04 standard deviations. One additional half-day of absence in an older sibling increases absences in the younger sibling by around 0.4 half-days, equivalent to an effect size of 0.03 standard deviations. Heterogeneity analysis reveals generally larger spillover effects for siblings who are closer in age but relatively comparable effects across different gender pairings. Pupils who are economically disadvantaged (ever eligible for free school meals) also have slightly larger spillover effects than their more advantaged peers, suggesting that we may underestimate the positive impact on inequalities of interventions that successfully improve the non-cognitive skills of disadvantaged pupils who are also older siblings. Though we are unable to fully disentangle the mechanisms of our spillover effects, the fact that we find larger sibling spillover effects for siblings who are excluded for the same reason and for siblings who are closer in age makes role modelling a convincing explanation of our results.

We contribute to the literature on the production of non-cognitive skills. Much of the existing research is focused on modelling the production of non-cognitive skills within dynamic skill formation frameworks (Cunha and Heckman, 2007, 2008; Cunha et al., 2010). In general, this literature finds that, compared to cognitive skills, family and parental inputs are more effective at raising non-cognitive skills, and do so at later ages. More recent studies have used experimental and quasi-experimental designs to identify the causal effects of interventions on non-cognitive skill development (Alan et al., 2021; Attanasio et al., 2020; Heckman et al., 2013; Kautz and Zanoni, 2024; Sorrenti et al., 2025). Syntheses of such evidence consistently conclude that, while non-cognitive skills are somewhat stable over the lifecourse, they can also be improved through targeted programs (Almlund et al., 2011; Borghans et al., 2008; Kautz et al., 2014). However, the vast majority of empirical research on non-cognitive skill formation has focused on school-based inputs and interventions, most likely reflecting both the accessibility of data and the policy relevance of the findings (Heckman et al., 2006; Jackson, 2018; Petek and Pope, 2023). By comparison, family inputs remain less studied, and especially at ages beyond the early years (Kautz et al., 2014).[2] We contribute to this literature by providing evidence of one mechanism by which family inputs contribute to the formation of non-cognitive skills, namely through spillovers between siblings.

We also contribute to a growing literature on sibling spillover effects. A number of studies have documented positive spillover effects from older-to-younger siblings in test scores (Landersø et al.,

[2] One exception is Moroni et al. (2019), who examine the role of parental inputs in the socio-emotional skill formation for children aged six to 11. They find evidence of complementarity between parental inputs and early socio-emotional skills in enriched environments, and substitutability in stressful environments.

2020; Nicoletti and Rabe, 2019), as well as spillover effects in education- and career-related choices such as school choice (Dustan, 2018), high-school track choice (Joensen and Nielsen, 2018), course choice (Gurantz et al., 2020), college choice (Altmejd et al., 2021), decisions to serve in the military (Bingley et al., 2021), for fathers to take parental leave (Dahl et al., 2014), and for mothers' labour supply (Nicoletti et al., 2018). However there is comparatively little existing research on sibling spillover effects in non-cognitive skills and behaviours. One exception is Altonji et al. (2017), who find that smoking, drinking, and cannabis use by an older sibling modestly increases the probability that a younger sibling engages in the same type of behaviour. Another exception is Anand and Kahn (2023), who find that individuals who have experienced a sibling or friend's teen pregnancy have a decreased likelihood of having unprotected sex and have fewer sexual partners. We contribute to this literature by investigating sibling spillover effects in a two previously unexamined outcomes: temporary exclusions and unauthorised absences.

Finally, we also contribute to the literature examining spillover effects in school-related behaviours among peer groups, much of which is based on the US context. One strand of this literature exploits individuals' exogenous exposure to badly behaved or 'crime-prone' schoolmates through school allocation or re-location procedures, identifying negative spillover effects in disciplinary problems, absenteeism, and criminal involvement (Deming, 2011; Imberman et al., 2012). A second strand uses the characteristics of peer groups as an instrument for their behaviour. For example Carrell and Hoekstra (2010) find that children exposed to domestic violence decrease their peers' test scores and increase misbehavior, while Gaviria and Raphael (2001) find strong evidence of peer effects in drug use, alcohol drinking, cigarette smoking, church going, and dropping out of school, using variation in peers' personal and family background characteristics. However a few recent papers have found novel ways to identify behavioural spillovers more directly; Bennett and Bergman (2021) use fine-grained attendance data to identify peer networks based on pupils who systematically miss classes together and find evidence of significant attendance spillovers along these networks using a randomised behavioural intervention, while Lenard and Silliman (2025) leverage idiosyncratic changes in school bus peer groups in the US and find larger spillover effects in behaviours (absences, lateness, and short-term suspensions) compared to academic outcomes. We add to this literature by implementing a different methodology which exploits deviations in the behaviour of the peer group with whom the older sibling is most likely to interact.

Our second motivation is to inform the design and evaluation of education and social policies. If an older sibling's misbehaviour causally influences that of their younger siblings, this may partly explain persistent inequalities in both behaviour-related sanctions and educational outcomes between social groups. Our results indicate that the misbehaviour of an older sibling propagates through their family network such that the challenges faced disproportionately by disadvataged pupils can have knock-on effects among their younger siblings. At the same time, our results also indicate that the effects of policies designed to reduce misbehaviours and improve non-cognitive skills can be compounding. Mentoring, social and emotional learning, and other school-based interventions which are effective in tackling school misbehaviour and nonattendance are likely to have larger effects than those typically measured in studies that only focus on treated pupils because such policies are likely to indirectly benefit younger siblings too. Parental or public investments into the non-cognitive skills of older siblings can therefore have multiplier effects, and these externalities should be factored in to estimations of the aggregate benefits of such policies and interventions.

The rest of this paper is structured as follows. Section 3.2 provides an overview of schooling in England, including trends in school temporary exclusions and unauthorised absences. Section 3.3 then describes the data used and provides descriptive evidence about sibling correlations in behaviours. Section 3.4 describes the empirical strategy. Results are found in Section 3.5, including robustness checks and heterogeneous effects. Section 3.6 concludes.

3.2 Education, school discipline and school attendance in England

3.2.1 The English education system

In England, full-time education is compulsory for all children between ages five and 16. There is no grade retention, and in the vast majority of cases pupils attend two schools: Primary school from ages five to 11, and Secondary school from ages 11 to 16.

School discipline

Schools in England have autonomy to set their own behavior policies within government guidelines. Behaviour policies are typically strict and emphasise discipline, punctuality, and uniform compliance. They are enforced through a range of disciplinary measures including behavior point systems, detentions, isolation rooms during lesson time, temporary exclusions (suspensions) and permanent exclusions. This paper focuses on exclusions, as they are one of the few sanctions that schools are legally required to report in an annual school census.[3] Other disciplinary measures are not observed in administrative data.

When deciding whether to temporarily exclude a pupil, schools are likely to consider trade-offs: the excluded pupil will learn less during their exclusion, but their peers may learn more in their absence (Pope and Zuo, 2023). Exclusions may also serve as a deterrent, or be motivated by safety concerns. Only the head teacher (principal) of a school can exclude a pupil and this must be on disciplinary grounds, with permitted reasons including physical or verbal assault against a pupil or adult, damage, theft, and bullying (Department for Education, 2022, and see Appendix Table C.1 for a complete list). Schools are permitted to exclude pupils either permanently or for a temporary period, which can be up to a maximum of 45 school days in a single academic year.[4] Permanent exclusions are rare, ever affecting around 0.05% of pupils in mainstream schools. This paper therefore focuses on temporary exclusions, which are experienced by around 11% of pupils in mainstream schools at least once during Secondary school.[5]

[3] Although exclusions are quite an extreme measure of behaviour at school, compared to less extreme measures they are also less likely to be subject to measurement error (schools are obliged by law to report exclusions), or to variations in the behaviour management strategies of particular classroom teachers (instead, head teachers (principals) decide whether to exclude a pupil according to pre-determined guidance).

[4] In cases where a temporary exclusion would bring the pupil's total number of school sessions out of school to more than 30 (15 days) in an academic term – or between 10 (5 days) and 30 (15 days) if parents make representations – then a school's governing body is obliged consider whether the suspended pupil should be reinstated.

[5] Non-mainstream schools, which constitute 1.42% of our data, include special schools, hospital schools, and pupil referral units. Additionally we do not include City Technology Colleges in this figure as these are not required to report exclusions under the same guidance as other, mainstream schools. These figures are for pupil cohorts completing their GCSE exams in 2017, 2018 and 2019.

School attendance

The academic year in England runs from the beginning of September to mid-July. Each day is divided into two sessions – one in the morning and one in the afternoon – and schools must open for 380 sessions (190 days) in each academic year.

Attendance registers are taken twice a day, at the start of the morning and afternoon sessions. Where a pupil is absent, the school must record whether their absence is authorised or unauthorised. Authorised absences include those for reasons including illness, medical appointments, or observing religious festivals. Unauthorised absences include all unexplained or unjustified absences, truancy, and late arrival after registration has closed (usually 30 minutes after the start of the session). These absences can lead to lower-level sanctions such as detentions. Schools also routinely follow-up with families to establish the reason for an absence, and may require parents to attend support services, to pay a fine, or to face court proceedings.[6] Approximately 75% of pupils receive an unauthorised absence at some point during their Secondary schooling.[7]

3.2.2 Exclusions and absences in England

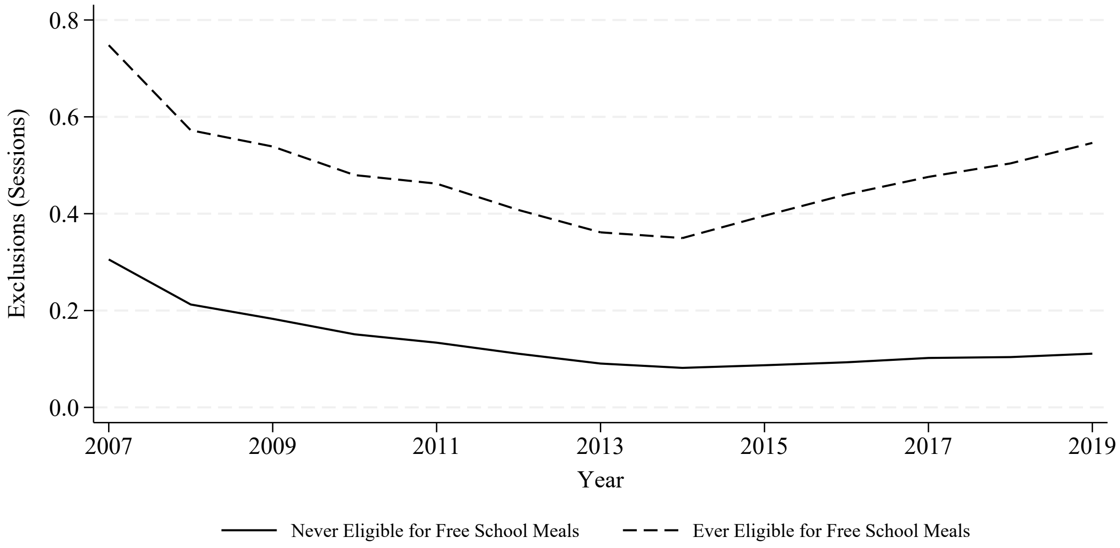
In England, there are pronounced disparities in exclusions and absences across pupil groups (see Appendix Table C.2). On average, pupils miss 0.2 sessions per year due to temporary exclusions and 3 sessions due to unauthorised absences. Boys experience nearly three times as many sessions due to exclusions as girls, though rates of absence are similar. Pupils with Special Educational Needs, those from Black and Gypsy, Roma and Traveller backgrounds, and those from more deprived households miss substantially more sessions than their peers in both measures. These disparities are concerning as they are likely to exacerbate existing educational inequalities and may undermine the effectiveness of broader investments in education.

Figures 3.1 and 3.2 show the trends in the mean number of exclusion and absence sessions in England between 2007 and 2019 separately for pupils who are ever and never eligible for free school

[6] Fines are between £80 and £160. If a pupil is absent three or more times within three years then their parents or carers may be taken to court.

[7] This figure is for pupil cohorts completing their GCSE exams in 2017, 2018 and 2019.

Figure 3.1 Line graph showing mean exclusion sessions by year and eligibility for free school meals, 2007 - 2019



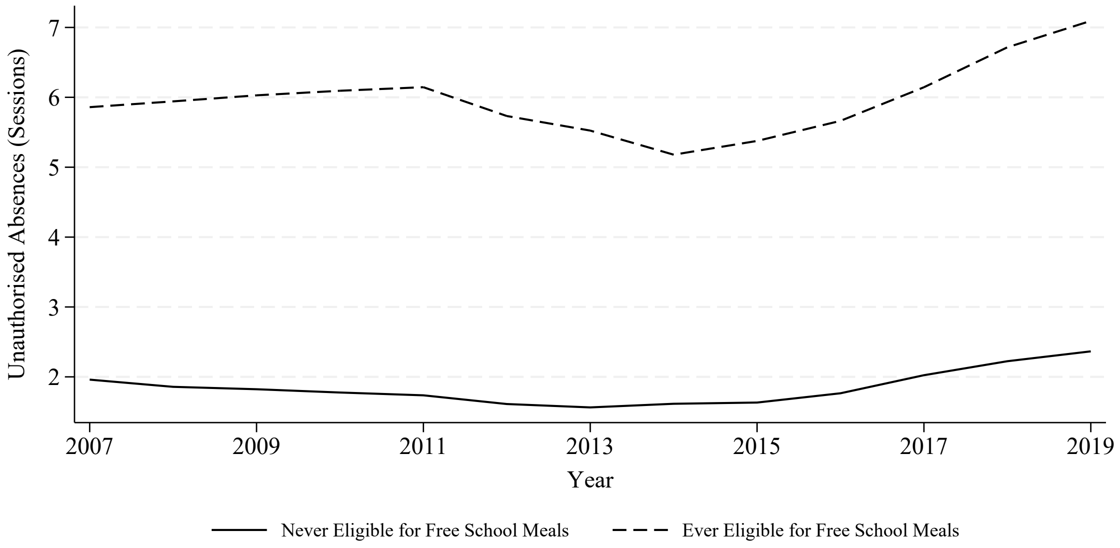
Notes: National Pupil Database, 2007 - 2019. Unrestricted sample, all grades. Temporary exclusions only. Sample size: 73,952,708.

meals, a commonly-used proxy for economic disadvantage.[8] Figure 3.1 shows that exclusion sessions have been increasing since 2014, following a period of decline, and that the rate of increase has been higher among more economically disadvantaged pupils. Figure 3.2 shows that the same is true of absences.

Finally, exclusions and absences occur more frequently in some school grades compared to others. Figure 3.3 shows the mean exclusion and absence sessions, based on data for 2007-2019, separately by school grade. (Note that there are separate y-axes for each outcome, and that mean sessions for absences are around ten times higher than for exclusions.) Mean exclusion sessions (left y-axis) increase substantially from Grade 7, when pupils begin Secondary schooling, and peak in Grade 10 when pupils are aged 14. Mean absence sessions (right y-axis) decrease slightly through primary schooling until Grade 6, and increase rapidly thereafter. In this paper we focus on exclusions and absences in grades 8, 9 and 10, the grades when exclusions are most prevalent. Grade 11 is the grade in which pupils sit high-stakes national exams (GCSEs), and when factors such as leaving

[8] Throughout this paper, we use a year to refer to the year of the January of the academic year. For example, 2007 refers to the academic year 2006/07.

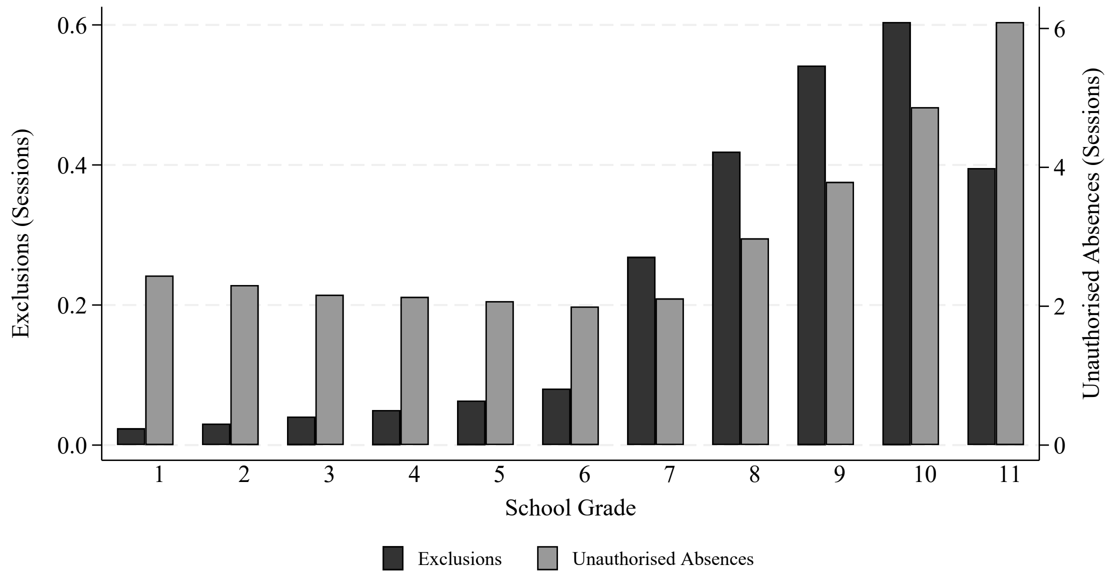
Figure 3.2 Line graph showing mean absence sessions by year and eligibility for free school meals, 2007 - 2019



Notes: National Pupil Database, 2007 - 2019. Unrestricted sample, all grades. Unauthorised absences only. Sample size: 73,952,708.

school following the exams and being granted discretionary study leave make measurement of school absences (and, to a lesser extent, exclusions) less reliable.

Figure 3.3 Exclusion and absence sessions by grade



Notes: National Pupil Database, 2007 - 2019 (combined). Exclusions are temporary exclusions only. Absences are unauthorised absences only. Unrestricted sample. Sample size: 73,952,708.

3.3 Data and descriptive statistics

3.3.1 Data

For this analysis we use the National Pupil Database (NPD), an administrative dataset collected by the Department for Education in England. All government-funded schools in England are required to submit these data by law, meaning that the data covers 93% of pupils in England and is highly accurate and complete.[9] Data are reported for every pupil for each year that they are educated in schools, and data tables can be linked to each other and across the years using an anonymous pupil identifier. We focus on the years 2007 to 2019, which are the years for which consistent data on exclusions and attendance are available.

Outcomes and observed background

We use NPD data to construct our main dependent and independent variables on exclusions and absences. For exclusions, we focus on temporary exclusions (suspensions), which account for the vast majority of incidents (97.78%).[10] Each exclusion includes a school-submitted reason, which we categorise into physical, verbal, disruptive behaviour, illicit, and other (see Appendix Table C.1 for a more detailed overview of these categorisations). In this analysis, we focus on the intensive margin, and therefore collate the number of sessions for which a pupil is temporarily excluded during the academic year of focus. For absences, we follow a similar approach, and collate the number of unauthorised absences a pupil receives per academic year. Using unauthorised absences ensures that absences for reasons such as illness and medical appointments are not included in the measure.[11]

The NPD also provides rich information on pupil and family background. We use indicators for gender, ethnicity, month and year of birth, whether the pupil speaks English as their first language

[9] 7% of pupils in England are educated in private schools, which are not government-funded and not required to submit data for inclusion in the National Pupil Database.

[10] Permanent exclusions are very rare (ever affecting around 0.05% of pupils) and therefore too extreme a margin of behaviour to be relevant population-wide. We also retain lunchtime exclusions, which mean that a pupil is forbidden to mix with their peers during unstructured lunchtime breaks. However these are relatively rare, accounting for 0.43% of the total exclusion incidents that we observe. Permanent exclusions account for 1.79%, and temporary exclusions 97.78%.

[11] There may be some measurement error in our unauthorised absences variable as some absences can be assigned ‘unauthorised’ when no reason is yet assigned, however the proportion is relatively small: in 2010, of the 17% of absences which were unauthorised, just 3% were due to having no reason yet assigned (Department for Education, 2011).

(EAL), and school type.[12] Economic disadvantage is proxied by whether a pupil is ever eligible for free school meals (FSM) during their years of schooling between 2006 and 2019, and special educational needs (SEN) status is defined similarly. In robustness checks, we also incorporate the Income Deprivation Affecting Children Index (IDACI), a neighbourhood-level measure of resource deprivation.

Sibling definition

The NPD contains a family group identifier, which identifies pupils in government-funded schools between ages four and 16 and who are living together at the same address. Step- and half-siblings are therefore identified if they live at the same address, and are indistinguishable from biological siblings (Nicoletti and Rabe, 2013, 2019). The family group identifier is available for just three years: 2008, 2013 and 2016. We use a ‘nearest year’ method to identify siblings in every year of our panel.[13]

Peer definition

As we will discuss in the next section, our identification strategy relies on idiosyncratic variation in the behaviour of a siblings’ homogenous (similar) peers, with whom they are most likely to interact. Pupil groupings in English secondary schools tend to be highly structured; pupils attend lessons only with those in the same school and grade, and are often grouped by prior academic attainment, particularly in earlier Secondary years. These groupings, as well as an inclination to associate with peers of the same gender, tend to determine a pupil’s friendship group. Indeed, in a survey of over 3,000 adolescents in England, Burgess et al. (2011) find that 82% of stated friends attended the same school, 78% were of the same gender, and that friends are strongly similar in academic performance and behaviour, though less so in socio-economic status.[14] In line with this evidence in support of the assumption

[12] Local authority schools include all schools funded by the local authority. Academies are funded centrally and have more freedoms compared to local authority schools, and are similar to charter schools in the US. We observe three types of academy: sponsored academies, which are usually previously low-performing schools; converter academies, which have opted to convert to academy status; and free schools, which are new academy schools set up by parents and local groups.

[13] The ‘nearest year’ method entails using each group identifier for years following the identifier year (not leading up to), to ensure that older siblings who are older than 16, and therefore may no longer be in the NPD in the identifier year, are correctly identified. Specifically, we use the 2008 identifier for cohorts who are in Grade 8 between 2010 and 2012, the 2013 identifier for those who are in Grade 8 between 2013 and 2016, and the 2016 identifier for those who are in Grade 8 in 2017 and 2018.

[14] Similarly, Bennett and Bergman (2021), using US data, finds strong homophily in truancy networks, particularly by gender and academic achievement.

of homophily in adolescent networks, we define a sibling's peer group as those in the same school, grade, gender, and prior attainment group as the sibling. We delineate prior attainment groups based on national test scores at age 11, with pupils assigned to within-school attainment terciles.[15] We then construct peer behaviour measures as the take-one-out mean number of sessions missed due to temporary exclusions or unauthorised absences for each homogeneous peer group in the academic year of interest.

Sample restrictions

The initial sample for our analysis includes all siblings attending mainstream schools in England between 2007 and 2019.[16] We restrict the sample to pupils for whom we observe age 11 (Grade 6) test scores, time-invariant demographics, and both absences and exclusions outcomes in at least one of grades 8, 9 or 10 in academic years 2010 to 2018.[17] As well as removing from the data all twins and siblings attending the same academic year, we remove all siblings who are just one academic year apart to ensure that three-year rolling average peer behaviour measures (which we will explain in Section 3.4) do not include the other sibling's behaviours.[18] To avoid any multiplier spillover effects we also retain only the eldest sibling pair per family group identifier and year. The resultant, initial sample comprises 987,021 sibling pairs.

As our identification relies on a measure of change in peer group behavioural outcomes, we next follow Lavy et al. (2012) and drop sibling pairs for whom either sibling attends a school that has a year-on-year change of entry cohort size of more than 75% or enrollments below 30 pupils (3.88%).[19][20] We also restrict our sample to sibling pairs for whom we observe the three-year school-level rolling

[15] Especially in earlier grades of Secondary school, pupils are grouped according to their test scores at the end of Primary school, at age 11. We therefore also use age 11 test scores to determine the prior attainment terciles.

[16] We drop pupils attending non-mainstream schools, which constitute 1.42% of our data. The non-mainstream schools are special schools, hospital schools, pupil referral units, and City Technology Colleges as these are not required to report exclusions under the same guidance as other, mainstream schools.

[17] Requiring Grade 6 test scores leaves us with 11 cohorts, with outcomes observed between 2009 and 2019. Then, as we will explain in Section 3.4, a three-year rolling average peer behaviour measure is a crucial control variable in our empirical strategy. As such, we do not include the cohorts at either end of this distribution as our focal pupils.

[18] As we are interested in sibling outcomes when siblings are in *the same grade*, this means that we observe older siblings in the academic years 2010 to 2016, and younger siblings in the academic years 2012 to 2018.

[19] We note that Lavy et al. (2012) drop schools with enrollments below 15 pupils. However, as our peer measures are within age 11 test score terciles and gender, we set a larger minimum cohort measure. 30 pupils is the approximate capacity of one classroom in England.

[20] Table 3.6 in Section 3.5.2 results without these cohort restrictions and finds that they differ little from the main results that we will present below.

averages for peers in the same school cohort, age 11 attainment group, and of the same gender. This is a more stringent restriction as it requires non-missing school-by-attainment-group-by-gender measures of behaviour within the same school for three consecutive years, and it is intermittent in our data for two reasons. First, during the observed years many Secondary schools converted to a more autonomous school type (Academy), and in our data it is not possible to follow these schools across, or in the years directly around, that conversion.[21] Second, in 2010 a quarter of Primary schools boycotted the age 11 attainment tests due to ideological differences regarding the nature of curriculum and assessment. These missing test scores exclude a non-incidental number of pupils from the sample, and also affect whether rolling averages can be constructed for the cohorts either side. For these reasons, implementing this second restriction results is a reduction of the sample by 32.5% of sibling pairs. The final sample comprises 1,142,734 observations by sibling pair and grade, including 627,959 unique sibling pairs and 3,498 schools. Observation counts by year, grade and sibling type are provided in Appendix Table C.3.

