Nurturing inequality: How structural and psychological processes create difference among primary school children

A thesis in three papers

Tammy Campbell

Department of Quantitative Social Science

UCL Institute of Education

University College London

University of London

PhD Quantitative Social Science for Social Policy

Declaration

I hereby declare that, except where explicit attribution is made, the work presented in this thesis is entirely my own.

Word count (exclusive of appendices and list of references): **37,850** words

Related work

This thesis draws upon five main publications produced exclusively during my PhD: three sole-authored working papers, and two sole-authored journal articles.

The working papers are:

> Campbell, T. (2013a). 'In-school ability grouping and the month of birth effect: Preliminary evidence from the Millennium Cohort Study.' [Online]. Available at: http://www.cls.ioe.ac.uk/shared/get-file.ashx?itemtype=document&id=1618 [Last accessed 5th May 2015.]
>
> Campbell, T. (2013b). 'Stereotyped at seven? Biases in teacher judgements of pupils' ability and attainment.' [Online]. Available at: http://www.cls.ioe.ac.uk/shared/get-file.ashx?itemtype=document&id=1715 [Last accessed 5th May 2015.]
>
> Campbell, T. (2014). 'Selected at seven: The relationship between teachers' judgements and assessments of pupils, and pupils' stream placements.' [Online]. Available at: http://repec.ioe.ac.uk/REPEc/pdf/qsswp1410.pdf [Last accessed 21st May 2015.]

The journal articles are:

> Campbell, T. (2014). 'Stratified at seven: in-class ability grouping and the relative age effect'. *British Educational Research Journal*, 40(5), 749-771.

> Campbell, T. (2015). ‘Stereotyped at Seven? Biases in Teacher Judgement of Pupils’ Ability and Attainment’. *Journal of Social Policy,* 44(3), 517-547.

An additional journal article, co-authored with a fellow PhD student, and exploring the cross-over between our fields of study, has also been produced:

> Shackleton, N. & Campbell, T. (2014). ‘Are teachers' judgements of pupils' ability influenced by body shape?’ *International Journal of Obesity*, 38(4), 520-524.

Several ongoing, related strands of work have been pursued throughout this PhD, and are in currently progress towards becoming working papers and / or journal articles. They include investigations of:

- Whether teachers’ perceptions of pupils’ behaviour can provide any explanation for biases in judgements of pupils’ ability and attainment.
- Whether available data can be used to establish a control for teacher perceptions of pupil behaviour, in order to allow robust exploration of whether there are disproportionalities and biases in perceptions of behaviour according to pupil characteristic.
- The extent to which young children’s behaviour / teachers’ perceptions of their behaviour, relative to their peers, can be implicated in the formation of the ‘month of birth effect’ in early primary school.
- The interaction between gestational age and month of birth in influencing children’s development and educational experiences, and implications of this for school starting age policies.

New areas of work have also begun to be explored during the course of this PhD, as familiarity with appropriate data has been built. They include:

- Investigations of the impact of mothers’ maternity and birth experiences on children’s outcomes and development throughout childhood.

- Investigations of correlates with, explanations for, and heterogeneity in the apparent associations between breastfeeding and children's cognitive development.

Abstract

This thesis makes a unique contribution to knowledge with three papers presenting new empirical evidence on factors involved in early educational inequalities.

The first explores whether streaming influences teachers' judgements of children. It investigates whether pupils who perform equivalently in cognitive tests, and who are similar according to a wide range of additional characteristics, are perceived differently by their teachers in line with their stream placement level. By testing associations across situations and subjects, consistent indications that stream placement has an effect on judgement are produced. Streaming is becoming more prevalent within early primary schooling, so this paper makes a timely contribution to the debate on whether the practice is efficient or equitable.

The second paper investigates bias and stereotyping in teachers' perceptions of pupils. It compares children's manifest performance to teachers' judgements of their ability and attainment, and indicates biases according to all key pupil-level characteristics documented as underpinning gaps in primary achievement. It therefore questions prevalent policy assumptions regarding the construction of early educational inequalities, and suggests that refocussing policy to include more understanding of the impact of bias and stereotyping could help tackle disparities.

The third paper examines whether early in-class ability grouping may play a part in forming the 'month of birth effect,' where children relatively younger in their year group attain lower academic levels than their comparatively older peers. It focusses on teachers' early judgements of children, and compares pupils in schools that in-class group to those not grouping in this way. It shows more polarisation by birth month in teachers' evaluations when grouping takes place. As teachers' judgements influence children's education both at an everyday level and through formal assessments, this suggests that early grouping might be important to birth month inequalities, and that cessation of the practice may increase parity.

Contents

List of tables

List of figures

Acknowledgements

I am very grateful to the Economic and Social Research Council for funding this PhD. It uses data from the Millennium Cohort Study, and I am grateful to the Centre for Longitudinal Studies, UCL Institute of Education for the use of these data, and to the UK Data Archive and Economic and Social Data Service for making them available.

Thanks

Many thanks to Lorraine Dearden, who has supervised this thesis throughout, offering invaluable support and critique, and without whom I would not have even applied for PhD candidature.

Thanks also to all the other enabling and motivating staff who taught me on the (then) MSc Policy Analysis and Evaluation at the Institute of Education – a course which enhanced and deepened my interest in this field, and opened the doors to further study.