3.3.2 Descriptive statistics

Summary statistics

Table 3.1 contains summary statistics for the final sample. On average, 9.3% of younger siblings and 7.9% of older siblings experience least one temporary exclusion in grades 8, 9 or 10, while just below half of pupils are ever absent without authorisation during these grades.[22] Reasons for exclusion are relatively balanced in proportion, with the most common reasons being verbal and physical abuse. Individual characteristics are very balanced across younger and older siblings. This is as we would expect given that siblings are likely to be largely biological and, by definition, living in the same household, and gives us confidence that the sibling identifier is accurate. Older siblings have higher prior attainment compared to younger siblings, which aligns with well-documented birth order effects (e.g., De Haan, 2010). Regarding sibling pair characteristics, age gaps of two, three and four or more

[21] However, not observing these schools in the years directly around a conversion does have the benefit of ensuring more consistency in the school's use of behavioural sanctions; some school conversions are associated with considerably higher rates of exclusions (Machin et al., 2020).

[22] Appendix Table C.4 additionally reports the mean values of the sessions excluded and sessions absent measures by grade and sibling type, and Appendix Figure C.1 displays the distributions for younger siblings, which are our dependent variables.

Table 3.1 Summary statistics by sibling type

	Younger sibling		Older sibling	
	Mean	Standard deviation	Mean	Standard deviation
pupils' outcomes: extensive (grades 8, 9 & 10)				
Ever excluded	0.093	0.291	0.079	0.269
Ever permanently excluded	0.002	0.047	0.001	0.035
Ever absent	0.499	0.500	0.473	0.499
Ever persistently absent	0.193	0.395	0.191	0.393
Ever excluded reason: Persistent disruptive behaviour	0.029	0.167	0.020	0.140
Ever excluded reason: Physical abuse	0.034	0.182	0.031	0.174
Ever excluded reason: Verbal abuse	0.036	0.187	0.030	0.171
Ever excluded reason: Illicit behaviour	0.015	0.121	0.014	0.116
Ever excluded reason: Other	0.031	0.173	0.024	0.152
pupils' outcomes: intensive (grades 8, 9 & 10)				
Exclusion sessions	0.049	0.170	0.038	0.144
Absence sessions	3.486	11.588	2.515	8.490
Exclusion reason: Persistent disruptive behaviour	0.117	1.274	0.068	0.838
Exclusion reason: Physical abuse	0.091	0.750	0.073	0.624
Exclusion reason: Verbal abuse	0.111	0.971	0.079	0.719
Exclusion reason: Illicit behaviour	0.038	0.441	0.032	0.390
Exclusion reason: Other	0.088	0.851	0.059	0.618
Invididual characteristics (may vary between siblings)				
Female	0.496	0.500	0.495	0.500
Ever eligbile for Free School Meals	0.264	0.441	0.274	0.446
Ever diagnosed with Special Educational Needs	0.333	0.471	0.331	0.471
English as an Additional Language	0.115	0.319	0.126	0.332
Ethnicity: White British or Irish	0.770	0.421	0.770	0.421
Ethnicity: Pakistani or Bangladeshi	0.060	0.237	0.060	0.237
Ethnicity: Indian	0.024	0.154	0.024	0.154
Ethnicity: Other Asian or Mixed White and Asian	0.024	0.151	0.023	0.151
Ethnicity: Black African or Mixed White and Black African	0.034	0.181	0.034	0.181
Ethnicity: Black Caribbean or Mixed White and Black Caribbean	0.021	0.143	0.021	0.143
Ethnicity: Other Black	0.004	0.067	0.005	0.067
Ethnicity: Chinese	0.003	0.052	0.003	0.052
Ethnicity: Gypsy, Roma and Traveller	0.001	0.036	0.001	0.034
Ethnicity: Other Ethnic Group	0.049	0.216	0.051	0.220
Ethnicity: Unknown	0.010	0.101	0.008	0.091
Prior attainment: Low	0.321	0.467	0.292	0.455
Prior attainment: Medium	0.337	0.473	0.334	0.472
Prior attainment: High	0.342	0.474	0.374	0.484
Prior attainment (standardised)	0.048	0.970	0.114	0.971
Sibling characteristics (same for both siblings)				
Age gap: 2 years	0.350	0.477		
Age gap: 3 years	0.314	0.464		
Age gap: 4+ years	0.335	0.472		
Brothers	0.257	0.437		
Younger brother, older sister	0.247	0.431		
Younger sister, older brother	0.248	0.432		
Sisters	0.247	0.432		
Same prior attainment group	0.455	0.498		
Same prior attainment group and gender	0.161	0.368		
Pair observed for 1 year	0.022	0.148		
Pair observed for 2 years	0.246	0.431		
Pair observed for 3 years	0.731	0.443		
No. of sibling pairs	627,959			
No. of observations including pooled sibling pairs across years	1,142,734			
No. of schools	3,498			

Notes: National Pupil Database, 2007 - 2019. Sample is restricted to the eldest two siblings per family and year, attending a mainstream Secondary school, with no missing covariates, with at least a two year grade gap, and observed in grades 8, 9 or 10. Unless otherwise stated, exclusions are temporary exclusions only. Absences are unauthorised absences only. One session is half a school day.

academic years account for approximately a third of the sample each, and sibling sex combinations are relatively evenly balanced. Finally, while 45.5% of sibling pairs are in the same prior attainment group, only 16.1% share both prior attainment group and gender.

Sibling correlations in outcomes

Table 3.2 reports the sample Pearson correlation coefficients for exclusion and absence sessions in the same grade between siblings, with the exclusion correlations estimated separately by reason in the second panel. Overall there are weak positive correlations between siblings in exclusion and absence sessions when siblings are the in the same grade. The correlation for absences is stronger (see Panel A), most likely reflecting that it is a more common event. Panel B shows the correlations in exclusion sessions by reason for exclusion. In general, there is a stronger association in exclusions *for the same reason* between sibling pairs, than in exclusions for different reasons. These stronger associations could be at least in part due to older siblings role modelling behaviours to younger siblings. The exception is illicit behaviours, which include sexual misconduct, drug and alcohol related events, damage, and theft, and have a low between-sibling correlation. This is interesting as exclusions due to illicit behaviours are relatively rare, and may represent a more extreme form of undesirable choices during adolescence. In the remainder of the paper, we address the important but empirically difficult question of whether any part of these sibling correlations are indeed due to a causal effect of the older sibling's behaviour on the younger.

Table 3.2 Sibling same grade correlations in behaviours

Panel A: Overall					
Sessions excluded	0.096				
Sessions absent	0.291				
Panel B: Sessions excluded by reason					
	Disruptive (younger)	Physical (younger)	Verbal (younger)	Ilicit (younger)	Other (younger)
Disruptive (older)	0.084	0.029	0.036	0.012	0.020
Physical (older)	0.032	0.037	0.035	0.016	0.022
Verbal (older)	0.034	0.031	0.044	0.013	0.024
Illicit (older)	0.018	0.019	0.015	0.008	0.012
Other (older)	0.023	0.022	0.026	0.009	0.035

Notes: National Pupil Database, 2007 - 2019. Sample is restricted to the eldest two siblings per family and year, attending a mainstream Secondary school, with no missing covariates, with at least a two year grade gap, and observed in grades 8, 9 or 10. Exclusions are temporary exclusions only. Absences are unauthorised absences only. One session is half a school day.

3.4 Empirical strategy

3.4.1 Identification of sibling spillover effects

To estimate the older-to-younger sibling spillover effect in temporary exclusions and unauthorised absences during Secondary school, we begin by considering the following equation:

$$Y_{1,igst} = \beta_0 + \beta_1 Y_{2,igs't'} + \beta_2 X_{1,i} + \beta_3 X_{2,i} + \beta_4 P_i + \omega_t + \lambda_g + \varepsilon_{1,igst} \tag{3.1}$$

where $Y_{1,igst}$ is the number of exclusion (or unauthorised absence) sessions of the younger sibling of sibling pair i in grade g in school s in year t, $Y_{2,igs't'}$ is the corresponding number of sessions for the older sibling observed in the same grade, in school s' and time t'. In the following we refer to the exclusion and absence sessions with the more generic term of misbehaviour. $X_{1,i}$ and $X_{2,i}$ are vectors of individual characteristics of the younger and older sibling, respectively, including standardised prior attainment in Grade 6 (age 11), FSM status, and EAL status. P_i is a vector of characteristics of the sibling pair i, including the sex combination and age gap. ω_t and λ_g are year and grade fixed effects (for the younger sibling). $\varepsilon_{1,igst}$ is the idiosyncratic error term.

We estimate the sibling spillover effect on misbehaviour for siblings observed at the same grade ($g =$ 8, 9 or 10 (ages 12-15)), who are at least two school grades apart (($(t' - t) \geq$ 2), and who may or may not attend the same school. Observing siblings in the same grade is relevant for gauging whether a younger sibling can be deemed to be 'on the same track' as an older sibling. Same-grade effects also mean that correlations in siblings' misbehaviours are less likely to be caused by shared family year-specific shocks, as these happen at different ages for the two siblings.[23] It also ensures that the reflection issue is not a major concern as that the older sibling's misbehaviour is observed at least two years before the younger sibling's, so there is no simultaneity.[24] Nonetheless identifying the sibling spillover effect in misbehaviour (β_1) still requires us to address two identification issues, apart from

[23] For example, intra-year household income instability has been found to predict school exclusions in the US (Gennetian et al., 2015).

[24] We also show in sensitivity analysis that there are no reverse effects running from the younger to the older sibling (see Section 3.5.2).

simultaneity, which threaten the identification of peer effects (Manski, 1993; Moffitt, 2001): correlated omitted variables and endogenous peer membership.

The issue of correlated omitted variables is a concern in equation (3.1) because there may remain unobserved characteristics, such as family attributes, that affect both siblings and lead to a spurious correlation. To account for this potential endogeneity we instrument the older sibling's school misbehaviour with the average misbehaviour across the older sibling's homogenous school peers, defined as pupils in the same school, grade, year, prior school attainment group, and gender, as the older sibling (see Section 3.3.1). We exclude the older sibling from the computation of their peers' average misbehaviour by constructing the leave-one-out mean, which we denote by $\overline{Y}_{2,-igs't'}$. This instrument is likely to be relevant because children tend to be influenced by the misbehaviour of peers with whom they spend the most time, such as those of the same gender and same level of ability. We also note that, if the older sibling is influenced by their school peers, then the younger sibling will also be influenced by their own school peers (von Hinke et al., 2019). Therefore, to avoid mis-specification, we also include in our equation the leave-one-out mean misbehaviour for the younger sibling's homogenous peers $\overline{Y}_{1,-igst}$, which we compute in the same way as for the older sibling.

The validity of the instrument $\overline{Y}_{2,-igs't'}$ may be compromised if school-specific unobservables – such as disciplinary policies, school-specific time trends in disciplinary actions, and implicit biases against pupils who are expected to be more misbehaved in class – are shared by both siblings.[25] This is a concern since many siblings attend the same school or sort into schools with similar environments. We address this endogeneity issue by using quasi-random variation across (academic) calendar years, t, in the number of exclusion (unauthorised absence) sessions of the homogeneous school peers. To do this we net out from the dependent variable, $Y_{1,igst}$, the expected number of exclusions (absences) for the peer group with the same shared characteristics but across the nearby years $(t-1)$, t and $(t+1)$. In other words, we include in equation (3.1) the leave-one-out homogeneous peers' mean misbehaviour across three years, $(\frac{1}{3}\sum_{r=t-1}^{t+1}\overline{Y}_{1,-igsr})$. Similarly, we net out from the older sibling's misbehaviour, $Y_{2,igs't'}$, the expected number of exclusions (absences) for similar peers by including

[25] For example, implicit biases against pupils who are expected to be more misbehaved in class may be directed towards pupils with low prior attainment and boys, who exceed girls in rates of externalizing problems throughout childhood and adolescence (Leadbeater et al., 1999).

$(\frac{1}{3}\sum_{r'=t'-1}^{t'+1}\overline{Y}_{2,-igs'r'})$. Essentially, we are assessing how much more (or less) misbehaved the younger sibling is than their near-in-time and homogenous peers, as a function of how much more (or less) misbehaved the older sibling is than their near-in-time and homogenous peers. We assume that the variation in near-in-time homogenous peers across peers is random.

Keeping for simplicity the same notation as in equation (3.1), the extended equation with the three additional covariates $(\frac{1}{3}\sum_{r=t-1}^{t+1}\overline{Y}_{1,-igsr})$, $(\frac{1}{3}\sum_{r'=t'-1}^{t'+1}\overline{Y}_{2,-igs'r'})$ and $\overline{Y}_{1,-igst}$ is:

$$\begin{aligned} Y_{1,igst} =& \beta_0 + \beta_1 Y_{2,igs't'} + \beta_2 X_{1,i} + \beta_3 X_{2,i} + \beta_4 P_i + \beta_5 \frac{1}{3}\sum_{r=t-1}^{t+1}\overline{Y}_{1,-igsr} + \beta_6 \frac{1}{3}\sum_{r'=t'-1}^{t'+1}\overline{Y}_{2,-igs'r'} \\ &+ \beta_7 \overline{Y}_{1,-igst} + \omega_t + \lambda_g + \varepsilon_{1,igst}. \end{aligned} \tag{3.2}$$

Including $(\frac{1}{3}\sum_{r=t-1}^{t+1}\overline{Y}_{1,-igsr})$ controls for school unobservable characteristics that explain the younger sibling's school misbehaviour, and is in principle similar to controlling for school fixed effects that vary by grade, age, gender and prior attainment groups over a three-year period and excluding the younger sibling themselves.[26] Therefore the ordinary least squares (OLS) estimation of equation (3.2) can be thought of as equivalent to a refined school fixed-effect estimation. Similarly, the two-stage least squared (2SLS) estimation of equation (3.2), which additionally instruments $Y_{2,igs't'}$ with $\overline{Y}_{2,-igs't'}$, can be thought of as a combined school fixed effect and instrumental variable estimation. Notice that because we control for $(\frac{1}{3}\sum_{r'=t'-1}^{t'+1}\overline{Y}_{2,-igs'r'})$, the instrument $\overline{Y}_{2,-igs't'}$ exploits only idiosyncratic variation in the misbehaviour of the older siblings' school peers in year t' compared to the peers in the same school, grade, prior attainment group, and gender, but in years $(t'-1)$, t' and $(t'+1)$.

The third identification issue in peer effects estimation is endogeneous peer group selection. This is less of an issue in sibling peer effects estimation as siblings cannot choose one another. Nonetheless endogenous peer group membership may be a concern for our instrumental variable $\overline{Y}_{2,-igs't'}$, the older sibling's homogeneous peers' average misbehaviour, as the effect of $\overline{Y}_{2,-igs't'}$ on $Y_{2,igs't'}$ in the first stage estimation can capture the effect of school unobserved characteristics that explain both sorting into schools and the likelihood to be excluded. Controlling for $(\frac{1}{3}\sum_{r'=t'-1}^{t'+1}\overline{Y}_{2,-igs'r'})$, which

[26] Controlling for $(\frac{1}{3}\sum_{r=t-1}^{t+1}\overline{Y}_{1,-igsr})$ has two main advantages over school fixed effects: (1) it allows us to consider a leave-out-mean rather than a mean including the focal sibling that can cause estimation bias (von Hinke et al., 2019); (2) it does not net out all unobserved school characteristics from the explanatory variables, but only the ones that could confound the spillover effect in misbehaviour, i.e. school characteristics that are correlated with the school peers' misbehavior.

takes account of unobserved variables that can be correlated with both sorting into schools and the homogeneous peers' behaviour, addresses this issue.

Finally, we discuss potential threats to the validity of our instrument. To be valid our instrument must satisfy the exclusion restriction, i.e. it must explain the younger sibling's behaviour only through the older sibling's behaviour. This restriction may not hold if the school peers of the older sibling have relevant interactions with the younger sibling. In England, cohorts are taught strictly separately and there is no grade retention, meaning that older sibling's peers have no direct interactions with a younger sibling in the learning environment. In addition, controlling for the average behaviour of the younger sibling's peers blocks the possible violation of this assumption caused by the older sibling's peers having younger siblings who are in the same grade and same year as the focal younger sibling. It may still be possible that an older sibling's peer might interact with a younger sibling out of school if they live in the same neighbourhood. However we are able to test explicitly for this possibility and our findings suggest that the sibling spillover effect is not driven by interactions between pupils living in the same neighbourhood (see Section 3.5.2).

3.4.2 Identifying variation

Table 3.3 shows the identifying variation in our dependent variables, the younger siblings' exclusion and absence sessions, and our instrumental variables, the older siblings' peers' exclusion and absence sessions. The top panel of Table 3.3 shows the mean and standard deviation of the younger siblings' behaviours, and then this variation net of various additional fixed effects and controls described in Section 3.4.1. The variation in the final specifications are not much reduced at all compared to the unconditional specifications. The bottom panel of Table 3.3 repeats this process but for the take-one-out-mean of the older sibling's homogenous peers' exclusions. Unsurprisingly, when using these more aggregated measures the additional fixed effects and controls do reduce the variation, but only to a modest degree; the standard deviation in the final specifications are about 64% and 56% of the unconditional standard deviations in the older sibling's peers'exclusion and absence sessions, respectively. Finally, Figure 3.4 shows the distributions of both instrumental variables.

Table 3.3 Identifying variation in dependent and instrumental variables

Outcome: Exclusion sessions	*Mean*	*SD*
Total variation	0.376	2.760
Total variation net year FE	-0.000	2.760
Total variation net year & grade FE	-0.000	2.759
Total variation net year & grade FE & individual characteristics	-0.000	2.729
Total variation net year & grade FE & individual & school characteristics	-0.000	2.702
Outcome: Absence sessions	*Mean*	*SD*
Total variation	3.142	12.329
Total variation net year FE	-0.000	12.329
Total variation net year & grade FE	0.000	12.316
Total variation net year & grade FE & individual characteristics	-0.000	11.983
Total variation net year & grade FE & individual & school characteristics	0.000	11.891
N	1,142,734	
IV: Peers' exclusion sessions	*Mean*	*SD*
Total variation	0.347	0.728
Total variation net year FE	-0.000	0.727
Total variation net year & grade FE	0.000	0.726
Total variation net year & grade FE & individual characteristics	0.000	0.680
Total variation net year & grade FE & individual & school characteristics	-0.000	0.465
IV: Peers' absence sessions	*Mean*	*SD*
Total variation	2.871	3.493
Total variation net year FE	-0.000	3.482
Total variation net year & grade FE	-0.000	3.451
Total variation net year & grade FE & individual characteristics	-0.000	3.128
Total variation net year & grade FE & individual & school characteristics	0.000	1.958
N	1,142,734	

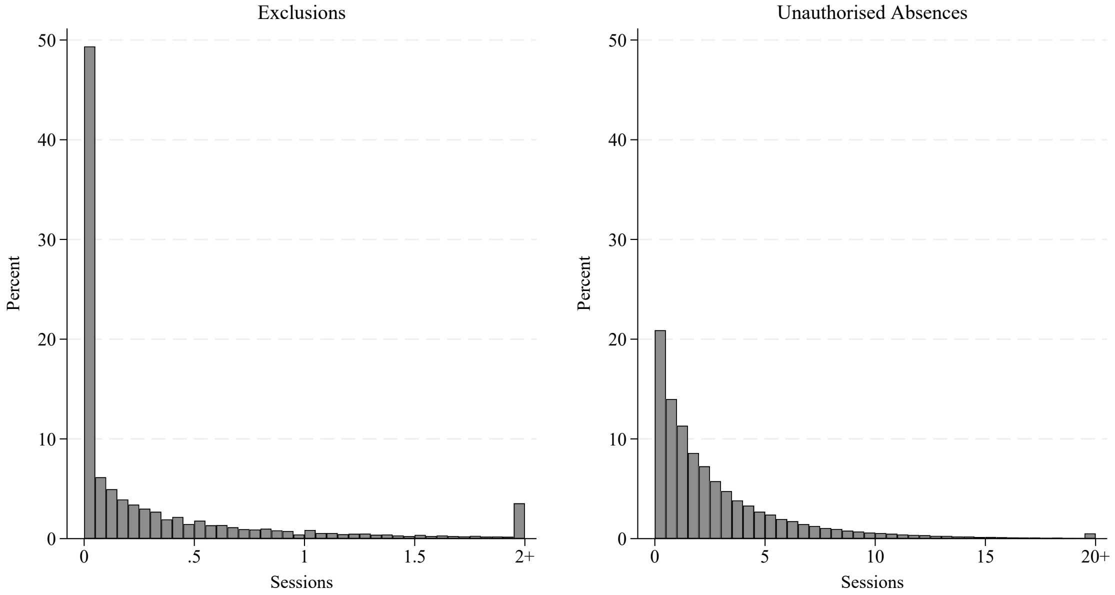
Notes: National Pupil Database, 2007 - 2019. Sample is restricted to the eldest two siblings per family and year, attending a mainstream Secondary school, with no missing covariates, with at least a two year grade gap, and observed in grades 8, 9 or 10. Exclusions are temporary exclusions only. Absences are unauthorised absences only. One session is half a school day. FE stands for Fixed Effect. Individual characteristics are prior attainment at age 11 (standardised), free school meals eligibility, whether first language is English, sibship sex combination, and sibship age gap. School characteristics are three-year rolling averages of homogeneous peers' misbehaviour for both siblings, and the younger sibling's own homogeneous peers' average misbehaviour. All peer measures are constructed using take-one-out means.

3.5 Results

3.5.1 Main results

Table 3.4 reports the main estimates of the older-to-younger sibling spillover effect in temporary exclusions and unauthorised absences using a pooled sample of pupils across grades 8, 9 and 10. The spillover effects are estimates of the impact of one additional session excluded or absent for the older sibling on the same outcome for the younger sibling when they are in the same grade. Column 1 shows the unconditional associations, while columns 2 to 5 display estimates of the spillover effects when successively adding year and grade fixed effects, individual and family controls, and the school peer behaviour rolling averages. Column 5 reports the estimates from Equation 3.2 without the instrumental variable, while column 6 reports the 2SLS estimates.

Figure 3.4 Distribution of older siblings' homogeneous peers' behaviours



Notes: National Pupil Database, 2007 - 2019. Sample is restricted to the eldest two siblings per family and year, attending a mainstream Secondary school, with no missing covariates, with at least a two year grade gap, and observed in grades 8, 9 or 10. Exclusions are temporary exclusions only. Absences are unauthorised absences only. One session is half a school day. Sample size: 1,142,734.

In column 1 of the first panel, the association between the siblings' exclusion sessions in the raw, uncontrolled specification is 0.125, which is about a third of the size of the sample mean 0.376 (see Appendix Table C.4). This means that an additional exclusion session of an older sibling is associated with an increase of 0.125 sessions in the younger sibling. Adding fixed effects for year (column 2) and grade (column 3) does not change the estimate. Column 4 further controls for observable, time-invariant characteristics of individual pupils and their families, and attenuates the estimate somewhat to 0.106. Adding the homogeneous peer behaviour rolling averages for both the older and younger siblings – which is similar to controlling for school fixed effects that vary by grade, age, gender and prior attainment groups over a three-year period and excluding the siblings themselves – attenuates it a little further, to 0.097 (column 5). All estimates are statistically significant at the 1% level.

We address the possibility that after adding this extensive set of fixed effects and controls there may still be unobservable characteristics shared between siblings which increase the likelihood of exclusions (or absences) for both siblings. We do this by instrumenting the older sibling's exclusions and absences using their homogenous (same school, grade, year, prior school attainment group, and gender) peers' related behaviours in a 2SLS estimation. The first stage regresses the older siblings' exclusion sessions

Table 3.4 Sibling spillover effects: Main estimates

Panel A	**Outcome: Exclusion (sessions)**					
	OLS					2SLS
	(1)	(2)	(3)	(4)	(5)	(6)
Spillover effect	0.125*** (0.009)	0.125*** (0.009)	0.125*** (0.009)	0.106*** (0.009)	0.097*** (0.008)	0.031 (0.058)
First-stage coefficient						0.181*** (0.013)
F-test first stage						209.36
Stock-Yogo Critical Value						16.38
Endogeneity test						1.30
Endogeneity test p-value						0.254
Year fixed effects		✓	✓	✓	✓	✓
Grade fixed effects			✓	✓	✓	✓
Individual and family controls				✓	✓	✓
School peer behaviour controls					✓	✓
Observations	1,142,734					
Panel B	**Outcome: Absence (sessions)**					
	OLS					2SLS
	(1)	(2)	(3)	(4)	(5)	(6)
Spillover effect	0.385*** (0.007)	0.385*** (0.007)	0.384*** (0.007)	0.333*** (0.007)	0.321*** (0.007)	0.208** (0.086)
First-stage coefficient						0.103*** (0.008)
F-test first stage						148.49
Stock-Yogo Critical Value						16.38
Endogeneity test						1.70
Endogeneity test p-value						0.192
Year fixed effects		✓	✓	✓	✓	✓
Grade fixed effects			✓	✓	✓	✓
Individual and family controls				✓	✓	✓
School peer behaviour controls					✓	✓
Observations	1,142,734					

Notes: National Pupil Database, 2007 - 2019. Sample is restricted to the eldest two siblings per family and year, attending a mainstream Secondary school, with no missing covariates, with at least a two year grade gap, and observed in grades 8, 9 or 10. Exclusions are temporary exclusions only. Absences are unauthorised absences only. One session is half a school day. The individual and family controls are prior attainment at age 11 (standardised), free school meals eligibility, whether first language is English, sibship sex combination, and sibship age gap. The school peer behaviour controls are three-year rolling averages of homogeneous peers' misbehaviour for both siblings, and the younger sibling's own homogeneous peers' average misbehaviour. The instrumental variable is the average misbehaviour of the older sibling's homogeneous peers. All peer measures are constructed using take-one-out means. * Significant at 10%; ** significant at 5%; *** significant at 1%.

on our instrument, the take-one-out mean of exclusion sessions of their homogeneous peers, as well as on all other controls including the three-year rolling average for the homogeneous peers' behaviour. The second stage then regresses the younger siblings' exclusions on all explanatory variables, with the older siblings' exclusions replaced by their predicted value from the first stage.

The F-statistic for the significance of the instrumental variable in the first stage is large and confirms the relevance of the instrument (third row of column 6). For exclusions, the first-stage coefficient is 0.181, which means that, holding constant the three-year rolling averages of the homogeneous peers, an increase of one session in the mean exclusion sessions of the older sibling's *own* homogeneous peers is associated with an increase of 0.181 exclusion sessions in the older sibling themselves. The resultant 2SLS estimate in column 6 is smaller than the estimate in column 5, though not statistically different from it, and larger standard errors make it less precise. The results of the endogeneity test reject conditional endogeneity of the older siblings' exclusions and therefore suggest that we cannot reject the equality of the fully controlled fixed effects specifications with and without the IV. The estimation without IV (column 5) is more efficient, and we therefore use it as our baseline specification to produce estimates of heterogeneous effects.

The results for absences follow a broadly similar pattern to those for exclusions. The raw correlation in column 1 in the second panel indicates that each additional session an older sibling is absent is associated with an increase of 0.385 sessions in the younger sibling, which is about a tenth of the size of the sample mean 3.14 (see Appendix Table C.4). Again, adding the fixed effects and controls attenuates the estimate, and the IV estimate in column 12 is smaller again. It indicates that a single absence session increase for an older sibling increases the absence sessions of a younger sibling, when they are in that grade, by 0.208 sessions, and is statistically significant at the 5% level. Again, the F-statistic for the significance of the instrumental variable in the first stage is sufficiently large for us to be reassured about the strength of the instrument. And, again, the endogeneity test rejects endogeneity of absence sessions of the older sibling conditional on controls, leading us to choose the fixed effects estimation in column 5 as our baseline estimation.[27]

Overall our results indicate that there are modest, positive sibling spillover effects in non-cognitive skills as proxied by adverse school-related behaviours such as temporary exclusions and unauthorised absences. These results suggest that parental or public investments (or lack of investments) have multiplier effects, and that the benefits of such investments could be underestimated if spillover effects are not included in relevant calculations. Using the means and standard deviations in Appendix Table C.4, the 0.097 spillover effect coefficient for exclusion sessions is equivalent to an effect size of 0.04 standard deviations, or a 25.8% increase compared to the sample mean. For absences, the spillover

[27] Appendix Table C.5 displays the fully controlled OLS specification and IV estimates separately by grade. The OLS estimates are quite consistent across grades for both outcomes, whereas the IV estimates decrease as the grade increases. This may indicate that sibling spillover effects are larger when siblings are younger, and decrease as siblings age. For each outcome and grade combination, the F-statistic for the significance of the instrumental variable in the first stage is sufficiently large and the endogeneity tests fails to reject the equality of the IV and OLS estimates.

effect coefficient of 0.321 is equivalent to an effect size of 0.03 standard deviations, or a change of 10.2% compared to the sample mean. As our measure is of the intensive margin, it is hard to compare to other literature measuring sibling spillover effects in behaviours which use binary outcomes (Altonji et al., 2017; Anand and Kahn, 2023). However, for comparison, Altonji et al. (2017) estimate that an older sibling engaging in substance use increases the probability of the younger sibling engaging by 10-20%. Our estimates therefore do not appear to diverge too much from those in the related literature.

3.5.2 Threats to identification and robustness checks

In this section we briefly consider threats the validity of our identification strategy and present robustness checks for our main specification.

Threats to identification

First, the exclusion restriction requires that the older sibling's homogeneous peers' exclusions and absences have no direct influence on the younger sibling's exclusions and absences. To account for possible violations of this exclusion restriction, in equation (3.2) we control for the misbehaviour of the homogeneous peers of the younger sibling, as the peers of the older and younger sibling may themselves be siblings. However, it may also be the case that some peers of the older sibling live in the same neighbourhood as, and interact with, the younger sibling, even if they do not belong to the same cohort as the younger sibling. Indeed, evidence from England finds that, while neighbourhood peer effects in school *achievement* have not been documented, neighbourhood composition does seem to exert a small effect on pupils' non-cognitive behavioural outcomes (Gibbons et al., 2013).[28]

In our data, we are able to define neighbourhoods based on Lower Level Super Output Areas, which are statistical geographies containing roughly 1,500 residents and 650 households. To test the possibility of neighbourhood interactions, we exclude from the computation of the instrumental variable any peers living in the same neighbourhood as the siblings. Columns 1 to 3 of Table 3.5 displays the results of this exercise. Excluding the older sibling's peers living in the same neighbourhood from the computation of the instrument produces results which are very comparable to

[28] Evidence from the US also finds that social interactions in informal settings outside of school – for example, on the school bus – can have ramifications for what occurs within the classroom (Lenard and Silliman, 2025).

the benchmark estimate, suggesting that direct interaction within neighbourhoods does not threaten our identifying assumption.[29]

Table 3.5 Sibling spillover effects: Estimates excluding same-neighbourhood peers and younger-to-older effects

Panel A	**Outcome: Exclusion (sessions)**					
	Excluding Same-Neighbourhood Peers			**Younger-to-Older**		
	OLS		2SLS	OLS		2SLS
	(1)	(2)	(3)	(4)	(5)	(6)
Spillover effect	0.125***	0.096***	0.065	0.073***	0.057***	-0.008
	(0.009)	(0.008)	(0.076)	(0.005)	(0.004)	(0.024)
First-stage coefficient			0.207***			0.252***
			(0.014)			(0.026)
F-test first stage			222.50			96.59
Stock-Yogo Critical Value			16.38			16.38
Endogeneity test			0.169			6.42
Endogeneity test p-value			0.681			0.011
Year fixed effects						
Grade fixed effects		x	x		x	x
Individual and family controls		x	x		x	x
School peer behaviour controls		x	x		x	x
Observations	1,142,716			1,142,734		

Panel B	**Outcome: Absence (sessions)**					
	Excluding Same-Neighbourhood Peers			**Younger-to-Older**		
	OLS		2SLS	OLS		2SLS
	(1)	(2)	(3)	(4)	(5)	(6)
Spillover effect	0.385***	0.321***	0.186***	0.219***	0.183***	-0.038
	(0.007)	(0.007)	(0.070)	(0.004)	(0.004)	(0.074)
First-stage coefficient			0.165***			0.089***
			(0.011)			(0.010)
F-test first stage			237.55			85.71
Stock-Yogo Critical Value			16.38			16.38
Endogeneity test			3.68			10.03
Endogeneity test p-value			0.055			0.002
Year fixed effects		x	x		x	x
Grade fixed effects		x	x		x	x
Individual and family controls		x	x		x	x
School peer behaviour controls		x	x		x	x
Observations	1,142,716			1,142,734		

Notes: National Pupil Database, 2007 - 2019. Sample is restricted to the eldest two siblings per family and year, attending a mainstream Secondary school, with no missing covariates, with at least a two year grade gap, and observed in grades 8, 9 or 10. Exclusions are temporary exclusions only. Absences are unauthorised absences only. One session is half a school day. The individual and family controls are prior attainment at age 11 (standardised), free school meals eligibility, whether first language is English, sibship sex combination, and sibship age gap. The school peer behaviour controls are three-year rolling averages of homogeneous peers' misbehaviour for both siblings, and the younger sibling's own homogeneous peers' average misbehaviour. The instrumental variable is the average misbehaviour of the older sibling's homogeneous peers. All peer measures are constructed using take-one-out means. Sample size is slightly smaller than in Table 3.4 due to a handful of pupils having no non-LSOA peers. * Significant at 10%; ** significant at 5%; *** significant at 1%.

[29] As mentioned, younger siblings may directly interact with their older sibling's schoolmates at school. In England there is no grade retention and cohorts are taught strictly separately, meaning that the older sibling's peers will be in different classes to the younger sibling. Because of our methodologically-driven sample restriction choices, siblings will also be at least two years apart. Although we have no way of testing this further, these contextual factors form the basis for our assumption that interactions that are relevant to exclusions and absences are unlikely to take place across cohorts while pupils are at school.

Another threat to identification highlighted in the peer effects literature is simultaneity (Manski, 1993): just as older siblings might affect younger siblings, younger siblings might too affect older siblings. Simultaneity is not a major concern in our analysis as the older sibling's misbehaviour is observed at least two years before the younger sibling's, ruling out reverse causality. Nonetheless for transparency we check this assumption explicitly by estimating younger-to-older effects. We use the same estimation strategy as outlined in Section 3.4 but with the superscripts '1' and '2' of the equations reversed to swap the role of the younger sibling with that of the older sibling. In other words we regress the exclusions of the older sibling on the exclusions of the younger sibling and the necessary controls, and we instrument the exclusion of the younger sibling with the take-one-out mean exclusion rate of their homogenous peers. As the effects that we estimate are for when siblings are in *the same grade*, our approach actually estimates effects that are backward in time. For example, we regress an older sibling's behaviour in Grade 8 on a younger sibling's behaviour in Grade 8, which occurred at least two years later. Nonetheless we still find this to be an informative check to ascertain whether the spillover effects occur only in the direction that we expect, or not.

Columns 4 to 6 of Table 3.5 displays the results of this check. While the raw and controlled OLS estimates show a correlation in the siblings' behaviours that is smaller than that in Table 3.4, the 2SLS estimates are very close to zero. The endogeneity tests do not reject endogeneity and therefore reject the equality of the 2SLS estimate and the OLS estimate at the 5% level for exclusion sessions and at the 1% level for absences. These results reassure us that there is no evidence of younger-to-older effects.

Finally, in our particular setting it is important to discuss a further possible component of the sibling spillover effects that we observe: deterrent effects. In particular, the implementation of sanctions or punishment following a misbehaviour is intended to serve as a deterrent for others to behave similarly. If there is a deterrent effect of an older sibling receiving an exclusion, or a punishment (or fine) for a high number of absences, then a younger sibling may be less likely to engage in similar behaviours. This would result in a downward bias on our results.[30] We acknowledge that this is a possiblity that we are unable to check explicitely.

[30] Of course, it may also be that there is a deterrent effect among homogenous peers, and as such as acknowledge that the first stages of our instrumental variable estimations may comprise a combination of deterrent and role modelling effects.

Robustness checks

First, we remove the cohort-size and change-in-enrollment sampling restrictions described in Section 3.3.1. Removing this restriction increases our sample size by approximately 7% (79,320 observations), but may also increase the volatility of our peer group behavioural measures as school cohorts which are small or highly variable in size are retained. The results are displayed in columns 1 and 2 of Table 3.6 and are almost identical to our main estimates.

Next, we check whether adding additional control variables affects our results. In columns 3 and 4 of Table 3.6 we use a time-variant measure of free school meal eligiblity, but find that it changes our estimates very little. In columns 5 and 6 we include quintiles of the Income Deprivation Affecting Children Index (IDACI), which is neighbourhood-based measure of the proportion of all children aged zero to 15 living in income deprived families. Again, we find that the inclusion of IDACI quintiles changes our results little. In columns 7 and 8 we include indicators for whether the pupil has been diagnosed with a Special Educational Need before Grade 8, which is the first grade that we include in our outcome measures. The OLS results are almost identical to our main estimates; the 2SLS estimate is, for exclusions, a little smaller, and, for absences, a little larger, but our main conclusions remain unchanged. Finally, we also include, in columns 9 and 10 of Table 3.6, our main specification with additional controls for ethnicity. Again, the results are virtually unchanged.

We also check whether our instrument remains relevant across different sub-groups. Appendix Table C.2 shows that, while absence sessions are similarly common among male and female pupils, exclusion sessions are considerably more common for male pupils. We may also be concerned that pupils with different levels of prior attainment – and who are therefore in different prior attainment groups – experience exclusions and absences to a greater and lesser extent. Appendix Table C.6 shows the first stage estimates and corresponding F-tests for each outcome and each attainment-by-gender group. As we might expect, the first stage estimates are larger for male compared to female groups in exclusions, and are slightly larger for lower compared to higher prior-attaining groups. Interestingly, an identical pattern exists in absences. However, in all cases the F-statistics for the significance of the instrumental variable in the first stage are greater than 10. This reassures us that our main estimates are not driven by variation in one attainment-by-gender group alone.

Table 3.6 Sibling spillover effects: Robustness checks

Panel A	**Outcome: Exclusion (sessions)**									
	Expanded sample		**Incl. contemp. FSM**		**Incl. IDACI quintile**		**Incl. SEN (Grade 7)**		**Incl. Ethnicity**	
	OLS	2SLS	OLS	2SLS	OLS	2SLS	OLS	2SLS	OLS	2SLS
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Spillover effect	0.097*** (0.008)	0.037 (0.054)	0.097*** (0.008)	0.029 (0.058)	0.096*** (0.008)	0.029 (0.058)	0.094*** (0.008)	0.013 (0.058)	0.096*** (0.008)	0.030 (0.058)
First-stage coefficient		0.179*** (0.012)		0.181*** (0.012)		0.181*** (0.013)		0.177*** (0.013)		0.181*** (0.013)
F-test first stage		225.73		209.50		209.18		182.73		209.13
Stock-Yogo Critical Value		16.38		16.38		16.38		16.38		16.38
Endogeneity test		1.25		1.38		1.33		1.91		1.31
Endogeneity test p-value		0.263		0.240		0.249		0.167		0.253
Year fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Grade fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Individual and family controls	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
School peer behaviour controls	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Observations	1,222,054		1,142,734		1,142,590		1,097,356		1,142,734	
Panel B	**Outcome: Absence (sessions)**									
	Expanded sample		**Incl. contemp. FSM**		**Incl. IDACI quintile**		**Incl. SEN (Grade 7)**		**Incl. Ethnicity**	
	OLS	2SLS	OLS	2SLS	OLS	2SLS	OLS	2SLS	OLS	2SLS
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Spillover effect	0.318*** (0.007)	0.213** (0.075)	0.321*** (0.007)	0.210** (0.086)	0.320*** (0.007)	0.202** (0.087)	0.323*** (0.007)	0.229** (0.091)	0.319*** (0.007)	0.206** (0.086)
First-stage coefficient		0.114*** (0.009)		0.103*** (0.008)		0.102*** (0.008)		0.099*** (0.009)		0.102*** (0.008)
F-test first stage		171.77		147.87		147.26		134.06		147.82
Stock-Yogo Critical Value		16.38		16.38		16.38		16.38		16.38
Endogeneity test		1.98		1.65		1.84		1.07		1.71
Endogeneity test p-value		0.160		0.199		0.175		0.300		0.191
Year fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Grade fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Individual and family controls	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
School peer behaviour controls	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Observations	1,222,054		1,142,734		1,142,590		1,097,356		1,142,734	

Notes: National Pupil Database, 2007 - 2019. Sample is restricted to the eldest two siblings per family and year, attending a mainstream Secondary school, with no missing covariates, with at least a two year grade gap, and observed in grades 8, 9 or 10. Exclusions are temporary exclusions only. Absences are unauthorised absences only. One session is half a school day. The individual and family controls are prior attainment at age 11 (standardised), free school meals eligibility, whether first language is English, sibship sex combination, and sibship age gap. The school peer behaviour controls are three-year rolling averages of homogeneous peers' misbehaviour for both siblings, and the younger sibling's own homogeneous peers' average misbehaviour. The instrumental variable is the average misbehaviour of the older sibling's homogeneous peers. All peer measures are constructed using take-one-out means. FSM stands for free school meals; IDACI stands for the Income Deprivation Affecting Children Index; SEN stands for Special Educational Needs; FE stands for Fixed Effect. * Significant at 10%; ** significant at 5%; *** significant at 1%.

Finally, we re-run our results using a Poisson model. The Poisson model is useful in our setting as both exclusion and absence sessions are count data, and heavily right-skewed. The results from the Poisson model are displayed in Appendix Table C.7. In the instrumental variable estimation for the exclusions outcome (column 6 in Panel A), some parameters were not identified. However, for the absences outcome, the instrumental variable estimate is very similar to the fully-controlled OLS estimate, which is in line with our main results indicating no difference between the 2SLS and OLS estimates. The spillover effect estimates are somewhat smaller than those in Table 3.4. This may indicate non-linearity or overdispersion in our underlying data, and that using OLS may bias our estimates upwards.

3.5.3 Heterogeneity and implications for inequality

In this section, we consider whether the relative magnitudes of sibling spillover effects in exclusions and absences differ by sibling characteristics.

First, we analyse the heterogeneity of sibling spillover effects according to the grade gap between the siblings.[31] The grade gap is an approximate measure of the age gap between siblings, as grades are determined by exact birth dates and school starting age cut-offs. Siblings who are closer in age (grade) might spend more time together and share a closer bond. However this is not the overwhelming finding in the existing literature; Altonji et al. (2017) find larger sibling spillover effects in smoking, drinking and marijuana use for siblings who are more than two years apart, while Nicoletti and Rabe (2019) find similar sibling spillover effects in test scores for siblings who are one, two, and three years apart.

Table 3.7 reports separate results by the grade gap combinations of the younger and older siblings, estimated on separate samples using our baseline specification (equation 3.2 without the instrumental variable). For both exclusion and absence sessions the effects are consistently larger for siblings who are closer in age. The estimates are 0.123, 0.086 and 0.072 exclusion sessions for siblings who are two, three, and four or more grades apart respectively (see columns 1, 2 and 3 of Panel A). For absence sessions they are 0.359, 0.334 and 0.265 (see columns 1, 2 and 3 of Panel B). Recall that siblings with a one-year grade gap were dropped from our sample.

[31] As our sample is restricted to the eldest two siblings per family group identifier and year, we cannot be certain that these are the first- and second-born children in each family but we can be certain that they are consecutively spaced. Given

Table 3.7 Sibling spillover effects in exclusions and absences: Heterogeneity analyses

Panel A	**Outcome: Exclusion (sessions)**								
	Age Gap			**Gender Composition**				**FSM Eligibility**	
	2 years	3 years	4+ years	Female Female	Male Female	Female Male	Male Male	Never-FSM	Ever-FSM
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Spillover effect	0.123*** (0.012)	0.086*** (0.008)	0.072*** (0.008)	0.102*** (0.011)	0.066*** (0.008)	0.132*** (0.019)	0.116*** (0.012)	0.067*** (0.005)	0.107*** (0.010)
Year fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓
Grade fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓
Individual controls	✓	✓	✓	✓	✓	✓	✓	✓	✓
School/peer controls	✓	✓	✓	✓	✓	✓	✓	✓	✓
Observations	443,545	355,602	343,587	283,050	284,432	281,448	293,804	841,601	301,133

Panel B	**Outcome: Absence (sessions)**								
	Age Gap			**Gender Composition**				**FSM Eligibility**	
	2 years	3 years	4+ years	Female Female	Male Female	Female Male	Male Male	Never-FSM	Ever-FSM
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Spillover effect	0.359*** (0.012)	0.334*** (0.014)	0.265*** (0.011)	0.323*** (0.012)	0.298*** (0.013)	0.320*** (0.013)	0.346*** (0.015)	0.247*** (0.010)	0.337*** (0.009)
Year fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓
Grade fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓
Individual controls	✓	✓	✓	✓	✓	✓	✓	✓	✓
School/peer controls	✓	✓	✓	✓	✓	✓	✓	✓	✓
Observations	443,545	355,602	343,587	283,050	284,432	281,448	293,804	841,601	301,133

Notes: National Pupil Database, 2007 - 2019. Sample is restricted to the eldest two siblings per family and year, attending a mainstream Secondary school, with no missing covariates, with at least a two year grade gap, and observed in grades 8, 9 or 10. Exclusions are temporary exclusions only. Absences are unauthorised absences only. One session is half a school day. The individual and family controls are prior attainment at age 11 (standardised), free school meals eligibility, whether first language is English, sibship sex combination, and sibship age gap. The school peer behaviour controls are three-year rolling averages of homogeneous peers' misbehaviour for both siblings, and the younger sibling's own homogeneous peers' average misbehaviour. All peer measures are constructed using take-one-out means. Estimates use the fully-controlled OLS specification. * Significant at 10%; ** significant at 5%; *** significant at 1%.

We also consider whether there is evidence of any heterogeneous sibling spillover effects according to the gender composition of the siblings. For example, Altonji et al. (2017) find that sibling spillover effects for behavioural outcomes (in their case, smoking and marijuana use) are substantially larger for sister pairs. Columns 4 to 7 of Table 3.7 presents the results, estimated using separate samples. In general, the spillover effects are larger when the younger sibling is male, although same-sex female pairings also exhibit relatively large effects, especially in absences. The smallest spillover effects are for an older male sibling and younger female sibling, for both outcomes.

Our final heterogeneity analysis checks whether sibling spillover effects differ by level of economic deprivation. This is an important consideration for equality and for policy. For example, Nicoletti and Rabe (2019) find that positive spillover effects in test scores are smaller for pupils from disadvantaged compared to affluent backgrounds, and that these differences are driven by more disadvantaged families

this, we are able to assess whether siblings who are closer in age have greater spillover effects than those less close in age without contamination by birth order effects (Joensen and Nielsen, 2018).

having, on average, lower-attaining older siblings. They therefore conclude that interventions aimed at increasing the attainment of disadvantaged older siblings will have a more equalising effect than interventions focused on changing the interactions between siblings in disadvantaged compared to non-disadvantaged families. Such considerations are also relevant here, as it could be that the higher average exclusion and absence sessions for disadvantaged pupils shown in Figures 3.1 and 3.2 are driven by risk factors not related to the nature of family interactions, or it could be that a higher likelihood of sibling spillover effects in more disadvantaged families is also playing a part.

Columns 8 and 9 of Table 3.7 show the estimated sibling spillover effects in both exclusions and absences separately for pupils who are ever and never eligible for free school meals. The spillover effect for exclusion sessions is 0.107 sessions for ever-FSM-eligible pupils and 0.067 for never-FSM-eligible pupils (see Panel A). For absence sessions it is 0.337 sessions for ever-FSM-eligible pupils and 0.247 sessions for never-FSM-eligible pupils (see Panel B). We therefore find higher spillovers among disadvantaged pupils than non-disadvantaged. This is likely to amplify inequalities between social groups, as older siblings of ever-FSM-eligible pupils are more often excluded (absent) and also pass this behaviour on to a greater extent than affluent pupils. As opposed to Nicoletti and Rabe (2019), who find the same spillovers across both groups, in our case possible policy responses could include both improving behaviours among disadvantaged pupils (because of the spillover of these onto their younger siblings) *and* addressing any reasons why the spillover is so large for this group.

3.5.4 Mechanisms

The literature identifies two potential mechanisms through which siblings might affect one another: parental investments and direct interaction (e.g. role modelling).

First, sibling spillovers might arise due to behavioural responses by parents. There is a large body of research evidence concerning how parents allocate resources among their children so as to reinforce or compensate initial endowments. For example parents might want to invest more – in terms of time, attention and/or financial resources – in better-endowed children, and thereby reinforce endowment differentials (Becker and Tomes, 1976). However parents may also seek to compensate endowment differentials among their children, and therefore invest more in less-endowed children (Behrman et al., 1982). Empirical evidence supporting one or the other theory is quite mixed and far from conclusive; some papers find evidence of reinforcing behaviour (Datar et al., 2010; Rosales-Rueda, 2014; Rosenzweig and Zhang, 2009), while others find evidence of compensating behaviour (Bharadwaj et al., 2018; Yi et al., 2015).

In our context parents would exhibit reinforcing behaviour by investing more in a non-excluded or non-absent sibling who may been viewed as 'succeeding' more in school. This behaviours would likely lessen the strength of the association in behaviour between siblings, as the non-excluded or non-absent sibling would be less likely to subsequently misbehave. On the other hand, parents would exhibit compensatory behaviour if they invest more in the excluded or absent sibling in an attempt to make subsequent exclusions or absences less likely. This behaviour would likely strengthen the association in behaviours between siblings, as the excluded or absent sibling would be less likely to subsequently misbehave. Parents may also respond to the instrument itself, for example by limiting the time that the siblings spend with any increasingly misbehaving peers. Unfortunately existing evidence about parental responses to the misbehaviour of a child's sibling or peers is scarce. We therefore acknowledge that parental investments may drive some part of the spillover effects that we observe, but remain agnostic as to the direction.

It is also possible that our measured sibling spillover effects in exclusions and absences are driven by direct interactions between siblings. Social learning theory (Bandura, 1977) suggests that siblings learn from one another due to observation and imitation. Through this process, siblings develop similar attributes, attitudes, and behaviours (Defoe et al., 2013; McHale et al., 2012; Whiteman et al., 2011). In social learning theory, older siblings are likely to have a greater influence on younger siblings than the reverse because younger siblings will consider their older sibling as a role model to observe, follow, and imitate (Brim, 1958; Buhrmester, 1992; Pepler et al., 1981).

We explore the direct interaction mechanism in two ways. First, we recall that Table 3.5 provides no evidence of younger-to-older spillover effects and Table 3.7 shows that older-to-younger spillover effects decrease as the grade (age) gap between siblings increases. Together, these results suggest that direct interaction may be a compelling mechanism for our findings as the existence of more prominent older-to-younger effects is supported by social learning theory and we might expect pupils who are closer in age to interact more. Second, we use the reason for exclusion to ascertain the extent to which the sibling spillover effects may be due to older siblings role modelling particular types of behaviour. Section 3.3.2 has already shown there is a stronger correlation in exclusions *for the same reason* between sibling pairs, than in exclusions for different reasons. We now probe this association by testing each combination of reasons using the fully-controlled OLS specification. The results are displayed in Table 3.8, and show that the larger associations are generally retained between siblings for the same reason for exclusion. Overall, though far from conclusive, these results give some further support for the direct interaction mechanism.

Table 3.8 Sibling spillover effects in exclusion sessions: By reason for exclusion

Older			**Younger**		
	Disruptive	Physical	Verbal	Illicit	Other
Disruptive	0.107***	0.021***	0.032***	0.005***	0.013***
	(0.033)	(0.003)	(0.005)	(0.001)	(0.003)
Physical	0.036***	0.034***	0.036***	0.009***	0.019***
	(0.005)	(0.003)	(0.004)	(0.002)	(0.003)
Verbal	0.037***	0.025***	0.044***	0.006***	0.019***
	(0.004)	(0.003)	(0.004)	(0.001)	(0.003)
Illicit	0.031***	0.025***	0.020***	0.007***	0.015***
	(0.006)	(0.005)	(0.004)	(0.002)	(0.004)
Other	0.025***	0.019***	0.027***	0.004***	0.038***
	(0.004)	(0.003)	(0.004)	(0.001)	(0.006)

Notes: National Pupil Database, 2007 - 2019. Sample is restricted to the eldest two siblings per family and year, attending a mainstream Secondary school, with no missing covariates, with at least a two year grade gap, and observed in grades 8, 9 or 10. Exclusions are temporary exclusions only. One session is half a school day. The individual and family controls are prior attainment at age 11 (standardised), free school meals eligibility, whether first language is English, sibship sex combination, and sibship age gap. The school peer behaviour controls are three-year rolling averages of homogeneous peers' misbehaviour for both siblings, and the younger sibling's own homogeneous peers' average misbehaviour. All peer measures are constructed using take-one-out means. Estimates use the fully-controlled OLS specification. * Significant at 10%; ** significant at 5%; *** significant at 1%.

Finally, in our particular setting it is possible that bias and stereotyping may drive part of our sibling spillover effect estimates. This would occur, for example, if teachers and headteachers (principals) develop beliefs about all members of a family based on the behaviour of one sibling belonging to that family. Such beliefs then could, for example, mean that a younger sibling is more likely to be excluded than a peer who does not have a misbehaving older sibling. However we are less concerned about bias driving our results as we find comparable results across both exclusions and absences, and it seems unlikely that bias plays a large role in younger siblings having high rates of absences from school.

3.6 Conclusion

In this paper we explore whether and to what extent inputs into the production of non-cognitive skills arise from sibling spillovers. We proxy non-cognitive skills using measures of temporary exclusions and unauthorised absences, and find evidence of modest sibling spillover effects in both. Having an older sibling who is excluded from school for one session increases the exclusion sessions of a younger sibling by 0.1 sessions. For absences, it is 0.3 sessions. As an effect size proportional to the sample means, the increase for exclusions is around 25%, and for absences around 10%. Both spillover effects are around 3-4% of a standard deviation, which is a small effect.

We also estimate heterogeneous sibling spillover effects in temporary exclusions and unauthorised absences. We find that siblings who are closer in age seem to have larger spillover effects. This makes sense given that siblings who are closer in age may spend more time together, be experiencing more similar stages of puberty, and/or have access to more similar opportunities, items or substances. However this is not a consistent finding in the wider literature. We also find that spillover effects are larger for disadvantaged pupils compared to non-disadvantaged pupils. Our results indicate that direct interaction (role modelling) is likely to be a convincing explanation of the mechanism of the spillover effects that we find. They also indicate that ignorance of such spillover effects may result in understating the positive externalities of successful policies and interventions which reduce adverse school-related behaviours among older siblings.

Finally, as shown in Figures 3.1 and 3.2, differences in non-cognitive skills across groups are an increasing social equity concern. Successful policy responses require a detailed understanding of the causal mechanisms behind the increasing rates and rising disparity in exclusions and absences between groups. This paper suggests one possible mechanism: pupils from the same family influence one another's non-cognitive skill development and behaviour. Further research is needed to uncover the reason for the larger spillover effects between groups, but at the very least policymakers should take note that there are likely to be positive multiplier effects to investments that reduce exclusions and/or absences for individuals who have younger siblings, and that these may help to reduce inequities between differently-advantaged groups. The findings from this paper also suggest that an older siblings' temporary exclusions and unauthorised absences may be a reliable signal for early intervention to prevent temporary exclusions and unauthorised absences in the younger sibling.

Conclusions

In *The Next Big Thing in School Improvement*, economist Rebecca Allen, headteacher Matthew Evans and classroom teacher Ben White explore why it is that the school system often resists our attempts to improve it. ‘[T]he further we sit from the complex realities of the classroom,’ they argue, ‘the more our attempts to influence schooling are mired by misunderstanding, presumption and over-simplification’ (Allen et al., 2021). Without acknowledging the constant interplay between schools and wider society, or the indelible complexities of how schools actually work, well-meaning policies are likely to lead to limited or even unintended consequences. And the further we move from that complexity, the more likely we are to get it wrong.

The three essays that constitute this thesis unite on a similar conclusion: while educational policies and institutional practices are often designed to promote learning and equality of opportunity, they often fall short because they fail to account for practical, familial or structural constraints. Chapter 1 shows that teacher-assigned grades can perpetuate biases, suggesting that even well-meaning policies (such as trusting teacher judgment) can produce unequal outcomes if social biases are not addressed. Chapter 2 finds that a large-scale, early childhood curriculum reform (the Foundation Phase) had no lasting effect on outcomes, challenging the idea that intended changes in pedagogy alone will lead to better or more equitable educational trajectories. And Chapter 3 illustrates how peer and family dynamics, such as sibling interactions, can influence an individual’s own educational experiences, underscoring the very large extent to which factors outside of the control of schools (such as the family environment) shape children’s outcomes.

This thesis also highlights how, while structural or interpersonal dynamics can limit the intended effects of education policy, the complexities of the schooling system also present challenges to the education policy evaluator.

First, in contrast to a supposed scientific ‘ideal’, the effects of education policies are context-dependent and rarely universal. Chapter 1 shows that, in England, many ethnic minority groups actually perform better than their White British peers in their end-of-compulsory-schooling examinations. Though this makes complete sense in the English context, what it means for the ways that teachers perceive their students is different to other contexts where comparable groups almost universally perform worse; the dynamics of teachers’ biases are not a universal truth. Chapter 2, meanwhile, finds that a play-based early childhood curriculum inspired by Reggio Emilia in Italy, Te Whãriki in New Zealand, and multiple programmes across Scandinavia, failed to lead to statistically significant, positive effects on schooling outcomes in Wales in the medium term. Whether this means that the policy itself has ‘failed’ depends on why no effects emerged. If the issue was one of training and resources, or even the way that the policy was viewed by schools and families, then assuming comparability to very different contexts may be misguided.

Second, schools are a far cry from experimental laboratories. Each one is a unique and dynamic institution, composed of individual teachers exercising professional judgment and whose established practices are not easily shifted by top-down reform. Spillovers are inevitable, both within schools and across localities; a change in how one subject is taught may influence others, just as shifts in one school’s hiring or admissions policies can ripple through neighbouring schools. Multiple overlapping policy reforms – some originating outside the education sector – simultaneously shape pupil outcomes. And researchers must frequently make do with outcome measures which are poor proxies for the constructs of real interest and which also change over time, disappear altogether, or suffer from ceiling effects that obscure variation. In short, the real-world complexity of schooling rarely lends itself to the tidy assumptions of policy evaluation.

Nonetheless, while uncovering causal effects under the level of complexity of school systems is difficult, this thesis attempts to do so. In Chapter 1, we use Gelbach decompositions alongside the exogeneous change in assessment methods during the Covid-19 pandemic to evaluate how the change in assessment methods affects grade gaps between students from different ethnic minority groups and White British students. Ruling out alternative explanations of the pattern of grade gap changes that we find leads us to conclude that teacher stereotyping is a convincing explanation of the grade gap changes that we observe. In Chapter 2, I use a difference-in-differences methodology to compare the

43 schools in which the Foundation Phase policy was piloted to all other schools in Wales, therein accounting both for differences between the groups of schools before the policy was introduced and differences over time that affected all schools. I show that crucial assumptions are met which allow us to interpret the results as causal. In Chapter 3, we implement a novel instrumental variable design to overcome the identification issues inherent in evaluating the effects of peers on pupils' outcomes by using a three-year rolling average of pupils' homogeneous school peers. In the future, this approach could be used more widely to help to control for selection into schools, or to achieve quasi-random assignment of peers (conditional on covariates).

Finally, there remain some substantial limitations and areas for future investigation in the research presented in this thesis. In Chapter 1, we refrain from concluding that the unexplained grade gap changes that we observe entirely constitute stereotyping effects for two reasons. First, unlike other papers on stereotyping in educational assessment (Carlana, 2019), we have no way of estimating a measure of bias at the teacher-level and are therefore unable to model bias directly. Second, a causal interpretation is based on the assumption that changes in the levels or returns to *unobserved* pupil characteristics across groups affect grades in the same direction across subjects. Though this is highly likely, it is not something that we can test. An interesting extension of the analysis in Chapter 1 would be to relate teachers' characteristics to the grade gap changes that we observe. However unfortunately the only data on teachers' characteristics that is available (such as teacher ethnicity, qualification and age) are only available at the school aggregate level, and cannot be linked directly with pupils or even to specific cohorts or subjects.

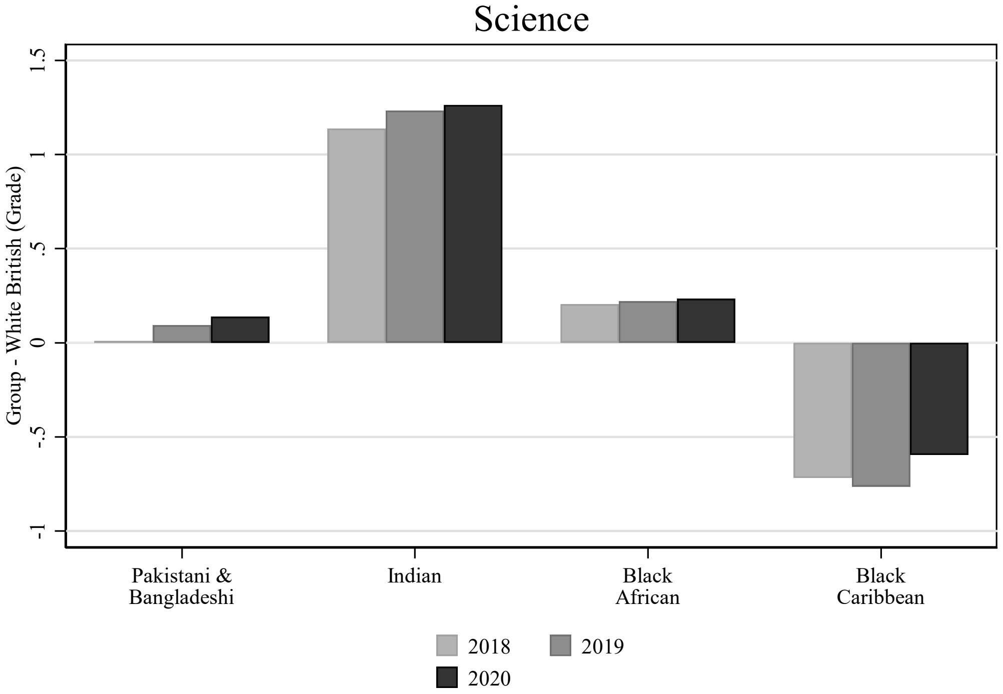
In Chapter 2 the primary limitation is data availability. First, as the first year of the SAIL Databank education dataset is not long before the implementation of the Foundation Phase begins, I have limited pre-intervention years for particular Pilot groups and outcomes. This is particularly the case for Pilot Group 1a and for the GCSE outcomes, and impacts how strongly we can interpret any pre-trends tests as evidence in support of the conditional parallel trends assumption. Second, additional data measures would also help to more fully explore the impacts of the Foundation Phase policy. These include subject-level GCSE examination data, which exists and has been requested. They also include data on Further Education College attendance, which exists, and comparable measures across both Pilot and control schools at age seven, which unfortunately does not. It would be both useful and fascinating to have access to survey data which might elucidate parental responses to the introduction of the Foundation Phase pilot in their children's schools. Unfortunately, though the Millenium Cohort Study contains such data, the sample size is too small to sufficiently power any statistical analyses.

In Chapter 3 we are primarily restricted in the types of behavioural outcomes that are available to researchers in administrative data. Having access to lower-level behavioural infringements such as detentions, for example, would allow us to estimate spillover effects at a more precise level. They would also likely give sufficient variation in Primary school for us to estimate value-added behavioural spillover effects using a sibling's exposure to new peers in Secondary school, which would be an alternative identification strategy to the one that we do employ. Our identification strategy in Chapter 3 also relies on assumptions about how schools organise pupils into classes and which pupils spend time together. Access to data such as lower-level behavioural outcomes and classroom groupings would therefore considerably expand the precision that researchers are able to reach in analyses of what happens within schools.

Ultimately, both designing education policy and designing education policy evaluations are complex tasks. This thesis, however, hopes to contribute some knowledge in the service of both.

A Appendix for Chapter 1

Figure A.1 Science grade gaps by ethnic group and year



Notes: National Pupil Database, 2017 - 2020. Sample comprises age 16 pupils in mainstream schools in England with no missing data. Exam grades are the May and June (end of year) examinations only. The Science grade is the mean of all Science grades awarded (‘double award’ or ‘triple award’). Grade gaps for 2017 are not included as reformed science GCSE qualifications were first awarded in 2018.

Table A.1 Cohort characteristics by year

	2017	2018	2019	2020
Pupil characteristics				
Male	0.498	0.500	0.500	0.502
	(0.500)	(0.500)	(0.500)	(0.500)
Ethnicity:				
White British	0.748	0.732	0.719	0.711
	(0.434)	(0.443)	(0.450)	(0.453)
Pakistani & Bangladeshi	0.056	0.061	0.062	0.060
	(0.230)	(0.240)	(0.242)	(0.238)
Indian	0.025	0.027	0.027	0.027
	(0.155)	(0.161)	(0.163)	(0.162)
Black African	0.029	0.031	0.033	0.035
	(0.169)	(0.175)	(0.179)	(0.183)
Black Caribbean	0.013	0.013	0.014	0.013
	(0.114)	(0.115)	(0.117)	(0.114)
Multiethnic	0.044	0.046	0.049	0.052
	(0.205)	(0.210)	(0.217)	(0.222)
Any Other Ethnic Group	0.085	0.088	0.095	0.102
	(0.278)	(0.284)	(0.293)	(0.302)
Free school meals	0.253	0.249	0.246	0.244
	(0.435)	(0.432)	(0.430)	(0.430)
Neighbourhood deprivation	0.192	0.192	0.194	0.196
	(0.138)	(0.138)	(0.139)	(0.139)
Special educational needs	0.109	0.109	0.113	0.120
	(0.311)	(0.312)	(0.317)	(0.325)
First language not English	0.131	0.138	0.141	0.141
	(0.337)	(0.345)	(0.348)	(0.349)
Maths prior attainment (age 11)	0.000	0.000	0.000	0.000
	(1.000)	(1.000)	(1.000)	(1.000)
English prior attainment (age 11)	0.000	0.000	0.000	0.000
	(1.000)	(1.000)	(1.000)	(1.000)
School characteristics				
Selective admissions	0.044	0.048	0.047	0.046
	(0.206)	(0.213)	(0.212)	(0.209)
Region:				
London	0.138	0.142	0.144	0.143
	(0.345)	(0.349)	(0.351)	(0.351)
East of England	0.117	0.116	0.118	0.115
	(0.321)	(0.321)	(0.322)	(0.319)
North East	0.150	0.150	0.148	0.148
	(0.357)	(0.357)	(0.355)	(0.355)
North West	0.141	0.138	0.142	0.139
	(0.348)	(0.345)	(0.349)	(0.346)
South East	0.159	0.161	0.165	0.162
	(0.366)	(0.367)	(0.371)	(0.368)
South West	0.098	0.095	0.083	0.092
	(0.298)	(0.293)	(0.276)	(0.289)
East Midlands	0.084	0.085	0.087	0.087
	(0.277)	(0.279)	(0.281)	(0.282)
West Midlands	0.112	0.113	0.113	0.114
	(0.315)	(0.317)	(0.317)	(0.318)
Urban	0.876	0.877	0.877	0.874
	(0.330)	(0.328)	(0.329)	(0.332)
Size of cohort	190	189	192	196
	(62.5)	(63.6)	(62.3)	(63.0)
School governance type:				
Local Authority Maintained	0.337	0.301	0.270	0.244
	(0.473)	(0.459)	(0.444)	(0.429)
Single Academy Trust	0.185	0.192	0.198	0.192
	(0.388)	(0.394)	(0.398)	(0.394)
Multi Academy Trust	0.416	0.454	0.502	0.541
	(0.493)	(0.498)	(0.500)	(0.498)
Other	0.062	0.054	0.030	0.023
	(0.241)	(0.225)	(0.171)	(0.151)
Proportion free school meals	0.252	0.246	0.243	0.240
	(0.151)	(0.148)	(0.144)	(0.140)
Average neighbourhood deprivation	0.193	0.193	0.194	0.196
	(0.089)	(0.088)	(0.089)	(0.089)
School value-added (lagged)	0.022	0.033	0.036	0.026
	(0.335)	(0.406)	(0.423)	(0.435)
N	449,673	439,279	444,842	489,748

Notes: National Pupil Database, 2017 - 2020. Sample comprises age 16 pupils in mainstream schools in England with no missing data. Free school meals indicates if a pupil is known to have been eligible for free school meals in the past six years. Neighbourhood deprivation is a rank based on the proportion of all children aged 0 to 15 living in income deprived households in the pupil's local area of residence (IDACI score). Prior attainment scores are standardised by year and subject with mean zero and standard deviation of one. School value-added refers to how many average grades higher or lower the pupils in that school achieve across eight qualifying GCSE subjects compared to pupils across the country who score comparatively at age 11. Standard deviations in parentheses.

Table A.2 Cohort characteristics by group

	WB	PB	I	BA	BC
Pupil characteristics					
Male	0.501	0.496	0.509	0.483	0.483
	(0.500)	(0.500)	(0.500)	(0.500)	(0.500)
Free school meals	0.213	0.387	0.144	0.494	0.466
	(0.410)	(0.487)	(0.351)	(0.500)	(0.499)
Neighbourhood deprivation	0.172	0.274	0.193	0.316	0.301
	(0.133)	(0.115)	(0.112)	(0.126)	(0.123)
Special educational needs	0.118	0.093	0.060	0.097	0.171
	(0.323)	(0.291)	(0.238)	(0.296)	(0.377)
First language not English	0.004	0.702	0.595	0.529	0.025
	(0.062)	(0.457)	(0.491)	(0.499)	(0.156)
Maths prior attainment (age 11)	0.006	-0.138	0.288	-0.099	-0.330
	(0.992)	(1.035)	(0.945)	(0.998)	(1.013)
English prior attainment (age 11)	0.040	-0.266	0.095	-0.110	-0.235
	(0.991)	(0.992)	(0.965)	(0.972)	(0.958)
School characteristics					
Selective admissions	0.041	0.039	0.151	0.050	0.013
	(0.198)	(0.194)	(0.358)	(0.217)	(0.115)
Region:					
London	0.056	0.262	0.291	0.600	0.640
	(0.230)	(0.440)	(0.454)	(0.490)	(0.480)
East of England	0.126	0.070	0.064	0.075	0.055
	(0.332)	(0.255)	(0.245)	(0.263)	(0.229)
North East	0.169	0.185	0.063	0.043	0.029
	(0.375)	(0.388)	(0.243)	(0.203)	(0.167)
North West	0.157	0.153	0.097	0.055	0.029
	(0.363)	(0.360)	(0.295)	(0.228)	(0.169)
South East	0.178	0.079	0.131	0.086	0.043
	(0.382)	(0.270)	(0.337)	(0.280)	(0.203)
South West	0.112	0.010	0.028	0.020	0.018
	(0.315)	(0.101)	(0.164)	(0.140)	(0.132)
East Midlands	0.095	0.039	0.133	0.042	0.036
	(0.293)	(0.192)	(0.339)	(0.201)	(0.186)
West Midlands	0.107	0.202	0.194	0.079	0.149
	(0.309)	(0.402)	(0.395)	(0.270)	(0.357)
Urban	0.843	0.988	0.977	0.985	0.989
	(0.364)	(0.108)	(0.151)	(0.121)	(0.105)
Size of cohort	192	199	199	185	180
	(64.0)	(58.6)	(69.4)	(54.5)	(54.4)
School governance type:					
Local Authority Maintained	0.274	0.362	0.264	0.356	0.345
	(0.446)	(0.481)	(0.441)	(0.479)	(0.475)
Single Academy Trust	0.188	0.153	0.254	0.208	0.192
	(0.391)	(0.360)	(0.436)	(0.406)	(0.394)
Multi Academy Trust	0.491	0.462	0.459	0.409	0.431
	(0.500)	(0.499)	(0.498)	(0.492)	(0.495)
Other	0.046	0.023	0.023	0.028	0.032
	(0.209)	(0.150)	(0.150)	(0.165)	(0.176)
Proportion free school meals	0.221	0.375	0.242	0.356	0.377
	(0.128)	(0.162)	(0.146)	(0.171)	(0.161)
Average neighbourhood deprivation	0.178	0.265	0.206	0.269	0.275
	(0.083)	(0.082)	(0.075)	(0.082)	(0.077)
School value-added (lagged)	-0.019	0.164	0.299	0.181	0.093
	(0.381)	(0.414)	(0.454)	(0.420)	(0.426)
N	1,326,331	109,634	48,111	58,848	24,388

Notes: National Pupil Database, 2017 - 2020. ‘WB’ stands for White British, ‘P&B’ stands for Pakistani and Bangladeshi, ‘I’ Indian, ‘BA’ Black African, ‘BC’ Black Caribbean. Sample comprises age 16 pupils in mainstream schools in England, with no missing data, in years 2017, 2018, 2019 or 2020. Free school meals indicates if a pupil is known to have been eligible for free school meals in the past six years. Neighbourhood deprivation is a rank based on the proportion of all children aged 0 to 15 living in income deprived households in the pupil’s local area of residence (IDACI score). Prior attainment scores are standardised by year and subject with mean zero and standard deviation of one. School value-added refers to how many average grades higher or lower the pupils in that school achieve across eight qualifying GCSE subjects compared to pupils across the country who score comparatively at age 11. Standard deviations in parentheses.

Table A.3 Detailed Gelbach decomposition of the change in grade gaps by ethnic group

	Maths				**English**			
	P&B	I	BA	BC	P&B	I	BA	BC
Raw gap change	0.101***	0.091***	0.115***	0.201***	-0.060***	-0.041	-0.103***	0.082**
	(0.023)	(0.028)	(0.024)	(0.035)	(0.020)	(0.028)	(0.023)	(0.036)
Explained gap change	0.065***	-0.009	0.018	0.056*	0.080***	0.036*	0.075***	0.104***
	(0.022)	(0.024)	(0.021)	(0.030)	(0.019)	(0.022)	(0.020)	(0.025)
Amount explained by:								
Male	0.001	0.000	0.002	0.003*	0.003	0.004	-0.004	0.005
	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)	(0.004)	(0.004)	(0.005)
Socio-economic status:								
Free school meals	-0.002	0.009***	0.002	-0.001	0.003	0.007***	0.012***	0.008*
	(0.002)	(0.002)	(0.003)	(0.004)	(0.002)	(0.002)	(0.003)	(0.005)
Neighbourhood deprivation	0.008***	0.003	0.014***	0.007	0.012***	0.004*	0.019***	0.011**
	(0.003)	(0.002)	(0.004)	(0.004)	(0.003)	(0.002)	(0.004)	(0.005)
Special educational needs	0.004***	0.007***	0.004***	-0.003	0.005***	0.007***	0.006***	0.000
	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	(0.004)
First language not English	0.011	0.008	0.005	0.000	0.011	0.009	0.006	0.000
	(0.009)	(0.008)	(0.007)	(0.001)	(0.010)	(0.008)	(0.007)	(0.001)
Prior attainment (age 11)	0.024	-0.026	-0.015	0.027	0.002	0.036***	0.030***	0.047***
	(0.016)	(0.019)	(0.016)	(0.026)	(0.010)	(0.014)	(0.012)	(0.016)
School characteristics:								
Selective admissions	0.000	-0.002	0.001	0.001	0.001	-0.005	0.001	0.002
	(0.001)	(0.004)	(0.001)	(0.001)	(0.001)	(0.004)	(0.001)	(0.001)
Region	0.007	-0.001	0.013	0.014	-0.012**	-0.019***	-0.025**	-0.024**
	(0.005)	(0.005)	(0.009)	(0.011)	(0.006)	(0.006)	(0.011)	(0.012)
Urban	0.003	0.003	0.003	0.003	-0.001	-0.001	-0.001	-0.001
	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)
Size of cohort	-0.003*	-0.002	0.003**	0.006***	-0.002	-0.002	0.002**	0.005***
	(0.001)	(0.002)	(0.001)	(0.002)	(0.001)	(0.002)	(0.001)	(0.002)
School governance type	-0.001	0.000	0.001	0.001	0.000	0.001	0.002	0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)
Prop. free school meals	0.019	0.001	0.016	0.020	0.048***	0.006**	0.043***	0.049***
	(0.015)	(0.002)	(0.014)	(0.016)	(0.016)	(0.003)	(0.015)	(0.017)
Avg. neigh'hood deprivation	-0.023	-0.006	-0.024	-0.026*	-0.003	-0.003	-0.003	-0.002
	(0.014)	(0.004)	(0.014)	(0.016)	(0.016)	(0.005)	(0.016)	(0.018)
School value-added (lagged)	0.015*	-0.003	-0.008	0.004	0.012	-0.008	-0.011	0.003
	(0.009)	(0.009)	(0.008)	(0.010)	(0.009)	(0.010)	(0.008)	(0.010)
Unexplained gap change	0.036*	0.100***	0.096***	0.145***	-0.140***	-0.077***	-0.179***	-0.022
	(0.019)	(0.020)	(0.019)	(0.026)	(0.020)	(0.025)	(0.021)	(0.030)
N	934,590				934,590			

Notes: National Pupil Database, 2017 - 2020. 'P&B' stands for Pakistani and Bangladeshi, 'I' Indian, 'BA' Black African, 'BC' Black Caribbean. Sample comprises age 16 pupils in mainstream schools in England with no missing data. Exam grades are the May and June (end of year) examinations only. The English grade is the highest of English Literature and English Language if pupils received grades for both. The gap is the gap in average grades between the ethnic minority group and White British pupils. Free school meals indicates if a pupil is known to have been eligible for free school meals in the past six years. Neighbourhood deprivation is a rank based on the proportion of all children aged 0 to 15 living in income deprived households in the pupil's local area of residence (IDACI score). Prior attainment scores are standardised by year and subject with mean zero and standard deviation of one. Region contains 9 categories and school governance type 4 categories. School value-added refers to how many average grades higher or lower the pupils in that school achieve across eight qualifying GCSE subjects compared to pupils across the country who score comparatively at age 11. Standard errors in parentheses, clustered at the school level. * Significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.4 Gelbach decomposition returns

	Maths		English	
	2019	*2020	2019	*2020
Constant	5.057***	0.468***	5.831***	0.134***
	(0.033)	(0.033)	(0.037)	(0.038)
Male	-0.089***	-0.130***	-0.610***	0.086***
	(0.006)	(0.007)	(0.006)	(0.007)
Free school meals	-0.355***	-0.050***	-0.418***	-0.018**
	(0.005)	(0.007)	(0.006)	(0.007)
Neighbourhood deprivation	-1.115***	0.079***	-1.138***	0.114***
	(0.023)	(0.027)	(0.024)	(0.028)
Special educational needs	-0.198***	-0.074***	-0.514***	-0.029***
	(0.009)	(0.010)	(0.009)	(0.011)
First language not English	0.170***	0.020	0.195***	0.022
	(0.011)	(0.013)	(0.012)	(0.014)
Prior attainment (age 11)	1.412***	-0.056***	0.981***	0.014***
	(0.004)	(0.003)	(0.003)	(0.004)
Selective admissions	0.420***	-0.019	0.435***	-0.042
	(0.032)	(0.025)	(0.030)	(0.027)
Region (ref: London):				
East of England	-0.022	-0.011	-0.231***	0.074***
	(0.023)	(0.022)	(0.025)	(0.025)
North East	0.003	-0.010	-0.156***	0.057**
	(0.022)	(0.021)	(0.024)	(0.023)
North West	-0.062***	-0.016	-0.135***	0.019
	(0.022)	(0.022)	(0.026)	(0.025)
South East	-0.022	-0.032	-0.225***	0.027
	(0.022)	(0.022)	(0.024)	(0.024)
South West	0.005	-0.029	-0.237***	0.074**
	(0.025)	(0.025)	(0.027)	(0.028)
East Midlands	0.009	-0.080***	-0.189***	0.024
	(0.027)	(0.026)	(0.028)	(0.030)
West Midlands	-0.030	-0.043*	-0.186***	0.022
	(0.023)	(0.022)	(0.027)	(0.025)
Urban	0.003	0.021	0.019	-0.008
	(0.017)	(0.015)	(0.017)	(0.019)
Size of cohort	-0.000	-0.000***	-0.000	-0.000***
	(0.000)	(0.000)	(0.000)	(0.000)
School governance type (ref: Community):				
Single Academy Trust	0.001	0.004	0.022	0.007
	(0.016)	(0.016)	(0.018)	(0.018)
Multi Academy Trust	0.025*	-0.008	0.044***	-0.013
	(0.013)	(0.014)	(0.014)	(0.014)
Other	0.000	0.041	-0.017	0.000
	(0.034)	(0.037)	(0.039)	(0.045)
Proportion free school meals	-0.023	0.137	-0.293***	0.322***
	(0.100)	(0.104)	(0.104)	(0.113)
Average neighbourhood deprivation	-0.067	-0.262	0.688***	-0.035
	(0.156)	(0.159)	(0.166)	(0.181)
School value-added (lagged)	0.615***	-0.035**	0.618***	-0.050***
	(0.016)	(0.017)	(0.018)	(0.018)
N	934,590		934,590	

Notes: National Pupil Database, 2017 - 2020. Columns labelled ‘2019’ refer to the returns to a characteristic in 2019. Columns labelled ‘*2020’ are the interaction terms for each characteristic and the 2020 indicator. Sample comprises age 16 pupils in mainstream schools in England with no missing data. Exam grades are the May and June (end of year) examinations only. The English grade is the highest of English Literature and English Language if pupils received grades for both. The gap is the gap in average grades between the ethnic minority group and White British pupils. Free school meals indicates if a pupil is known to have been eligible for free school meals in the past six years. Neighbourhood deprivation is a rank based on the proportion of all children aged 0 to 15 living in income deprived households in the pupil’s local area of residence (IDACI score). Prior attainment scores are standardised by year and subject with mean zero and standard deviation of one. School value-added refers to how many average grades higher or lower the pupils in that school achieve across eight qualifying GCSE subjects compared to pupils across the country who score comparatively at age 11. Standard errors in parentheses, clustered at the school level. * Significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.5 Mean grades by subject, group and year

	WB	PB	I	BA	BC
Maths					
2017	4.651	4.614	5.797	4.804	3.948
	(2.003)	(2.045)	(2.063)	(1.987)	(1.883)
2018	4.699	4.655	5.813	4.809	3.829
	(1.996)	(2.037)	(2.046)	(1.978)	(1.858)
2019	4.697	4.734	5.920	4.787	3.802
	(1.995)	(2.060)	(2.071)	(1.999)	(1.826)
2020	4.974	5.111	6.287	5.177	4.279
	(1.959)	(2.011)	(1.964)	(1.908)	(1.784)
N	1,326,331	109,634	48,111	58,848	24,388
English (highest)					
2017	5.146	5.331	6.125	5.558	4.900
	(1.898)	(1.813)	(1.767)	(1.762)	(1.786)
2018	5.132	5.296	6.088	5.525	4.761
	(1.897)	(1.819)	(1.766)	(1.772)	(1.793)
2019	5.130	5.340	6.118	5.493	4.698
	(1.899)	(1.836)	(1.793)	(1.776)	(1.761)
2020	5.326	5.475	6.273	5.585	4.976
	(1.827)	(1.754)	(1.679)	(1.683)	(1.646)
N	1,326,331	109,634	48,111	58,848	24,388
English Literature					
2017	4.720	4.986	5.810	5.253	4.550
	(1.973)	(1.910)	(1.856)	(1.837)	(1.892)
2018	4.749	5.011	5.817	5.253	4.453
	(1.971)	(1.905)	(1.839)	(1.850)	(1.898)
2019	4.790	5.100	5.853	5.232	4.420
	(1.960)	(1.900)	(1.849)	(1.826)	(1.829)
2020	5.089	5.292	6.089	5.403	4.764
	(1.881)	(1.809)	(1.724)	(1.724)	(1.702)
N	1,268,229	103,987	46,455	57,367	23,776
English Language					
2017	4.758	4.809	5.548	4.942	4.387
	(1.842)	(1.730)	(1.742)	(1.695)	(1.682)
2018	4.756	4.783	5.521	4.929	4.245
	(1.838)	(1.726)	(1.740)	(1.702)	(1.658)
2019	4.754	4.822	5.593	4.928	4.207
	(1.840)	(1.767)	(1.762)	(1.725)	(1.666)
2020	5.133	5.202	5.988	5.288	4.711
	(1.803)	(1.724)	(1.670)	(1.652)	(1.605)
N	1,313,822	108,428	47,653	58,209	24,189
English mean					
2017	4.731	4.892	5.676	5.095	4.462
	(1.826)	(1.726)	(1.702)	(1.663)	(1.693)
2018	4.742	4.893	5.665	5.089	4.348
	(1.825)	(1.727)	(1.697)	(1.680)	(1.691)
2019	4.756	4.945	5.710	5.077	4.306
	(1.826)	(1.750)	(1.719)	(1.687)	(1.670)
2020	5.096	5.234	6.031	5.339	4.731
	(1.807)	(1.723)	(1.654)	(1.646)	(1.609)
N	1,326,331	109,634	48,111	58,848	24,388

Notes: National Pupil Database, 2017 - 2020. 'WB' stands for White British, 'P&B' for Pakistani and Bangladeshi, 'I' Indian, 'BA' Black African, 'BC' Black Caribbean. Sample comprises age 16 pupils in mainstream schools in England with no missing data. Exam grades are the May and June (end of year) examinations only. Standard deviations in parentheses. * Significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.6 Gelbach decomposition by English outcome

	English Language				**English Literature**			
	P&B	I	BA	BC	P&B	I	BA	BC
Raw gap change	0.001	0.016	-0.020	0.125***	-0.108***	-0.063**	-0.128***	0.044
	(0.021)	(0.028)	(0.023)	(0.035)	(0.023)	(0.030)	(0.024)	(0.037)
Explained gap change	0.091***	0.066***	0.091***	0.104***	0.066***	0.041*	0.070***	0.092***
	(0.019)	(0.021)	(0.019)	(0.024)	(0.022)	(0.023)	(0.021)	(0.026)
Amount explained by:								
Male	0.004	0.004	-0.003	0.006	0.004	0.006	-0.005	0.005
	(0.003)	(0.004)	(0.004)	(0.005)	(0.003)	(0.005)	(0.004)	(0.006)
Socio-economic status	0.008*	0.012***	0.019***	0.008	0.015***	0.011***	0.033***	0.021***
	(0.004)	(0.003)	(0.005)	(0.006)	(0.005)	(0.004)	(0.006)	(0.007)
Special educational needs	0.006***	0.010***	0.007***	-0.002	0.004	0.004*	0.005**	0.001
	(0.002)	(0.002)	(0.002)	(0.004)	(0.002)	(0.002)	(0.002)	(0.004)
First language not English	0.027***	0.022***	0.017**	0.001	0.010	0.007	0.004	0.000
	(0.009)	(0.008)	(0.007)	(0.001)	(0.011)	(0.009)	(0.008)	(0.001)
Prior attainment (age 11)	-0.001	0.037***	0.030***	0.045***	-0.009	0.032**	0.026**	0.036**
	(0.010)	(0.014)	(0.012)	(0.016)	(0.011)	(0.014)	(0.012)	(0.016)
School characteristics	0.035***	-0.017**	0.030***	0.043***	0.021**	-0.023***	0.012	0.020
	(0.008)	(0.008)	(0.011)	(0.012)	(0.009)	(0.008)	(0.012)	(0.013)
School value-added (lagged)	0.014*	-0.002	-0.008	0.004	0.021*	0.005	-0.005	0.008
	(0.008)	(0.009)	(0.008)	(0.009)	(0.011)	(0.011)	(0.010)	(0.011)
Unexplained gap change	-0.091***	-0.051**	-0.111***	0.021	-0.174***	-0.105***	-0.199***	-0.047
	(0.020)	(0.024)	(0.020)	(0.029)	(0.022)	(0.027)	(0.023)	(0.031)
N	924,345				881,399			

Notes: National Pupil Database, 2017 - 2020. ‘P&B’ stands for Pakistani and Bangladeshi, ‘I’ Indian, ‘BA’ Black African, ‘BC’ Black Caribbean. Sample comprises age 16 pupils in mainstream schools in England with no missing data. Exam grades are the May and June (end of year) examinations only. The gap is the gap in average grades between the ethnic minority group and White British pupils. Socio-economic status includes free school meals indicator and rank based on the proportion of all children aged 0 to 15 living in income deprived households in the pupil’s local area of residence (IDACI score). Prior attainment (age 11) includes subject-specific age 11 attainment score, standardised by year with mean zero and standard deviation of one. School characteristics are indicators for selectivity, urban, region (9 categories), and school governance type (4 categories), and continuous measures of cohort size, proportion of pupils eligible for free school meals, and average neighbourhood deprivation (IDACI) score of the school. School value-added refers to how many average grades higher or lower the pupils in that school achieve across eight qualifying GCSE subjects compared to pupils across the country who score comparatively at age 11. Standard errors in parentheses, clustered at the school level. * Significant at 10%; ** significant at 5%; *** significant at 1%.

B Appendix for Chapter 2

B.1 Data and sample construction

B.1.1 Datasets used and years covered

Table B.1 SAIL Databank datasets used and years covered

Dataset	**Years available**
Pupil (PLASC)	2003/04 – 2020/21
Attendance	2007/08 – 2018/19
Exclusions	2011/12 – 2019/20
Key Stage 1	2004/05 – 2010/11
Key Stage 2	2005/06 – 2018/19
Key Stage 4	2013/14 – 2018/19
Mergers	2002/03 – 2021/22
School	2003/04 – 2020/21
Pupil Teacher Ratio	2003/04 – 2020/21
Healthcare datasets	2003 - 2022

B.1.2 Data cleaning and variable construction

Sample selection

I begin with all pupils up to age 16 who are observed in the Pupil Level Annual School Census (January collection) during all available years, 2003 to 2021. This contains 4,818,708 observations, with each pupil observed for multiple years.

I retain pupils with school enrolment status “Current” or “Main” to remove duplicate entries by pupil and year, dropping 22,152 observations (0.5%). Within the Education Wales datasets, pupils are anonymously identified using a unique identifier (IRN). I drop observations for which the same IRN appears for multiple separate observations in the same year (20 observations, <0.001%), and for multiple separate observations for the same age pupil (1,416 observations, 0.03%). I also drop observations with the same IRN code but for which the age and year variables do not align throughout the period for which the pupil is observed (5,331 observations, 0.1%).

In order to link pupils in the Education Wales dataset with further demographic and health data, each pupil IRN code is matched to an ALF code using individuals' forename, surname, date of birth, sex, and postcode. I am able to link 100% of pupil IRN codes to a corresponding ALF code. Sometimes multiple IRN codes refer to the same ALF code. In such cases I assume that they correspond to the same pupil in all but two cases for which I recode the ALF code as missing: first, where the age and year variables do not align throughout the period for which the pupil is observed (626 observations, 0.01%); second, if the age of the pupils is the same, suggesting that two different people may have been linked to the same ALF code (1,714 observations, 0.04%). I merge the pupil-by-year dataset to data on attendance, exclusions, and academic assessments, by IRN, year, and unique school identifier (LAESTAB). I then link to school-level data by unique school identifier (LAESTAB) and year, and, where possible, to demographic and healthcare data by ALF code (99.8%). My fully-linked sample at the pupil level contains 443,859 individual pupils.

I retain pupils who are of school starting age (4) from the academic years 2001/02 and 2008/09 (291,811), and for whom I observe all relevant characteristics outlined in Table 2.1. I observe 248,125 (85.0%) complete cases for pupils at age 7 (Grade 2). 11.5% of pupils are dropped as they are not observed at all at age 7. A further 1.26% are missing WIMD, and a further 2.7% missing at least one demographic characteristic (sex, ethnicity, first language). I observe 238,436 (81.71%) complete cases for pupils at ages 7 (Grade 2) and 11 (Grade 6). 15.26% of pupils are dropped because they are not observed at both age 7 and age 11. A further 1.18% are missing WIMD, a further 1.1% are missing demographic characteristics, a further 0.99% are missing KS2 CSI, and a further 0.35% just attendance data. I observe 226,340 (77.56%) complete cases for pupils at ages 7 (Grade 2) and 16 (Grade 11). 20.04% of pupils are dropped because they are not observed at both age 7 and age 16. A further 1.13% are missing WIMD, a further 0.89% are missing demographic characteristics, a further 0.36% are missing KS4 Level 2, and a further 0.65% just attendance. I observe 224,891 (77.07%)

complete cases for pupils at ages 7 (Grade 2) and 16 and who are no longer of compulsory schooling age (Grade 12, or equivalent). In addition to the above sample, a further 1,449 (0.494%) pupils are dropped as they have no recorded WIMD at age 16/Grade 12, suggesting they may have moved out of Wales. For the ADHD outcome, as not all General Practices submit data, I restrict the sample to pupils who were registered with a contributing GP practice consistently from birth to age seven (47.43% of the existing age 7 sample).

Finally, I remove from the samples any pupils who attend a special school at any age (1.50%). This leaves the following final sample numbers. Age 7: 245,669 (84.19%). Age 11: 236,706 (81.12%). Age 16: 225,744 (77.36%). Age 16 including sixth form: 224,300 (76.86%). Age 7 including GP coverage: 141,098 (48.35%).

Variable construction

FSM refers to whether a pupil is recorded, by their school, as being in receipt of Free School Meals.

I code Pupil Special Educational Need (SEN) provision as a dichotomous variable, with codes “N” (“No special provision” as zero and “A” (“School Action”), “P” (“School Action Plus”), “Q” (“School Action Plus & Statutory Assessment”) and “S” (“Statemented”) as one.

Ethnicity data is coded using the valid ethnic background extended codes (see PLASC Technical Completion notes). I code ethnicity as missing if it is missing or recorded as “REFU” (information refused), or “NOBT” (information not obtained). I code as White British all pupils with recorded ethnicity “WBRI”. I code as Asian and Mixed White and Asian all pupils with recorded ethnicity beginning with A, beginning with C, “MWAS”, or “MWCH”. I code as Black and Mixed White and Black all pupils with recorded ethnicity beginning with B, “MWBA”, or “MWBC”. Remaining ethnicities are coded as Other Ethnic Groups. The male variable is coded as one if gender is recorded as “M” and zero if “F”.

I code the language variable using both EALACQUISITION and HOMEWELSH. First, I create a variable (“EAL”) which is equal to zero if EALACQUISITION is recorded as zero, and equal to one if EALAQUISITION is recorded as “A” (“New to English”), “B” (“Early acquisition”), “C” (“Developing competence”), “D” (“Competent”), or “E” (“Fluent”). I then code pupils’ language as English if EAL is equal to zero, HOMEWELSH is equal to zero (“Does not speak Welsh at home”), or

HOMEWELSH is equal to two (“Not applicable”). I code language as Non-Native if EAL is equal to one, and as Welsh if HOMEWELSH is equal to one (“Speaks Welsh at home”).

I code a pupil as attending sixth form if they appear in the Pupil Level Annual School Census at age 16, as the Pupil Level Annual School Census does not include Further Education colleges.

Where there are inconsistencies in the male, ethnicity or language variables within the same pupil across years, I make these consistent. For male, I replace any inconsistent entries with the sex as coded in the ALF table or, if unavailable, with the modal entry (or missing if two modes exist). For ethnicity, I replace any inconsistent entries with the modal entry or missing if two modes exist. For language, I replace any inconsistent entries with the modal entry or the maximum if two modes exist. This follows the logic that pupils will likely be either Welsh or Non-Native first-language speakers if they have been recorded as such as many times as they have been recorded as neither.

Attendance is a self-created measure, made by dividing the number of half day sessions that a pupil has attended by the number of half day sessions that a school is open. Absences, whether authorised or unauthorised, are not included in ‘attended’ sessions. I then standardise by year.

The exclusions variable measures unique fixed-term exclusions. Permanent exclusions (0.95%) are not included.

The KS2 Core Subject Indicator (CSI) identifies pupils who achieve the expected level (Level 4) or above, based on teacher assessments, in English or Welsh (first language), mathematics and science in combination at the end of Key Stage 2, aged 11. I code “N” as zero and “Y” as one. If pupils have multiple Key Stage 2 assessments at different ages, I retain the one completed at age 11 (dropping 771, 0.2%).

The KS4 Level 2 variable identifies pupils who achieve at least five GCSEs graded A* to C.

Over the period studied, a number of school merged, in particular Infant and Junior schools to create Primary schools. To optimise the number of schools that I can observe consistently throughout the treatment period, I treat these schools as one school. I therefore generate a new School ID variable which is the same for schools before and after merging. I generate this variable using the MERGERS data table. 13.5% of schools experienced a merger between 2004 and 2022.

Pupil-Teacher-Ratio (PTR) is used for descriptive purposes only. It is a self-created measure. From the PTR table I measure the number of adults assigned to different groups of children. There are some challenges to using this data: Nursery and Reception classes are categorised as Key Stage "F" and Grades 1 and 2 as Key Stage "1" until the national roll out of the Foundation Phase, when Key Stage "F" expands to include Grade 1 (in 2011) and Grade 2 (in 2012). I therefore keep observations with Key Stage "1" and grades "1", "2", or "M" (mixed), and drop the final cohort which would be impacted by the definition change. I count the number of pupils in grades "1" and "2" by school and year, and divide this number by the number of teachers assigned to any of the above categories. I collapse the data by school and year. I then drop any schools with any resultant ratios above the legal infant class size limit of 30 pupils (48%). My resultant sample contains 679 schools, 26 being Pilot schools – 7 in Pilot Group 1a, 4 in Pilot Group 1b, 15 in Pilot Group 2.

The Welsh Index of Multiple Deprivation (WIMD) variable measures the decile of local resource deprivation in the Lower Super Output Area (LSOA) of a pupils' address of residence using the 2014 index. More information on the measures contained in the index can be found in Welsh Government (2014). I use the Welsh Demographic Service Dataset to measure pupils' WIMD of residence between ages zero and 16.

The Attention Deficit Hyperactivity Disorder (ADHD) variable is generated using healthcare data from hospitals and General Practices (GPs) in Wales. I used Read V2 to identify diagnoses of ADHD in the General Practice dataset, and ICD-10 codes to identify diagnoses of ADHD in the Patient Episode and Outpatient datasets. The precise codes used are documented in Appendix B.1.3.

School Value-Added is made by regressing KS4 Level 2 on KS2 CSI and adjusting for sex, ethnicity, language, WIMD, FSM at age 15, PDG funding, and secondary school indicators, separately by year. For each year, I then extract the coefficients for each secondary school and standardise.

B.1.3 Read v2 and ICD-10 diagnosis codes used to identify records of ADHD in primary or secondary care

I use validated Read V.2 codes to identify children with primary care records for Attention Deficit Hyperactivity Disorder (ADHD) – including diagnoses, symptoms and prescriptions – and ICD-10 diagnostic codes to identify children with secondary care records for ADHD. The code lists are those used by John et al. (2022) and stored in the SAIL Databank Concept Library (C2708/9951 and C2931/9919).

Table B.2 Read v2 and ICD-10 codes for ADHD diagnoses

Read v2 codes for ADHD diagnosis

Read v2 code	**Description**
dw1I.	concerta xl 27mg m/r tablets
dc11.	*dexedrine 5mg tablets
dw42.	GUANFACINE 1mg m/r tablets
dw25.	strattera 60mg capsules
dw2y.	atomoxetine 18mg capsules
dw13.	*equasym 5mg tablets
E2E2.	hyperkinetic conduct disorder
dw2w.	atomoxetine 40mg capsules
dw1v.	methylphenidate 36mg m/r tabs | methylphenidate hydrochloride 36mg m/r tablets
dw11.	methylphenidate hcl 10mg tabs | methylphenidate hydrochloride 10mg tablets
E2E0.	child attention deficit disord | child attention deficit disorder
dw1u.	methylphenidate 30mg m/r caps | methylphenidate hydrochloride 30mg m/r capsules
dw21.	strattera 10mg capsules
dw1B.	*tranquilyn 20mg tablets
dw18.	concerta xl 36mg m/r tablets
dc1w.	dexamfetamine sulph 5mg tabs | dexamfetamine sulphate 5mg tablets
9Ngp.	on drug therapy for adhd (attention deficit hyperactivity disorder)
8BPT.	drug therapy for adhd (attention deficit hyperactivity disorder)
6A61.	adhd annual review | attention deficit hyperactivity disorder annual review
dw44.	GUANFACINE 2mg m/r tablets
Eu90z	[x]hyperkinetic disorder, unsp | [x]hyperkinetic disorder, unspecified
E2E01	attention deficit +hyperactive | attention deficit with hyperactivity
Eu900	[x]disturbance activity/attntn | [x]disturbance of activity and attention
dw1J.	medikinet 5mg tablets
dw3z.	lisdexamfetamine dim 30mg caps | lisdexamfetamine dimesylate 30mg capsules
dw1..	methylphenidate
dw2v.	atomoxetine 60mg capsules
dw1w.	methylphenidate 18mg m/r tabs | methylphenidate hydrochloride 18mg m/r tablets
dw1C.	equasym xl 10mg m/r capsules
9OlA.	attention deficit hyperactivity disorder monitoring invitation third letter
E2E..	childhood hyperkinetic syndr. | childhood hyperkinetic syndrome
Eu901	[x]hyperkinetic conduct disord | [x]hyperkinetic conduct disorder
dw3y.	lisdexamfetamine dim 50mg caps | lisdexamfetamine dimesylate 50mg capsules
dc14.	*durophet 20mg m/r capsules
dw12.	ritalin 10mg tablets
dw1q.	methylphenidate 5mg m/r caps | methylphenidate hydrochloride 5mg m/r capsules
dw41.	INTUNIV 1mg m/r tablets
dw1F.	medikinet xl 20mg m/r capsules
dw47.	INTUNIV 4mg m/r tablets

dw4..	Guanfacine
dw2z.	atomoxetine 10mg capsules
9Ol9.	attention deficit hyperactivity disorder monitoring invitation second letter
dw43.	INTUNIV 2mg m/r tablets
dw32.	elvanse 50mg capsules
dw1L.	medikinet 20mg tablets
E2E00	attention deficit-not hyperact | attention deficit without hyperactivity
9Ol8.	attention deficit hyperactivity disorder monitoring invitation first letter
dw1y.	methylphenidate hcl 5mg tabs | methylphenidate hydrochloride 5mg tablets
dw1D.	equasym xl 30mg m/r capsules
9Ngp0	on stimulant drug therapy for adhd (attention deficit hyperactivity disorder)
8BPT0	stimulant drug therapy for adhd (attention deficit hyperactivity disorder)
dw2u.	atomoxetine 80mg capsules
E2E0z	child attent.deficit dis.nos | child attention deficit disorder nos
E2Ey.	other hyperkinetic manifestat. | other hyperkinetic manifestation
dw17.	concerta xl 18mg m/r tablets
dw2x.	atomoxetine 25mg capsules
dw2..	atomoxetine
dw1r.	methylphenidate 27mg m/r tabs | methylphenidate hydrochloride 27mg m/r tablets
dc13.	*durophet 12.5mg m/r capsules
dw1M.	medikinet xl 5mg m/r capsules
dw1E.	medikinet xl 10mg m/r capsules
Eu90y	[x]oth hyperkinetic disorders | [x]other hyperkinetic disorders
dw1t.	methylphenidate 10mg m/r caps | methylphenidate hydrochloride 10mg m/r capsules
ZS91.	attention deficil disorder
dw19.	*tranquilyn 5mg tablets
dc1x.	*dexamphetamine 7.5mg m/r caps | dexamphetamine sulphate 7.5mg m/r capsules
9Ngp1	on non-stimulant drug therapy for adhd (attention deficit hyperactivity disorder)
E2E1.	hyperkinesis+development delay | hyperkinesis with developmental delay
dw46.	GUANFACINE 3mg m/r tablets
Eu90.	[x]hyperkinetic disorders
dw33.	elvanse 70mg capsules
dw16.	equasym xl 20mg m/r capsules
dw1A.	*tranquilyn 10mg tablets
Eu902	[x]def atten motor cont percep | [x]deficits in attention, motor control and perception
dw1z.	methylphenidate hcl 20mg tabs | methylphenidate hydrochloride 20mg tablets
dw1H.	medikinet xl 40mg m/r capsules
dw3x.	lisdexamfetamine dim 70mg caps | lisdexamfetamine dimesylate 70mg capsules
dw3..	lisdexamfetamine
dc1y.	*dexamphetamine 12.5mg caps | dexamphetamine sulphate 12.5mg m/r capsules
dw15.	*equasym 10mg tablets
dw45.	INTUNIV 3mg m/r tablets

dw26.	strattera 80mg capsules
dw1K.	medikinet 10mg tablets
dc1..	dexamfetamine sulphate
E2Ez.	hyperkinetic syndrome nos
dw22.	strattera 18mg capsules
dw1G.	medikinet xl 30mg m/r capsules
dw14.	*equasym 20mg tablets
dw24.	strattera 40mg capsules
dc1z.	*dexamphetamine 20mg m/r caps | dexamphetamine sulphate 20mg m/r capsules
dw23.	strattera 25mg capsules
Eu9y7	[x]attention deficit disorder
dw48.	GUANFACINE 4mg m/r tablets
dc12.	*durophet 7.5mg m/r capsules
dw1x.	methylphenidate 20mg m/r caps | methylphenidate hydrochloride 20mg m/r capsules
dw31.	elvanse 30mg capsules
8BPT1	non-stimulant drug therapy for adhd (attention deficit hyperactivity disorder)
dw1s.	methylphenidate 40mg m/r caps | methylphenidate hydrochloride 40mg m/r capsules

ICD-10 diagnosis codes for ADHD

ICD-10 code	**Description**
F90.1	hyperkinetic conduct disorder
F90.9	hyperkinetic disorder, unspecified
ICD10	DESCRIPTION
F90.0	disturbance of activity and attention
F90.2	attention-deficit hyperactivity disorder, combined type
F90.8	other hyperkinetic disorders
F90	hyperkinetic disorders

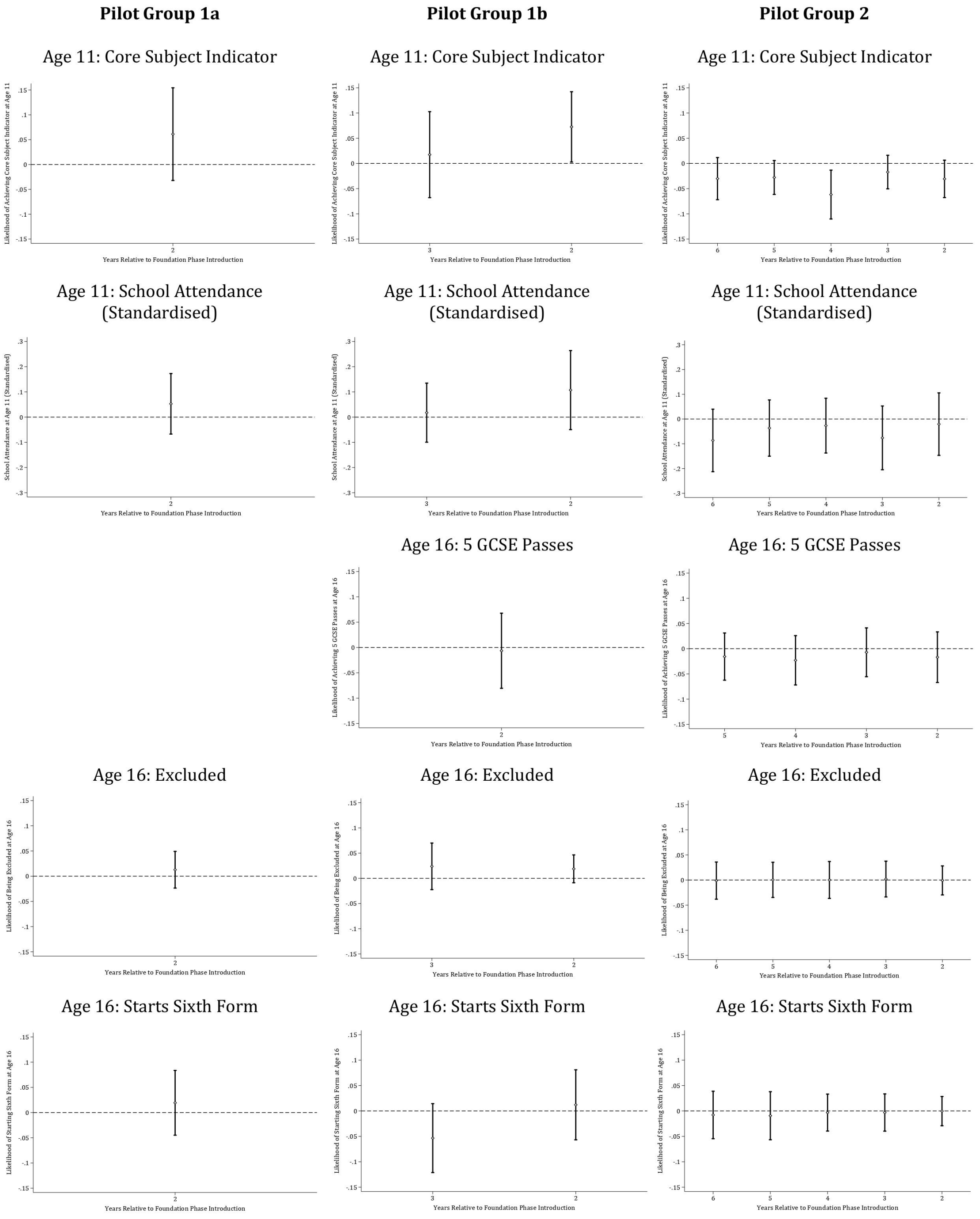
B.2 Supplementary tables and figures

Table B.3 Mobility rates for pupils during selected years in Primary school

	(1)	(2)	(3)
	Mobility rate ages 5 to 6	Mobility rate ages 5 to 6	Mobility rate ages 6 to 7
Pilot group	0.00430	0.00724	
	(0.00341)	(0.00457)	
Pilot group # Post		-0.00625	
		(0.00555)	
Primary merge			0.0243***
			(0.00247)
Constant	0.0451***	0.0451***	0.0355***
	(0.000863)	(0.000863)	(0.000815)
Mean	0.0453	0.0453	0.0412
N	210319	210319	241840

Notes: SAIL Databank data. Sample comprises pupils who attend mainstream schools in Wales and have no missing data. Table shows the results from regressions estimating the mobility rate of pupils between Primary schools at different ages. In columns (1) and (2) mobility is defined as a pupil's registered school in the January school census when they are five years old (Grade 1) being different to their registered school in the January school census when they are six years old (Grade 2), accounting for school mergers. In column (3) it is the same but for ages six and seven. The variable 'Primary merge' indicates schools which merge at some point during the years observed. The sample in columns 1 and 2 is a sub-sample of pupils who can be observed from age five. Standard errors in parentheses, clustered at the school level. * Significant at 10%; ** significant at 5%; *** significant at 1%.

Figure B.1 Pre-trends graphs by pilot group



Pilot Group 1a

Pilot Group 1b

Pilot Group 2

Age 11: Core Subject Indicator

Age 11: School Attendance (Standardised)

Age 16: 5 GCSE Passes

Age 16: Excluded

Age 16: Starts Sixth Form

Notes: SAIL Databank data. Sample comprises pupils who can be observed from age six, attend mainstream schools in Wales and have no missing data. Estimates from equation 2.2 on never-treated and pre-treated observations of pupils in treated schools only. Additional controls in underlying regression: gender, ethnicity, language, Welsh Index of Multiple Deprivation at age six, Free School Meal eligibility at the age of measurement, Free School Meal eligibility at the age of measurement interacted with Pupil Deprivation Grant funding, Primary school fixed effects, year fixed effects. Standard errors clustered at the school level.

Table B.4 Tests for parallel pre-trends by pilot group

	(1) Age 11: Core Subject Indicator	(2) Age 11: School Attendance (Standardised)	(3) Age 16: 5 GCSE Passes	(4) Age 16: Excluded	(5) Age 16: Starts Sixth Form
Pilot group 1a					
2 year horizon					
F	1.657	0.741		0.478	0.349
P(F)	0.198	0.390		0.489	0.555
N	225832	225832	159462	189277	188049
Pilot Group 1b					
2 year horizon					
F	4.163	1.798	0.029	1.781	0.118
P(F)	0.042	0.180	0.865	0.182	0.731
3 year horizon					
F	2.089	1.444		1.095	6.901
P(F)	0.124	0.236		0.335	0.001
N	226090	226090	159712	189512	188282
Pilot Group 2					
2 year horizon					
F	2.628	0.102	0.432	0.002	0.000
P(F)	0.105	0.750	0.511	0.962	0.982
3 year horizon					
F	1.352	0.735	0.216	0.024	0.015
P(F)	0.259	0.480	0.806	0.976	0.986
4 year horizon					
F	2.303	0.501	0.466	0.016	0.016
P(F)	0.075	0.682	0.706	0.997	0.997
5 year horizon					
F	1.806	0.396	0.367	0.015	0.046
P(F)	0.125	0.811	0.832	1.000	0.996
6 year horizon					
F	1.445	0.475		0.020	0.039
P(F)	0.205	0.795		1.000	0.999
N	230691	230691	163377	193839	192588

Notes: SAIL Databank data. Sample comprises pupils who can be observed from age six, attend mainstream schools in Wales and have no missing data. Pilot schools are split into Groups 1a, 1b and 2, which each implement the Foundation Phase in different years. Estimates are from equation 2.2 on never-treated and pre-treated observations of pupils in treated schools only. Additional controls in underlying regression: gender, ethnicity, language, Welsh Index of Multiple Deprivation at age six, Free School Meal eligibility at the age of measurement, Free School Meal eligibility at the age of measurement interacted with Pupil Deprivation Grant funding, Primary school fixed effects, year fixed effects. Standard errors clustered at the school level.

Table B.5 Main results using TWFE estimator

	(1) Age 11: Core Subject Indicator	(2) Age 11: School Attendance (Standardised)	(3) Age 16: 5 GCSE Passes	(4) Age 16: Excluded	(5) Age 16: Starts Sixth Form
Original estimator: BJS					
Treatment Effect (ATT)	0.014	-0.019	-0.014	0.002	-0.017
	(0.012)	(0.031)	(0.015)	(0.006)	(0.018)
Mean	0.831	0.020	0.690	0.037	0.408
Proportion treated	0.048	0.048	0.048	0.048	0.048
N	236706	236706	167190	198211	196934
Additional estimator: TWFE					
Treatment Effect (ATT)	0.022	-0.010	-0.004	0.001	-0.017
	(0.012)	(0.032)	(0.015)	(0.006)	(0.014)
Mean	0.831	0.020	0.690	0.037	0.408
Proportion treated	0.048	0.048	0.048	0.048	0.048
N	236706	236706	167190	198211	196934

Notes: SAIL Databank data. Sample comprises pupils who can be observed from age six, attend mainstream schools in Wales and have no missing data. Original estimates are from equation (2.1) using the BJS (Borusyak, Jaravel and Spiess) estimator outlined in Borusyak et al. (2024). Additional estimates are from equation (2.1) using a TWFE (Two Way Fixed Effects) estimator. Controls: gender, ethnicity, language, Welsh Index of Multiple Deprivation at age six, Free School Meal eligibility at the age of measurement, Free School Meal eligibility at the age of measurement interacted with Pupil Deprivation Grant funding, Primary school fixed effects, year fixed effects. Standard errors in parentheses, clustered at the school level. * Significant at 10%; ** significant at 5%; *** significant at 1%.

Table B.6 Tests for compositional change by sample and pupil characteristic

	Age 6	Age 11	Age 16	Age 16 with sixth form	Age 7 with GP data
Sex					
Male	0.017	0.016	0.015	0.015	0.012
	(0.013)	(0.013)	(0.013)	(0.013)	(0.017)
Ethnicity					
White British	-0.010	-0.011	-0.010	-0.010	-0.009
	(0.010)	(0.009)	(0.009)	(0.009)	(0.010)
Asian and Mixed White and Asian	0.003	0.006	0.008	0.009	0.006
	(0.004)	(0.007)	(0.008)	(0.008)	(0.009)
Black and Mixed White and Black	-0.009*	-0.008*	-0.010*	-0.010*	-0.004
	(0.005)	(0.004)	(0.006)	(0.006)	(0.003)
Other Ethnic Groups	0.016	0.013	0.011	0.011	0.007
	(0.011)	(0.008)	(0.008)	(0.008)	(0.005)
First language					
English	-0.005	-0.007	-0.003	-0.004	-0.002
	(0.016)	(0.015)	(0.015)	(0.015)	(0.012)
Welsh	-0.008	-0.008	-0.009	-0.009	-0.007
	(0.012)	(0.011)	(0.011)	(0.011)	(0.009)
Non-Native	0.013	0.014*	0.013	0.013	0.009
	(0.009)	(0.008)	(0.008)	(0.008)	(0.008)
Free School Meals at age six	-0.017	-0.014	-0.015	-0.016	-0.002
	(0.019)	(0.017)	(0.016)	(0.017)	(0.015)
WIMD decile at age six					
(Least deprived) 1	-0.000	-0.001	-0.001	-0.001	-0.005
	(0.004)	(0.004)	(0.004)	(0.004)	(0.007)
2	-0.003	-0.003	-0.003	-0.002	-0.007
	(0.004)	(0.004)	(0.004)	(0.004)	(0.006)
3	-0.005	-0.006	-0.007	-0.007	-0.012
	(0.006)	(0.006)	(0.006)	(0.006)	(0.007)
4	0.008	0.009	0.010	0.010	0.007
	(0.007)	(0.007)	(0.007)	(0.007)	(0.006)
5	0.008*	0.007*	0.008*	0.008*	0.010*
	(0.004)	(0.004)	(0.004)	(0.004)	(0.006)
6	-0.017**	-0.016**	-0.017**	-0.016**	-0.013*
	(0.006)	(0.007)	(0.006)	(0.006)	(0.007)
7	-0.003	-0.003	-0.002	-0.003	0.001
	(0.004)	(0.004)	(0.004)	(0.004)	(0.010)
8	0.007	0.010**	0.010*	0.010*	0.005
	(0.005)	(0.005)	(0.005)	(0.005)	(0.007)
9	0.017**	0.014*	0.015**	0.016**	0.012
	(0.007)	(0.007)	(0.007)	(0.007)	(0.009)
(Most deprived) 10	-0.012*	-0.011*	-0.012**	-0.013**	0.003
	(0.006)	(0.006)	(0.005)	(0.006)	(0.008)
N	245699	236706	225744	224300	141098

Notes: SAIL Databank data. Sample comprises pupils who can be observed from age six, attend mainstream schools in Wales and have no missing data. Estimates from equation (2.1) using the estimator proposed by Borusyak et al. (2024) but using each characteristic as the outcome variable and only Primary school fixed effects and year fixed effects as controls. Column headers refer to different samples used for analyses. WIMD stands for the Welsh Index of Multiple Deprivation of the pupil's home address, 2014 index, which combines data on eight domains of income, employment, health, education, access to services, community safety, physical environment, and housing. Standard errors in parentheses, clustered at the school level. * Significant at 10%; ** significant at 5%; *** significant at 1%.

Table B.7 Non-attrition rates by sample

	(1) Pilot Group	(2) Rest of Wales	(3) Test
Age 11			
Non-attritted	0.960	0.964	-0.004**
	(0.187)	(0.196)	
N	11443	225263	236706
N of which treated	4619	0	4619
Schools	42	1344	1388
Age 16			
Non-attritted	0.904	0.920	-0.016***
	(0.272)	(0.295)	
N	10770	214974	225744
N of which treated	4314	0	4314
Schools	42	1344	1388
Age 16 with sixthform			
Non-attritted	0.898	0.914	-0.015***
	(0.281)	(0.302)	
N	10708	213592	224300
N of which treated	4286	0	4286
Schools	42	1344	1388
Age 7 with GP data			
Non-attritted	0.560	0.575	-0.015***
	(0.494)	(0.496)	
N	6675	134423	141098
N of which treated	2996	0	2996
Schools	42	1302	1346

Notes: SAIL Databank data. Sample comprises pupils who can be observed from age six, attend mainstream schools in Wales and have no missing data. Columns 1 and 2 report the proportion of observations which were in the baseline sample at age six and are still observed at the given age. Standard errors in parentheses, clustered at the school level. * Significant at 10%; ** significant at 5%; *** significant at 1%.

Table B.8 Tests for parallel pre-trends and ATT estimates for sample excluding Flying Start cohort

	(1)	(2)	(3)	(4)	(5)
	Age 11: Core Subject Indicator	Age 11: School Attendance (Standardised)	Age 16: 5 GCSE Passes	Age 16: Excluded	Age 16: Starts Sixth Form
Panel A: Pre-Trends Tests					
2 year horizon					
F	0.207	0.203	0.443	0.335	0.228
P(F)	0.649	0.653	0.506	0.563	0.633
3 year horizon					
F	0.159	1.042	0.222	0.202	0.813
P(F)	0.853	0.353	0.801	0.817	0.444
4 year horizon					
F	1.503	0.697	0.463	0.147	0.565
P(F)	0.212	0.554	0.708	0.932	0.638
5 year horizon					
F	1.144	0.567	0.363	0.118	0.451
P(F)	0.334	0.686	0.835	0.976	0.772
6 year horizon					
F	0.964	0.601		0.106	0.377
P(F)	0.438	0.699		0.991	0.865
N	204163	204163	164159	195180	193921
Panel B: ATT Estimates					
Treatment Effect (ATT)	0.006	-0.019	-0.014	0.002	-0.017
	(0.012)	(0.032)	(0.015)	(0.006)	(0.018)
Mean	0.821	0.021	0.690	0.037	0.408
Proportion treated	0.049	0.049	0.048	0.048	0.048
N	207408	207408	167190	198211	196934

Notes: SAIL Databank data. Sample comprises pupils who can be observed from age six, attend mainstream schools in Wales and have no missing data. Sample also excludes the final cohort. Panel 1 estimates are from equation (2.2) on never-treated and pre-treated observations of pupils in treated schools only. Panel two estimates from equation (2.1) using the estimator proposed by Borusyak et al. (2024). Additional controls in underlying regression: gender, ethnicity, language, Welsh Index of Multiple Deprivation at age six, Free School Meal eligibility at the age of measurement, Free School Meal eligibility at the age of measurement interacted with Pupil Deprivation Grant funding, Primary school fixed effects, year fixed effects. Standard errors clustered at the school level. * Significant at 10%; ** significant at 5%; *** significant at 1%.

Table B.9 Tests for parallel pre-trends and ATT estimates for combined Pilot Group 1a and Pilot Group 1b

	(1) Age 11: Core Subject Indicator	(2) Age 11: School Attendance (Standardised)	(3) Age 16: 5 GCSE Passes	(4) Age 16: Excluded	(5) Age 16: Starts Sixth Form
Panel A: Pre-Trends Tests					
2 year horizon					
F	0.617	0.254	0.336	0.018	0.045
P(F)	0.432	0.614	0.562	0.892	0.832
3 year horizon					
F	0.309	0.445	0.168	0.149	0.510
P(F)	0.734	0.641	0.845	0.862	0.601
4 year horizon					
F	1.658	0.297	0.418	0.100	0.356
P(F)	0.174	0.828	0.740	0.960	0.785
5 year horizon					
F	1.246	0.258	0.327	0.083	0.290
P(F)	0.289	0.905	0.860	0.988	0.885
6 year horizon					
F	1.010	0.445		0.078	0.232
P(F)	0.410	0.817		0.996	0.949
N	232356	232356	164422	195443	194183
Panel B: ATT Estimates					
Treatment Effect (ATT)	0.013	-0.004	-0.013	-0.001	-0.022
	(0.012)	(0.030)	(0.015)	(0.006)	(0.017)
Mean	0.831	0.020	0.690	0.037	0.408
Proportion treated	0.048	0.048	0.048	0.048	0.048
N	236706	236706	167190	198211	196934

Notes: SAIL Databank data. Sample comprises pupils who can be observed from age six, attend mainstream schools in Wales and have no missing data. Pilot Groups 1a and 1b combined and assumed to introduce the Foundation Phase at the same time as Pilot Group 1b. Panel 1 estimates are from equation (2.2) on never-treated and pre-treated observations of pupils in treated schools only. Panel two estimates from equation (2.1) using the estimator proposed by Borusyak et al. (2024). Additional controls in underlying regression: gender, ethnicity, language, Welsh Index of Multiple Deprivation at age six, Free School Meal eligibility at the age of measurement, Free School Meal eligibility at the age of measurement interacted with Pupil Deprivation Grant funding, Primary school fixed effects, year fixed effects. Standard errors clustered at the school level. * Significant at 10%; ** significant at 5%; *** significant at 1%.

Table B.10 Tests for parallel pre-trends and ATT estimates for sample treated ages five to seven

	(1)	(2)	(3)	(4)	(5)
	Age 11: Core Subject Indicator	Age 11: School Attendance (Standardised)	Age 16: 5 GCSE Passes	Age 16: Excluded	Age 16: Starts Sixth Form
Panel A: Pre-Trends Tests					
2 year horizon					
F	0.002	0.203	0.482	0.101	0.109
P(F)	0.968	0.652	0.488	0.751	0.742
3 year horizon					
F	0.263	1.212	0.289	0.083	0.058
P(F)	0.769	0.298	0.749	0.920	0.943
4 year horizon					
F	1.730	0.832	0.874	0.069	0.057
P(F)	0.159	0.476	0.454	0.976	0.982
5 year horizon					
F	1.304	0.624	0.665	0.054	0.054
P(F)	0.266	0.646	0.617	0.995	0.994
6 year horizon					
F	1.304	0.624		0.054	0.054
P(F)	0.266	0.646		0.995	0.994
N	189663	189663	155513	155513	154522
Panel B: ATT Estimates					
Treatment Effect (ATT)	0.015	-0.018	-0.015	0.007	-0.021
	(0.011)	(0.034)	(0.015)	(0.005)	(0.022)
Mean	0.845	0.032	0.696	0.035	0.407
Proportion treated	0.048	0.048	0.048	0.048	0.048
N	194034	194034	158405	158405	157398

Notes: SAIL Databank data. Sample comprises pupils who can be observed from age six, attend mainstream schools in Wales and have no missing data. Sample also restricted to pupils who can be observed at age five and do not move school between ages five and six. Panel 1 estimates are from equation (2.2) on never-treated and pre-treated observations of pupils in treated schools only. Panel two estimates from equation (2.1) using the estimator proposed by Borusyak et al. (2024). Additional controls in underlying regression: gender, ethnicity, language, Welsh Index of Multiple Deprivation at age six, Free School Meal eligibility at the age of measurement, Free School Meal eligibility at the age of measurement interacted with Pupil Deprivation Grant funding, Primary school fixed effects, year fixed effects. Standard errors clustered at the school level. * Significant at 10%; ** significant at 5%; *** significant at 1%.

Table B.11 Tests for parallel pre-trends and ATT estimates for balanced sample of Primary schools

	(1)	(2)	(3)	(4)	(5)
	Age 11: Core Subject Indicator	Age 11: School Attendance (Standardised)	Age 16: 5 GCSE Passes	Age 16: Excluded	Age 16: Starts Sixth Form
Panel A: Pre-Trends Tests					
2 year horizon					
F	0.260	0.348	0.203	0.178	0.147
P(F)	0.610	0.555	0.652	0.673	0.701
3 year horizon					
F	0.589	1.037	0.102	0.123	0.980
P(F)	0.555	0.355	0.903	0.884	0.376
4 year horizon					
F	1.735	0.698	0.369	0.092	0.681
P(F)	0.158	0.553	0.776	0.964	0.564
5 year horizon					
F	1.312	0.580	0.291	0.074	0.532
P(F)	0.263	0.678	0.884	0.990	0.712
6 year horizon					
F	1.126	0.615		0.073	0.441
P(F)	0.344	0.689		0.996	0.820
N	227614	227614	161138	191354	190132
Panel B: ATT Estimates					
Treatment Effect (ATT)	0.014	-0.017	-0.021	0.001	-0.020
	(0.012)	(0.032)	(0.014)	(0.006)	(0.019)
Mean	0.832	0.021	0.691	0.037	0.409
Proportion treated	0.048	0.048	0.048	0.048	0.048
N	232126	232126	164061	194277	193038

Notes: SAIL Databank data. Sample comprises pupils who can be observed from age six, attend mainstream schools in Wales and have no missing data. Sample also restricted to balanced Primary schools. Panel 1 estimates are from equation (2.2) on never-treated and pre-treated observations of pupils in treated schools only. Panel two estimates from equation (2.1) using the estimator proposed by Borusyak et al. (2024). Additional controls in underlying regression: gender, ethnicity, language, Welsh Index of Multiple Deprivation at age six, Free School Meal eligibility at the age of measurement, Free School Meal eligibility at the age of measurement interacted with Pupil Deprivation Grant funding, Primary school fixed effects, year fixed effects. Standard errors clustered at the school level. * Significant at 10%; ** significant at 5%; *** significant at 1%.

Table B.12 Tests for parallel pre-trends and ATT estimates for sample excluding merger schools

	(1)	(2)	(3)	(4)	(5)
	Age 11: Core Subject Indicator	Age 11: School Attendance (Standardised)	Age 16: 5 GCSE Passes	Age 16: Excluded	Age 16: Starts Sixth Form
Panel A: Pre-Trends Tests					
2 year horizon					
F	1.283	0.275	1.168	0.424	0.270
P(F)	0.258	0.600	0.280	0.515	0.603
3 year horizon					
F	0.707	2.022	1.481	0.365	0.191
P(F)	0.493	0.133	0.228	0.695	0.826
4 year horizon					
F	3.094	1.517	1.010	0.244	0.155
P(F)	0.026	0.208	0.388	0.866	0.927
5 year horizon					
F	2.537	1.183	1.061	0.183	0.396
P(F)	0.039	0.317	0.375	0.947	0.812
6 year horizon					
F	2.707	0.947		0.268	0.346
P(F)	0.019	0.449		0.931	0.885
N	176780	176780	124131	147235	146274
Panel B: ATT Estimates					
Treatment Effect (ATT)	0.012	-0.021	-0.025	0.001	-0.023
	(0.013)	(0.038)	(0.019)	(0.008)	(0.021)
Mean	0.838	0.041	0.706	0.035	0.428
Proportion treated	0.045	0.045	0.045	0.045	0.045
N	180386	180386	126478	149582	148604

Notes: SAIL Databank data. Sample comprises pupils who can be observed from age six, attend mainstream schools in Wales and have no missing data. Sample also restricted to Primary schools which do not experience school mergers during the years observed. Panel 1 estimates are from equation (2.2) on never-treated and pre-treated observations of pupils in treated schools only. Panel two estimates from equation (2.1) using the estimator proposed by Borusyak et al. (2024). Additional controls in underlying regression: gender, ethnicity, language, Welsh Index of Multiple Deprivation at age six, Free School Meal eligibility at the age of measurement, Free School Meal eligibility at the age of measurement interacted with Pupil Deprivation Grant funding, Primary school fixed effects, year fixed effects. Standard errors clustered at the school level. * Significant at 10%; ** significant at 5%; *** significant at 1%.

Table B.13 Tests for parallel pre-trends and ATT estimates for additional outcomes

	(1)	(2)	(3)	(4)
	Age 7: Special Educational Need	Age 7: ADHD Diagnosis	Age 7: ADHD Diagnosis (Excluding OPDW)	Age 16: School Value-Added (Standardised)
Panel A: Pre-Trends Tests				
2 year horizon				
F	0.091	1.323	1.323	0.009
P(F)	0.762	0.250	0.250	0.925
3 year horizon				
F	0.078	1.019	1.019	0.006
P(F)	0.925	0.361	0.361	0.994
4 year horizon				
F	0.090	0.716	0.716	0.196
P(F)	0.965	0.542	0.542	0.899
5 year horizon				
F	0.143	1.063	1.063	0.147
P(F)	0.966	0.374	0.374	0.964
6 year horizon				
F	0.470	4.542	4.532	0.147
P(F)	0.799	0.000	0.000	0.964
N	240797	138102	138102	164159
Panel B: ATT Estimates				
Treatment Effect (ATT)	0.005	-0.000	-0.000	0.124
	(0.020)	(0.002)	(0.002)	(0.118)
Mean	0.270	0.006	0.006	-0.000
Proportion treated	0.049	0.047	0.047	0.048
N	245669	141098	141098	167190

Notes: SAIL Databank data. Sample comprises pupils who can be observed from age six, attend mainstream schools in Wales and have no missing data. ADHD stands for Attention Deficit Hyperactivity Disorder. OPDW stands for the Outpatient Database for Wales, excluded as a sensitivity check because diagnosis codes in these data are less reliable than the other healthcare datasets. School Value-Added is a self generated measure (see Appendix B.1.2). Panel 1 estimates are from equation (2.2) on never-treated and pre-treated observations of pupils in treated schools only. Panel two estimates from equation (2.1) using the estimator proposed by Borusyak et al. (2024). Additional controls in underlying regression: gender, ethnicity, language, Welsh Index of Multiple Deprivation at age six, Free School Meal eligibility at the age of measurement, Free School Meal eligibility at the age of measurement (FSM) interacted with Pupil Deprivation Grant funding, Primary school fixed effects, year fixed effects. Standard errors clustered at the school level. * Significant at 10%; ** significant at 5%; *** significant at 1%.

C

Appendix for Chapter 3

Table C.1 Categorisation of reason for exclusion

Category	Reason for exclusion
Disruptive behaviour	DB = Persistent disruptive behaviour
Physical	PP = Physical assault against a pupil
	PA = Physical assault against an adult
Verbal	VP = Verbal abuse/threatening behaviour against a pupil
	VA = Verbal abuse/threatening behaviour against an adult
	BU = Bullying
	RA = Racist abuse
Illicit	SM = Sexual misconduct
	DA = Drug and alcohol related
	DM = Damage
	TH = Theft
Other	OT = Other

Table C.2 Mean exclusion and absence sessions per year by pupil characteristics

	Exclusion sessions		Absence sessions	
	Mean	Standard deviation	Mean	Standard deviation
Female	0.122	1.590	3.019	11.437
Male	0.337	2.713	2.974	11.386
Ever-Special Educational Need	0.524	3.455	4.723	15.860
Never-Special Educational Need	0.085	1.211	2.134	8.214
Language English	0.247	2.330	2.945	11.793
Language Not English	0.146	1.610	3.251	9.052
Ethnicity:				
White	0.237	2.294	2.864	11.783
Pakistani & Bangladeshi	0.155	1.617	3.645	9.139
Indian	0.058	0.957	2.069	6.669
Asian Other	0.118	1.465	2.517	8.829
Black African	0.256	2.158	2.162	7.140
Black Caribbean	0.524	3.361	4.429	13.239
Black Other	0.349	2.615	3.223	10.027
Chinese	0.030	0.686	1.287	5.124
Gypsy, Roma and Traveller	0.815	4.455	17.282	28.368
Other	0.173	1.838	3.541	10.584
Ever-Free School Meals	0.477	3.293	5.951	17.256
Never-Free School Meals	0.136	1.644	1.855	7.800
IDACI quintile 1 (least deprived)	0.098	1.361	1.309	6.373
IDACI quintile 2	0.146	1.702	1.876	8.315
IDACI quintile 3	0.216	2.109	2.768	10.697
IDACI quintile 4	0.302	2.582	3.917	13.110
IDACI quintile 5 (most deprived)	0.394	3.008	5.112	15.676
Prior attainment: Low	0.387	2.894	4.192	13.936
Prior attainment: Medium	0.180	1.872	2.468	9.626
Prior attainment: High	0.071	1.112	1.368	6.264
N	73,952,708			

Notes: National Pupil Database, 2007 - 2019. Unrestricted sample, all grades. Exclusions are temporary exclusions only. Absences are unauthorised absences only. Sample size: 73,952,708. IDACI stands for the Income Deprivation affecting Children Index, and is neighbourhood-based measure of the proportion of all children aged 0 to 15 living in income deprived families. Prior attainment terciles are derived using average age 11 test scores in English and maths, standardised by year.

Table C.3 Sample cohort distribution

	Younger sibling			Older sibling		
Year	Grade 8	Grade 9	Grade 10	Grade 8	Grade 9	Grade 10
2010	0	0	0	104106	0	0
2011	0	0	0	71,530	76047	0
2012	13,987	0	0	37,513	57968	63186
2013	52,242	13195	0	88,084	59210	86973
2014	59,976	42450	12278	84,474	80252	57891
2015	69,160	57680	51333	63,437	62124	53598
2016	80,731	71512	66462	33403	34711	28227
2017	98,001	81001	75238	0	0	0
2018	108,450	104474	84564	0	0	0
N	1,142,734					

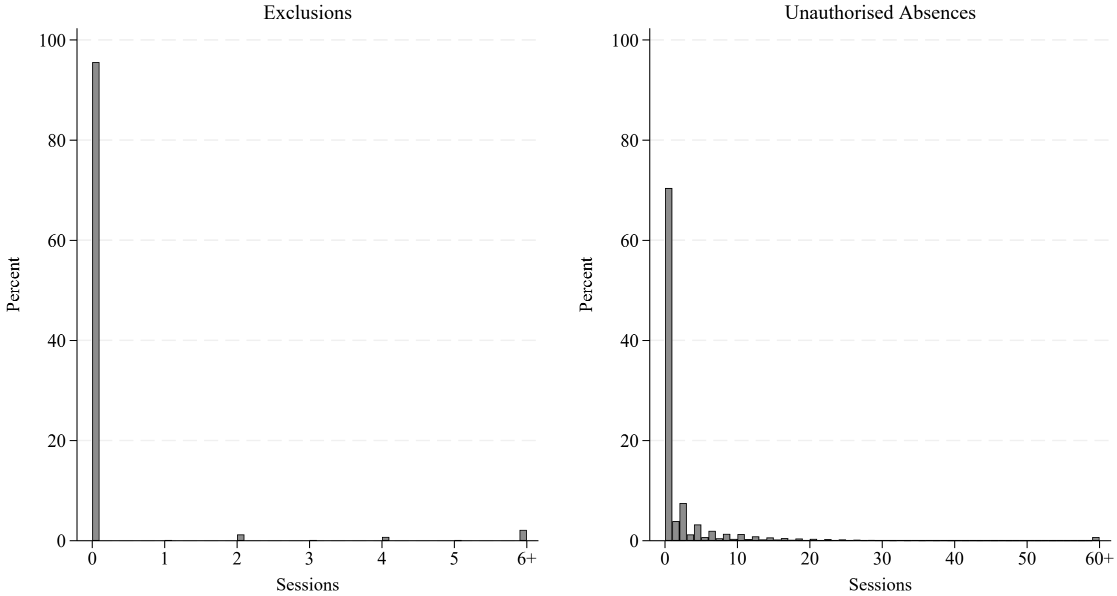
Notes: National Pupil Database, 2007 - 2019. Sample is restricted to the eldest two siblings per family and year, attending a mainstream Secondary school, with no missing covariates, with at least a two year grade gap, and observed in grades 8, 9 or 10.

Table C.4 Sample means by grade and sibling type

	Combined	Grade 8	Grade 9	Grade 10
Panel A: Exclusion (sessions)				
Younger sibling	0.376	0.318	0.404	0.437
	(2.76)	(2.60)	(2.87)	(2.87)
Older sibling	0.257	0.209	0.278	0.310
	(2.11)	(1.94)	(2.22)	(2.25)
Panel B: Absence (sessions)				
Younger sibling	3.14	2.61	3.26	3.87
	(12.33)	(10.34)	(12.59)	(14.76)
Older sibling	2.26	1.93	2.33	2.74
	(9.30)	(7.58)	(9.44)	(11.46)

Notes: National Pupil Database, 2007 - 2019. Sample is restricted to the eldest two siblings per family and year, attending a mainstream Secondary school, with no missing covariates, with at least a two year grade gap, and observed in grades 8, 9 or 10. Exclusions are temporary exclusions only. Absences are unauthorised absences only. One session is half a school day.

Figure C.1 Distribution of outcome variables (younger siblings only)



Notes: National Pupil Database, 2007 - 2019. Sample is restricted to the eldest two siblings per family and year, attending a mainstream Secondary school, with no missing covariates, with at least a two year grade gap, and observed in grades 8, 9 or 10. Exclusions are temporary exclusions only. Absences are unauthorised absences only. One session is half a school day. Sample size: 1,142,734.

Table C.5 Sibling spillover effects: Main estimates by grade

	Outcome: Exclusion (sessions)					
	Grade 8		**Grade 9**		**Grade 10**	
	OLS	2SLS	OLS	2SLS	OLS	2SLS
	(1)	(2)	(3)	(4)	(5)	(6)
Spillover effect	0.099***	0.087	0.108***	0.016	0.079***	-0.017
	(0.009)	(0.107)	(0.013)	(0.097)	(0.008)	(0.088)
First-stage coefficient		0.164***		0.190***		0.190***
		(0.021)		(0.021)		(0.022)
F-test first stage		58.87		79.55		73.60
Stock-Yogo Critical Value		16.38		16.38		16.38
Endogeneity test		0.014		0.934		1.14
Endogeneity test p-value		0.907		0.334		0.285
Year fixed effects	✓	✓	✓	✓	✓	✓
Grade fixed effects	✓	✓	✓	✓	✓	✓
Individual and family controls	✓	✓	✓	✓	✓	✓
School peer behaviour controls	✓	✓	✓	✓	✓	✓
Observations	482,547		370,312		289,875	
	Outcome: Absence (sessions)					
	Grade 8		**Grade 9**		**Grade 10**	
	OLS	2SLS	OLS	2SLS	OLS	2SLS
	(7)	(8)	(9)	(10)	(11)	(12)
Spillover effect	0.313***	0.274***	0.316***	0.181	0.329***	0.161
	(0.009)	(0.099)	(0.010)	(0.143)	(0.011)	(0.162)
First-stage coefficient		0.125***		0.097***		0.090***
		(0.011)		(0.012)		(0.015)
F-test first stage		121.26		63.20		36.49
Stock-Yogo Critical Value		16.38		16.38		16.38
Endogeneity test		0.151		0.911		1.07
Endogeneity test p-value		0.698		0.340		0.302
Year fixed effects	✓	✓	✓	✓	✓	✓
Grade fixed effects	✓	✓	✓	✓	✓	✓
Individual and family controls	✓	✓	✓	✓	✓	✓
School peer behaviour controls	✓	✓	✓	✓	✓	✓
Observations	482,547		370,312		289,875	

Notes: National Pupil Database, 2007 - 2019. Sample is restricted to the eldest two siblings per family and year, attending a mainstream Secondary school, with no missing covariates, with at least a two year grade gap, and observed in grades 8, 9 or 10. Exclusions are temporary exclusions only. Absences are unauthorised absences only. One session is half a school day. The individual and family controls are prior attainment at age 11 (standardised), Free School Meals eligibility, whether first language is English, sibship sex combination, and sibship age gap. The school peer behaviour controls are three-year rolling averages of homogeneous peers' misbehaviour for both siblings, and the younger sibling's own homogeneous peers' average misbehaviour. The instrumental variable is the average misbehaviour of the older sibling's homogeneous peers. All peer measures are constructed using take-one-out means. * Significant at 10%; ** significant at 5%; *** significant at 1%.

Table C.6 First stage estimates by older sibling prior attainment and gender groups

	Outcome: Exclusion (sessions)					
Prior attainment group:	Low		Medium		High	
Gender:	Female	Male	Female	Male	Female	Male
	(1)	(2)	(3)	(4)	(5)	(6)
First-stage coefficient	0.197***	0.201***	0.122***	0.200***	0.066***	0.137***
	(0.032)	(0.021)	(0.024)	(0.032)	(0.020)	(0.023)
F-test first stage	38.76	38.63	34.36	26.68	10.79	34.36
Stock-Yogo Critical Value (10%)	16.38	16.38	16.38	16.38	16.38	16.38
Observations	167,996	178,063	189,201	192,691	207,301	207,482
	Outcome: Absence (sessions)					
Prior attainment group:	Low		Medium		High	
Gender:	Female	Male	Female	Male	Female	Male
	(7)	(8)	(9)	(10)	(11)	(12)
First-stage coefficient	0.093***	0.128***	0.089***	0.114***	0.077***	0.092***
	(0.017)	(0.021)	(0.019)	(0.019)	(0.016)	(0.022)
F-test first stage	38.41	31.86	22.62	34.32	22.54	17.02
Stock-Yogo Critical Value (10%)	16.38	16.38	16.38	16.38	16.38	16.38
Observations	167,996	178,063	189,201	192,691	207,301	207,482

Notes: National Pupil Database, 2007 - 2019. Sample is restricted to the eldest two siblings per family and year, attending a mainstream Secondary school, with no missing covariates, with at least a two year grade gap, and observed in grades 8, 9 or 10. Exclusions are temporary exclusions only. Absences are unauthorised absences only. One session is half a school day. The individual and family controls are prior attainment at age 11 (standardised), Free School Meals eligibility, whether first language is English, sibship sex combination, and sibship age gap. The school peer behaviour controls are three-year rolling averages of homogeneous peers' misbehaviour for both siblings, and the younger sibling's own homogeneous peers' average misbehaviour. The instrumental variable is the average misbehaviour of the older sibling's homogeneous peers. All peer measures are constructed using take-one-out means. * Significant at 10%; ** significant at 5%; *** significant at 1%.

Table C.7 Main estimates: Poisson models

	Outcome: Exclusion (sessions)					
	OLS					2SLS
	(1)	(2)	(3)	(4)	(5)	(6)
Spillover effect	0.023***	0.023***	0.023***	0.016***	0.010***	-0.000***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)
Year fixed effects		✓	✓	✓	✓	✓
Grade fixed effects			✓	✓	✓	✓
Individual and family controls				✓	✓	✓
School peer behaviour controls					✓	✓
Observations	1,142,734					
	Outcome: Absence (sessions)					
	OLS					2SLS
	(1)	(2)	(3)	(4)	(5)	(6)
Spillover effect	0.057***	0.057***	0.056***	0.041***	0.039***	0.041***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.011)
Year fixed effects		✓	✓	✓	✓	✓
Grade fixed effects			✓	✓	✓	✓
Individual and family controls				✓	✓	✓
School peer behaviour controls					✓	✓
Observations	1,142,734					

Notes: National Pupil Database, 2007 - 2019. Sample is restricted to the eldest two siblings per family and year, attending a mainstream Secondary school, with no missing covariates, with at least a two year grade gap, and observed in grades 8, 9 or 10. Exclusions are temporary exclusions only. Absences are unauthorised absences only. One session is half a school day. The individual and family controls are prior attainment at age 11 (standardised), Free School Meals eligibility, whether first language is English, sibship sex combination, and sibship age gap. The school peer behaviour controls are three-year rolling averages of homogeneous peers' misbehaviour for both siblings, and the younger sibling's own homogeneous peers' average misbehaviour. The instrumental variable is the average misbehaviour of the older sibling's homogeneous peers. All peer measures are constructed using take-one-out means. Note that in the instrumental variable estimation for the exclusions outcome (column 6 in the top panel) some parameters were not identified. * Significant at 10%; ** significant at 5%; *** significant at 1%.

References

Alan, S., Baysan, C., Gumren, M., and Kubilay, E. (2021). Building social cohesion in ethnically mixed schools: An intervention on perspective taking. *The Quarterly Journal of Economics*, 136(4):2147–2194.

Alan, S., Ertac, S., and Mumcu, I. (2018). Gender stereotypes in the classroom and effects on achievement. *Review of Economics and Statistics*, 100(5):876–890.

Alesina, A., Carlana, M., La Ferrara, E., and Pinotti, P. (2018). Revealing stereotypes: Evidence from immigrants in schools. Working Paper 25333, National Bureau of Economic Research.

Allen, R., Evans, M., and White, B. (2021). *The Next Big Thing in School Improvement*. John Catt.

Almlund, M., Duckworth, A. L., Heckman, J., and Kautz, T. (2011). Personality psychology and economics. In Hanushek, E. A., Machin, S., and Woessmann, L., editors, *Handbook of The Economics of Education*, volume 4, chapter 1, pages 1–181. Elsevier.

Altmejd, A., Barrios-Fernández, A., Drlje, M., Goodman, J., Hurwitz, M., Kovac, D., Mulhern, C., Neilson, C., and Smith, J. (2021). O brother, where start thou? Sibling spillovers on college and major choice in four countries. *The Quarterly Journal of Economics*, 136(3):1831–1886.

Altonji, J. G., Cattan, S., and Ware, I. (2017). Identifying sibling influence on teenage substance use. *Journal of Human Resources*, 52(1):1–47.

Anand, P. and Kahn, L. B. (2023). The effect of a peer's teen pregnancy on sexual behavior. Working Paper 31228, National Bureau of Economic Research.

Attanasio, O., Cattan, S., Fitzsimons, E., Meghir, C., and Rubio-Codina, M. (2020). Estimating the production function for human capital: Results from a randomized controlled trial in colombia. *American Economic Review*, 110(1):48–85.

Bailey, D. H., Duncan, G. J., Cunha, F., Foorman, B. R., and Yeager, D. S. (2020). Persistence and fade-out of educational-intervention effects: Mechanisms and potential solutions. *Psychological Science in the Public Interest*, 21(2):55–97.

Baird, J. (1998). What's in a name? Experiments with blind marking in A-Level examinations. *Educational Research*, 40(2):191–202.

Baker, A. C., Larcker, D. F., and Wang, C. C. (2022). How much should we trust staggered difference-in-differences estimates? *Journal of Financial Economics*, 144(2):370–395.

Bandura, A. (1977). *Social Learning Theory*. Prentice-Hall.

Becker, G. S. and Tomes, N. (1976). Child endowments and the quantity and quality of children. *Journal of Political Economy*, 84(4):S143–S162.

Behrman, J. R., Pollak, R. A., and Taubman, P. (1982). Parental preferences and provision for progeny. *Journal of Political Economy*, 90(1):52–73.

Bennett, M. and Bergman, P. (2021). Better together? Social networks in truancy and the targeting of treatment. *Journal of Labor Economics*, 39(1):1–36.

Bertrand, M. and Pan, J. (2013). The trouble with boys: Social influences and the gender gap in disruptive behavior. *American Economic Journal: Applied Economics*, 5(1):32–64.

Bharadwaj, P., Eberhard, J. P., and Neilson, C. A. (2018). Health at birth, parental investments, and academic outcomes. *Journal of Labor Economics*, 36(2):349–394.

Bingley, P., Lundborg, P., and Lyk-Jensen, S. V. (2021). Brothers in arms: Spillovers from a draft lottery. *Journal of Human Resources*, 56(1):225–268.

Biroli, P., Del Boca, D., Heckman, J. J., Heckman, L. P., Koh, Y. K., Kuperman, S., Moktan, S., Pronzato, C. D., and Ziff, A. L. (2018). Evaluation of the Reggio approach to early education. *Research in Economics*, 72(1):1–32.

Black, M. M., Walker, S. P., Fernald, L. C., Andersen, C. T., DiGirolamo, A. M., Lu, C., McCoy, D. C., Fink, G., Shawar, Y. R., Shiffman, J., et al. (2017). Early childhood development coming of age: Science through the life course. *The Lancet*, 389(10064):77–90.

Black, N. and de New, S. C. (2020). Short, heavy and underrated? Teacher assessment biases by children's body size. *Oxford Bulletin of Economics and Statistics*, 82(5):961–987.

Blanden, J., Del Bono, E., McNally, S., and Rabe, B. (2016). Universal pre-school education: The case of public funding with private provision. *The Economic Journal*, 126(592):682–723.

Blinder, A. S. (1973). Wage discrimination: Reduced form and structural estimates. *Journal of Human Resources*, pages 436–455.

Bordalo, P., Coffman, K., Gennaioli, N., and Shleifer, A. (2016). Stereotypes. *The Quarterly Journal of Economics*, 131(4):1753–1794.

Borghans, L., Duckworth, A. L., Heckman, J. J., and Ter Weel, B. (2008). The economics and psychology of personality traits. *Journal of Human Resources*, 43(4):972–1059.

Borusyak, K., Jaravel, X., and Spiess, J. (2024). Revisiting event-study designs: Robust and efficient estimation. *Review of Economic Studies*, 91(6):3253–3285.

Botelho, F., Madeira, R. A., and Rangel, M. A. (2015). Racial discrimination in grading: Evidence from brazil. *American Economic Journal: Applied Economics*, 7(4):37–52.

Brim, O. G. (1958). Family structure and sex role learning by children: A further analysis of helen koch's data. *Sociometry*, 21(1):1–16.

Broughton, T., Langley, K., Tilling, K., and Collishaw, S. (2023). Relative age in the school year and risk of mental health problems in childhood, adolescence and young adulthood. *Journal of Child Psychology and Psychiatry*, 64(1):185–196.

Buhrmester, D. (1992). The developmental courses of sibling and peer relationships. In Boer, F., Dunn, J., and Dunn, J., editors, *Children's Sibling Relationships: Developmental and Clinical Issues*, chapter 2. Psychology Press.

Burgess, S. and Greaves, E. (2013). Test scores, subjective assessment, and stereotyping of ethnic minorities. *Journal of Labor Economics*, 31(3):535–576.

Burgess, S., Hauberg, D. S., Rangvid, B. S., and Sievertsen, H. H. (2022). The importance of external assessments: High school math and gender gaps in stem degrees. *Economics of Education Review*, 88. Article 102267.

Burgess, S. M., Sanderson, E., and Umaña-Aponte, M. (2011). School ties: An analysis of homophily in an adolescent friendship network. Working Paper 11/267, Centre for Market and Public Organisation.

Callaway, B. and Sant'Anna, P. H. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225(2):200–230.

Campbell, T. (2015). Stereotyped at seven? Biases in teacher judgement of pupils' ability and attainment. *Journal of Social Policy*, 44(3):517–547.

Carlana, M. (2019). Implicit stereotypes: Evidence from teachers' gender bias. *The Quarterly Journal of Economics*, 134(3):1163–1224.

Carrell, S. E. and Hoekstra, M. L. (2010). Externalities in the classroom: How children exposed to domestic violence affect everyone's kids. *American Economic Journal: Applied Economics*, 2(1):211–228.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., and Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from Project STAR. *The Quarterly Journal of Economics*, 126(4):1593–1660.

Codiroli Mcmaster, N. (2017). Who studies STEM subjects at A level and degree in England? An investigation into the intersections between students' family background, gender and ethnicity in determining choice. *British Educational Research Journal*, 43(3):528–553.

Conti, G. and Gupta, S. (2024). Early childhood interventions to reduce intergenerational inequality. In Kilpi-Jakonen, E., Blanden, J., Erola, J., and Macmillan, L., editors, *Research Handbook on Intergenerational Inequality*, chapter 26, pages 342–356. Edward Elgar Publishing.

Cornelissen, T. and Dustmann, C. (2019). Early school exposure, test scores, and noncognitive outcomes. *American Economic Journal: Economic Policy*, 11(2):35–63.

Cornelissen, T., Dustmann, C., Raute, A., and Schönberg, U. (2018). Who benefits from universal child care? Estimating marginal returns to early child care attendance. *Journal of Political Economy*, 126(6):2356–2409.

Council of Skills Advisors (2022). Learning and skills for economic recovery, social cohesion and a more equal britain. Technical report, Labour Party.

Cunha, F. and Heckman, J. (2007). The technology of skill formation. *American Economic Review*, 97(2):31–47.

Cunha, F. and Heckman, J. (2008). Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation. *Journal of Human Resources*, 43(4):738–782.

Cunha, F., Heckman, J. J., and Schennach, S. M. (2010). Estimating the technology of cognitive and noncognitive skill formation. *Econometrica*, 78(3):883–931.

Dahl, G. B., Løken, K. V., and Mogstad, M. (2014). Peer effects in program participation. *American Economic Review*, 104(7):2049–2074.

Datar, A., Kilburn, M., and Loughran, D. (2010). Endowments and parental investments in infancy and early childhood. *Demography*, 47(1):145–162.

Davies, O. R., Taylor, C., Maynard, T., Rhys, M., Waldron, S., and Blackaby, D. (2013). Evaluating the Foundation Phase: The outcomes of Foundation Phase pupils (Report 1). Technical report, Wales Institute of Social and Economic Research and Data.

Davies, O. R., Taylor, C., Rhys, M., Waldron, S., and Blackaby, D. (2015). Evaluating the Foundation Phase: The outcomes of Foundation Phase pupils up to 2011/12 (Report 2). Technical report, Wales Institute of Social and Economic Research and Data.

De Benedetto, M. A. and De Paola, M. (2023). Immigration and teacher bias towards students with an immigrant background. *Economic Policy*, 38(113):107–154.

De Chaisemartin, C. and d'Haultfoeuille, X. (2023). Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey. *The Econometrics Journal*, 26(3):C1–C30.

De Haan, M. (2010). Birth order, family size and educational attainment. *Economics of Education Review*, 29(4):576–588.

Dee, T. S. (2005). A teacher like me: Does race, ethnicity, or gender matter? *American Economic Review*, 95(2):158–165.

Dee, T. S. and Sievertsen, H. H. (2017). The gift of time? School starting age and mental health. *Health Economics*, 27(5):781–802.

Defoe, I. N., Keijsers, L., Hawk, S. T., Branje, S., Dubas, J. S., Buist, K., Frijns, T., A. G. van Aken, M., Koot, H. M., van Lier, P. A. C., and Meeus, W. (2013). Siblings versus parents and friends: Longitudinal linkages to adolescent externalizing problems. *Journal of Child Psychology and Psychiatry*, 54(8):881–889.

Del Bono, E., Etheridge, B., and Garcia, P. (2024). The economic value of childhood socio-emotional skills. ISER Working Paper Series 2024-01, Institute for Social and Economic Research.

Del Bono, E., Fumagalli, L., Holford, A., and Rabe, B. (2022). University access: The role of background and covid-19 throughout the application process. ISER Working Paper Series 2022-07, Institute for Social and Economic Research.

Deming, D. J. (2011). Better schools, less crime? *The Quarterly Journal of Economics*, 126(4):2063–2115.

Deming, D. J. (2022). Four facts about human capital. *Journal of Economic Perspectives*, 36(3):75–102.

Department for Education (2011). A profile of pupil absence in England. Research report, Department for Education.

Department for Education (2020a). Ad-hoc notice: Early entry into GCSE examinations in England. Technical report, Department for Education.

Department for Education (2020b). Secondary accountability measures: Guide for maintained secondary schools, academies and free schools. Technical report, Department for Education.

Department for Education (2021). GCSE attainment and lifetime earnings. Research report, Department for Education.

Department for Education (2022). Post-qualification admissions consultation response. Technical report, Department for Education.

Dhuey, E., Figlio, D., Karbownik, K., and Roth, J. (2019). School starting age and cognitive development. *Journal of Policy Analysis and Management*, 38(3):538–578.

Dietrichson, J., Lykke Kristiansen, I., and Viinholt, B. A. (2020). Universal preschool programs and long-term child outcomes: A systematic review. *Journal of Economic Surveys*, 34(5):1007–1043.

Duncan, G., Kalil, A., Mogstad, M., and Rege, M. (2023). Investing in early childhood development in preschool and at home. In Hanushek, E. A., Woessmann, L., and Machin, S. J., editors, *Handbook of the Economics of Education*, volume 6, chapter 1, pages 1–77. Elsevier.

Dustan, A. (2018). Family networks and school choice. *Journal of Development Economics*, 134:372–391.

Dustmann, C., Machin, S., and Schönberg, U. (2010). Ethnicity and educational achievement in compulsory schooling. *The Economic Journal*, 120(546):F272–F297.

Eivers, E., Worth, J., and Ghosh, A. (2020). Home learning during covid-19: Findings from the understanding society longitudinal study. Technical report, National Foundation for Educational Research.

Engel, A., Barnett, W. S., Anders, Y., and Taguma, M. (2015). Norway: Early childhood education and care policy review. Technical report, Organisation for Economic Co-operation and Development.

Esponda, I., Oprea, R., and Yuksel, S. (2023). Seeing what is representative. *The Quarterly Journal of Economics*, 138(4):2607–2657.

Farquharson, C., McNally, S., and Tahir, I. (2024). Education inequalities. *Oxford Open Economics*, 3:i760–i820.

Fidjeland, A., Rege, M., Solli, I. F., and Størksen, I. (2023). Reducing the gender gap in early learning: Evidence from a field experiment in norwegian preschools. *European Economic Review*, 154. Article 104413.

Fredriksson, P. and Öckert, B. (2014). Life-cycle effects of age at school start. *The Economic Journal*, 124(579):977–1004.

Fryer Jr, R. G. (2017). The production of human capital in developed countries: Evidence from 196 randomized field experiments. In *Handbook of Economic Field Experiments*, volume 2, pages 95–322. Elsevier.

García, J. L., Bennhoff, F. H., Leaf, D. E., and Heckman, J. J. (2021). The dynastic benefits of early childhood education. Working Paper 29004, National Bureau of Economic Research.

Gaviria, A. and Raphael, S. (2001). School-based peer effects and juvenile behavior. *Review of Economics and Statistics*, 83(2):257–268.

Gelbach, J. B. (2016). When do covariates matter? And which ones, and how much? *Journal of Labor Economics*, 34(2):509–543.

Gelber, A. and Isen, A. (2013). Children's schooling and parents' behavior: Evidence from the Head Start Impact Study. *Journal of Public Economics*, 101:25–38.

Gennetian, L. A., Wolf, S., Hill, H. D., and Morris, P. A. (2015). Intrayear household income dynamics and adolescent school behavior. *Demography*, 52(2):455–483.

Gibbons, S. and Chevalier, A. (2008). Assessment and age 16+ education participation. *Research Papers in Education*, 23(2):113–123.

Gibbons, S., Silva, O., and Weinhardt, F. (2013). Everybody needs good neighbours? Evidence from students' outcomes in England. *The Economic Journal*, 123(571):831–874.

Glyn, E., Harries, S., Lane, J., and Lewis, S. (2019). Evaluation of the early implementation of the childcare offer for Wales: Year two. Technical report, Welsh Government.

Gonzalez, K. E. (2020). Within-family differences in Head Start participation and parent investment. *Economics of Education Review*, 74. Article 101950.

Gray-Lobe, G., Pathak, P. A., and Walters, C. R. (2023). The long-term effects of universal preschool in Boston. *The Quarterly Journal of Economics*, 138(1):363–411.

Greaves, E., Hussain, I., Rabe, B., and Rasul, I. (2023). Parental responses to information about school quality: Evidence from linked survey and administrative data. *The Economic Journal*, 133(654):2334–2402.

Gurantz, O., Hurwitz, M., and Smith, J. (2020). Sibling effects on high school exam taking and performance. *Journal of Economic Behavior & Organization*, 178:534–549.

Hanna, R. N. and Linden, L. L. (2012). Discrimination in grading. *American Economic Journal: Economic Policy*, 4(4):146–168.

Hanney, M. (2000). Early years provision for three year olds. Schools and Early Learning Committee, National Assembly for Wales.

Hanushek, E. A. (2020). Education production functions. In Bradley, S. and Green, C., editors, *The Economics of Education Second Edition*, chapter 13, pages 161–170. Elsevier.

Harjunen, O., Izadi, R., and Sarvimäki, M. (2022). Impact of preschool on school preparedness: Evidence from a nationwide RCT in Finland. Pre-Analysis Plan AEARCTR-0008061, AAEA RCT Registry.

Havnes, T. and Mogstad, M. (2015). Is universal child care leveling the playing field? *Journal of Public Economics*, 127:100–114.

Heckman, J., Pinto, R., and Savelyev, P. (2013). Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. *American Economic Review*, 103(6):2052–2086.

Heckman, J. J. and Kautz, T. (2012). Hard evidence on soft skills. *Labour Economics*, 19(4):451–464.

Heckman, J. J., Moon, S. H., Pinto, R., Savelyev, P. A., and Yavitz, A. (2010). The rate of return to the HighScope Perry Preschool Program. *Journal of Public Economics*, 94(1-2):114–128.

Heckman, J. J. and Rubinstein, Y. (2001). The importance of noncognitive skills: Lessons from the GED testing program. *American Economic Review*, 91(2):145–149.

Heckman, J. J., Stixrud, J., and Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor economics*, 24(3):411–482.

Hirsh-Pasek, K., Golinkoff, R. M., Berk, L. E., and Singer, D. (2009). *A mandate for playful learning in preschool: Applying the scientific evidence*. Oxford University Press.

Holmes, S., Churchward, D., Howard, E., Keys, E., Leahy, F., Tonin, D., and Black, B. (2021). Centre judgements: Teaching staff survey and interviews, Summer 2020. Technical report, Office of Qualifications and Examinations Regulation.

Imberman, S. A., Kugler, A. D., and Sacerdote, B. I. (2012). Katrina's children: Evidence on the structure of peer effects from hurricane evacuees. *American Economic Review*, 102(5):2048–2082.

Jackson, C. K. (2018). What do test scores miss? The importance of teacher effects on non–test score outcomes. *Journal of Political Economy*, 126(5):2072–2107.

Joensen, J. S. and Nielsen, H. S. (2018). Spillovers in education choice. *Journal of Public Economics*, 157:158 – 183.

John, A., Friedmann, Y., Del Pozo-Banos, M., Frizzati, A., Ford, T., and Thapar, A. (2022). Association of school absence and exclusion with recorded neurodevelopmental disorders, mental disorders, or self-harm: A nationwide, retrospective, electronic cohort study of children and young people in Wales, UK. *The Lancet Psychiatry*, 9(1):23–34.

Joint Council for Qualifications (2020). GCSE (Full Course) outcomes for main grade set for each jurisdiction (age 16): Results Summer 2020. Technical report, Joint Council for Qualifications.

Kautz, T., Heckman, J. J., Diris, R., Ter Weel, B., and Borghans, L. (2014). Fostering and measuring skills: Improving cognitive and non-cognitive skills to promote lifetime success. Working Paper 20749, National Bureau of Economic Research.

Kautz, T. and Zanoni, W. (2024). Measurement and development of noncognitive skills in adolescence: Evidence from Chicago public schools and the OneGoal program. *Journal of Human Capital*, 18(2):272–304.

Labaree, D. F. (1997). Public goods, private goods: The American struggle over educational goals. *American Educational Research Journal*, 34(1):39–81.

Landersø, R. K., Nielsen, H. S., and Simonsen, M. (2020). Effects of school starting age on the family. *Journal of Human Resources*, 55(4):1258–1286.

Landersø, R., Nielsen, H. S., and Simonsen, M. (2017). School starting age and the crime-age profile. *The Economic Journal*, 127(602):1096–1118.

Lavy, V. (2008). Do gender stereotypes reduce girls' or boys' human capital outcomes? Evidence from a natural experiment. *Journal of Public Economics*, 92(10-11):2083–2105.

Lavy, V. and Megalokonomou, R. (2019). Persistency in teachers' grading bias and effects on longer-term outcomes: University admissions exams and choice of field of study. Working Paper 26021, National Bureau of Economic Research.

Lavy, V. and Sand, E. (2018). On the origins of gender gaps in human capital: Short- and long-term consequences of teachers' biases. *Journal of Public Economics*, 167:263–279.

Lavy, V., Silva, O., and Weinhardt, F. (2012). The good, the bad, and the average: Evidence on ability peer effects in schools. *Journal of Labor Economics*, 30(2):367–414.

Leadbeater, B. J., Kuperminc, G. P., Blatt, S. J., and Hertzog, C. (1999). A multivariate model of gender differences in adolescents' internalizing and externalizing problems. *Developmental Psychology*, 35(5):1268–1282.

Lenard, M. A. and Silliman, M. (2025). Informal social interactions, academic achievement and behavior: Evidence from peers on the school bus. *The Economic Journal*. Article ueaf013.

Li, W., Duncan, G. J., Magnuson, K., Schindler, H. S., Yoshikawa, H., Leak, J., et al. (2020). Timing in early childhood education: How cognitive and achievement program impacts vary by starting age, program duration, and time since the end of the program. EdWorkingPaper 20-201, Annenberg Institute for School Reform at Brown University.

Lindahl, E. (2007). Comparing teachers' assessments and national test results: Evidence from Sweden. Working Paper 2007:24, Institute for Labour Market Policy Evaluation.

Lubotsky, D. and Kaestner, R. (2016). Do 'skills beget skills'? Evidence on the effect of kindergarten entrance age on the evolution of cognitive and non-cognitive skill gaps in childhood. *Economics of Education Review*, 53:194–206.

Machin, S., McNally, S., and Ruiz-Valenzuela, J. (2020). Entry through the narrow door: The costs of just failing high stakes exams. *Journal of Public Economics*, 190. Article 104224.

Macmillan, L. and Tominey, E. (2023). Parental inputs and socio-economic gaps in early child development. *Journal of Population Economics*, 36(3):1513–1543.

Manski, C. (1993). Identification of endogenous social effects: The reflection problem. *Review of Economic Studies*, 60(3):531–542.

Maynard, T., Taylor, C. M., Waldron, S., Rhys, M., Smith, R., Power, S., and Clement, J. (2013). Evaluating the Foundation Phase: Policy logic model and programme theory. Technical report, Wales Institute of Social and Economic Research and Data.

McHale, S. M., Updegraff, K. A., and Whiteman, S. D. (2012). Sibling relationships and influences in childhood and adolescence. *Journal of Marriage and the Family*, 74(5):913–930.

Moffitt, R. (2001). Policy interventions, low-level equilibria, and social interactions. In Durlauf, S. and Young, P., editors, *Social Dynamics*, chapter 3, pages 45–82. MIT Press.

Moroni, G., Nicoletti, C., Tominey, E., et al. (2019). Child socio-emotional skills: The role of parental inputs. IZA Discussion Papers 12432, Institute of Labor Economics.

Murphy, R. and Wyness, G. (2020). Minority report: the impact of predicted grades on university admissions of disadvantaged groups. *Education Economics*, 28(4):333–350.

National Assembly for Wales (2000a). Desirable outcomes for children's learning before compulsory school age. Technical report, National Assembly for Wales.

National Assembly for Wales (2000b). The National Curriculum in Wales. Technical report, Qualifications, Curriculum and Assessment Authority for Wales (ACCAC) on behalf of the National Assembly for Wales. http://www.accac.org.uk/publications/ncorders.html.

National Assembly for Wales (2003). The learning country: Foundation phase, 3 to 7 years. Technical report, National Assembly for Wales.

National Assembly for Wales (2008). Framework for children's learning for 3 to 7-year-olds in Wales. Technical report, Department for Children, Education, Lifelong Learning and Skills, National Assembly for Wales.

Nicodemo, C., Nicoletti, C., and Vidiella-Martin, J. (2024). Starting school and ADHD: When is it time to fly the nest? IZA Discussion Papers 17091, Institute of Labor Economics.

Nicoletti, C. and Rabe, B. (2013). Inequality in pupils' test scores: How much do family, sibling type and neighbourhood matter? *Economica*, 80(318):197–218.

Nicoletti, C. and Rabe, B. (2019). Sibling spillover effects in school achievement. *Journal of Applied Econometrics*, 34(4):482–501.

Nicoletti, C., Salvanes, K. G., and Tominey, E. (2018). The family peer effect on mothers' labor supply. *American Economic Journal: Applied Economics*, 10(3):206–234.

Noble, K. G., Houston, S. M., Brito, N. H., Bartsch, H., Kan, E., Kuperman, J. M., Akshoomoff, N., Amaral, D. G., Bloss, C. S., Libiger, O., et al. (2015). Family income, parental education and brain structure in children and adolescents. *Nature Neuroscience*, 18(5):773–778.

Oaxaca, R. (1973). Male-female wage differentials in urban labor markets. *International Economic Review*, pages 693–709.

Office for National Statistics (2022). Population of england and wales. Technical report, Office for National Statistics. https://www.ethnicity-facts-figures.service.gov.uk/uk-population-by-ethnicity/national-and-regional-populations/population-of-england-and-wales.

Ofqual (2020a). Appeals for GCSE, AS, A Level and Project: 2018 to 2019 academic year. Technical report, Office of Qualifications and Examinations Regulation.

Ofqual (2020b). Provisional Entries for GCSE, AS and A level: Summer 2020 exam series. Technical report, Office of Qualifications and Examinations Regulation.

Ofqual (2020c). Student-level equalities analyses for GCSE and A level: Summer 2020. Technical report, Office of Qualifications and Examinations Regulation.

Ofqual (2020d). Summer 2020 grades for GCSE, AS and A level, Extended Project Qualification and Advanced Extension Award in maths: Information for Heads of Centre, Heads of Department/subject eads and teachers on the submission of centre assessment grades. Updated 22 May 2020. Technical report, Office of Qualifications and Examinations Regulation.

Ofqual (2022). Consultation outcome: Guidance on designing and developing accessible assessments. Technical report, Office of Qualifications and Examinations Regulation.

Ofqual (2023). Guidance on collecting evidence of student performance to ensure resilience in the qualifications system. Technical report, Office of Qualifications and Examinations Regulation.

Papageorge, N. W., Ronda, V., and Zheng, Y. (2019). The economic value of breaking bad: Misbehavior, schooling and the labor market. Working Paper 25602, National Bureau of Economic Research.

Pepler, D. J., Abramovitch, R., and Corter, C. M. (1981). Sibling interaction in the home: A longitudinal study. *Child Development*, 52:1344–1347.

Petek, N. and Pope, N. G. (2023). The multidimensional impact of teachers on students. *Journal of Political Economy*, 131(4):1057–1107.

Phelps, E. S. (1972). The statistical theory of racism and sexism. *American Economic Review*, 62(4):659–661.

Pope, N. G. and Zuo, G. W. (2023). Suspending suspensions: The education production consequences of school suspension policies. *The Economic Journal*, 133(653):2025–2054.

Power, S., Rhys, M., Taylor, C., and Waldron, S. (2019). How child-centred education favours some learners more than others. *Review of Education*, 7(3):570–592.

Rege, M., Størksen, I., Solli, I. F., Kalil, A., McClelland, M. M., Ten Braak, D., Lenes, R., Lunde, S., Breive, S., Carlsen, M., et al. (2024). The effects of a structured curriculum on preschool effectiveness: A field experiment. *Journal of Human Resources*, 59(2):576–603.

Rosales-Rueda, M. F. (2014). Family investment responses to childhood health conditions: Intrafamily allocation of resources. *Journal of Health Economics*, 37:41–57.

Rosenzweig, M. R. and Zhang, J. (2009). Do population control policies induce more human capital investment? Twins, birth weight and China's "One-Child" policy. *The Review of Economic Studies*, 76(3):1149–1174.

Roth, J. (2022). Pretest with caution: Event-study estimates after testing for parallel trends. *American Economic Review: Insights*, 4(3):305–22.

Roth, J., Sant'Anna, P. H., Bilinski, A., and Poe, J. (2023). What's trending in difference-in-differences? A synthesis of the recent econometrics literature. *Journal of Econometrics*, 235(2):2218–2244.

Siraj-Blatchford, I., Milton, E., Sylva, K., Laugharne, J., and Charles, F. (2007). Developing the Foundation Phase for 3–7-year-olds in Wales. *Wales Journal of Education*, 14(1):43–68.

Siraj-Blatchford, I., Sylva, K., Laugharne, J., Milton, E., and Charles, F. (2005). Monitoring and evaluation of the effective implementation of the Foundation Phase project across Wales. Technical report, Welsh Government.

Sorrenti, G., Zölitz, U., Ribeaud, D., and Eisner, M. (2025). The causal impact of socio-emotional skills training on educational success. *Review of Economic Studies*, 92(1):506–552.

Starck, J. G., Riddle, T., Sinclair, S., and Warikoo, N. (2020). Teachers are people too: Examining the racial bias of teachers compared to other american adults. *Educational Researcher*, 49(4):273–284.

StatsWales (2023a). School census results, 1974 to 2005. Technical report, StatsWales. https://statswales.gov.wales/Catalogue/Education-and-Skills/Schools-and-Teachers/Schools-Census/1974-to-2005/Pupils.

StatsWales (2023b). Schools by local authority, region and welsh medium type. Technical report, StatsWales. https://statswales.gov.wales/Catalogue/Education-and-Skills/Schools-and-Teachers/Schools-Census/Pupil-Level-Annual-School-Census/Welsh-Language/Schools-by-LocalAuthorityRegion-WelshMediumType.

Sun, L. and Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, 225(2):175–199.

Tanjitpiyanond, P., Jetten, J., and Peters, K. (2022). How economic inequality shapes social class stereotyping. *Journal of Experimental Social Psychology*, 98. Article 104248.

Taylor, C., Joshi, H., and Wright, C. (2015a). Evaluating the impact of early years educational reform in Wales to age seven: The potential use of the UK Millennium Cohort Study. *Journal of Education Policy*, 30(5):688–712.

Taylor, C., Rhys, M., Waldron, S., Davies, R., Power, S., Maynard, T., Moore, L., Blackaby, D., and Plewis, I. (2015b). Evaluating the Foundation Phase: Final report. Technical report, Wales Institute of Social and Economic Research and Data.

Terrier, C. (2020). Boys lag behind: How teachers' gender biases affect student achievement. *Economics of Education Review*, 77. Article 101981.

Van Ewijk, R. (2011). Same work, lower grade? Student ethnicity and teachers' subjective assessments. *Economics of Education Review*, 30(5):1045–1058.

von Hinke, S., Leckie, G., and Nicoletti, C. (2019). The use of instrumental variables in peer effects models. *Oxford Bulletin of Economics and Statistics*, 81(5):1179–1191.

Waldron, S., Rhys, M., and Taylor, C. M. (2014). Evaluating the Foundation Phase: Key findings on pedagogy and understanding. Technical report, Wales Institute of Social and Economic Research and Data.

Welsh Government (2009). School admissions code. Technical report, Welsh Government.

Welsh Government (2014). Welsh Index of Multiple Deprivation (WIMD) 2014: Revised. Technical report, Welsh Government.

Whiteman, S. D., McHale, S. M., and Soli, A. (2011). Theoretical perspectives on sibling relationships. *Journal of Family Theory & Review*, 3(2):124–139.

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

Wilton, J. and Davies, R. (2017). Flying start evaluation: Educational outcomes. Technical report, Wales Institute of Social and Economic Research and Data.

Yi, J., Heckman, J. J., Zhang, J., and Conti, G. (2015). Early health shocks, intra-household resource allocation and child outcomes. *The Economic Journal*, 125(588):F347–F371.