Thanks to Lucinda Platt for supervising the initial stages of my PhD with unswerving rigour and insight, and to Kirstine Hansen for stepping in at the end to complete my supervision with useful advice and suggestions.

Thanks to the many staff and students in the Department for Quantitative Social Science and in the wider Institute and research community, from whom I have learned greatly and been challenged extensively. Thought-provoking feedback and questions from a variety of perspectives, exchange of knowledge and ideas, and camaraderie and friendship have all enhanced the PhD experience.

Lastly, thanks to my family – especially to Sam, for interminable proofreading and for being a willing sounding board, and to Cal and Isaac for allowing Mummy’s many days playing Angry Birds on the computer. (Or something like that. It’s difficult to explain quantitative research to a small child).

All views and analyses presented in this thesis are mine, and none of the individuals, groups or organisations mentioned above are responsible for any of them. Any errors remain my own.

Chapter 1

Introduction

‘Gaps,’ ‘equity,’ and educational attainment

The uniting purpose of the three empirical papers presented in this thesis is to add to a working understanding of the factors that contribute to inequalities during early schooling. This is an aim that is of enduring relevance to educational policy-making, and that has been validated and corroborated by the academic and research community. Though there has been some ambivalence regarding trends, variations, and patterns in disparities, the persistent and ongoing presence and importance of attainment inequities is undisputed (Lupton & Thomson, 2015; Kerr & West, 2010; Kendall *et al*, 2008; Whitty & Anders, 2014; Sullivan *et al*, 2011).

What, then, are the ‘gaps’ upon which researchers and policy-makers have focussed in recent years? Firstly, ‘Ethnic minority pupils’ were proposed by Labour in 1997 as an underperforming ‘group’ (Department for Education and Employment, 1997, p.34), alongside children with special educational needs (SEN) (ibid, p.33), and boys (ibid, p.39). In the intervening years, regular national census data have begun to be compiled, measuring academic attainment against five main pupil-level characteristics: ethnicity, SEN, gender, free school meals (FSM) eligibility (a proxy for family income-level), and English as a first / second language (EAL). (Department for Children, Schools and Families, 2007; 2008a; Department for Education, 2014a; Department for Education, 2014b).

More recently, attainment in the early years has also been reported according to pupil month of birth (Department for Education, 2014c). This follows extensive internal departmental and externally commissioned analysis which clearly indicated a relationship between birth month and academic attainment. During early primary school, this association has in fact been more consistent over time than some of the relationships between achievement and the five factors already highlighted (Crawford *et al*, 2007; Department for Education, 2010b).

Tackling disparities

In order to formulate interventions, to engender change, and to close these longstanding ‘gaps,’ continuing empirical research is necessary to examine the factors that might be instrumental in the creation of difference within schooling.

The three papers presented within this thesis therefore contribute to continuing debate on how better potentially to create equity within the education system. They unpick key psychological and structural elements that seem to be instrumental in contributing to disparities during the primary phrase, and discuss the ways that changes to related processes may diminish inequalities. The first paper focusses on the effects of streaming, the second on perceptual bias and stereotyping, and the third on the role of early ability grouping in forming differences among pupils according to their month of birth.

Streaming

Streaming of young primary children has increased in recent years (Hallam et al, 2013), alongside governmental endorsement of early in-school ability groupings (Boaler, 1997; Conservative Party, 2007; Department for Children, Schools, and Families, 2008; Department for Education, 1992; Department for Education, 2010; Department for Education and Skills, 2005). This is despite a body of evidence which, upon systematic examination, suggests that grouping neither raises overall average attainment nor leads to greater parity or equality of opportunity or achievement (Dunne et al, 2007; Slavin, 1990; Sutton Trust / Educational Endowment Foundation, 2014).

Francis et al (2016) have argued that there is an apparent dearth, to date, of impact of research on streaming on policy-making – a contention backed by the recent increase in prevalence of the practice. This, they argue, can be explained by ‘cultural investments in discourses of “natural order” and hierarchy’ (p 1): that is, driving, historically underpinned, ‘common sense’ notions that children are of different ‘types,’ and can, as such, be sorted into streams.

So Chapter 3 of this thesis presents a timely challenge to this discourse and to the proliferation of ability grouping in early primary school. It examines explicitly the possible effects of streaming, exploring a mechanism for its potential impacts – teacher perceptions. It presents evidence that suggests a contribution of early streaming to attainment inequalities and differences between pupils, and it contends that the resurgence of fixed grouping structures, and of streaming in particular, may be detrimental to the original ideals of the comprehensive system: to equity, and to the lessening of between-pupil 'gaps.'

Perceptual bias and stereotyping

The second paper of this thesis focusses upon the part played by judgement biases and stereotyping in creating between-group variation in average pupil attainment levels. It presents analyses examining patterns of bias and stereotyping in teachers' judgements of primary pupils' ability and attainment, and discusses implications of these patterns for the construction and measurement of 'attainment' and for policy-making and implementation.

Longstanding evidence on the importance of teacher perceptions, and their effects on children's progress, provides a rationale for this Chapter (e.g. Rosenthal & Jacobsen, 1968; Rubie-Davies, 2010), alongside previous studies suggesting a tendency both to error and to bias in teacher judgements (e.g. Harlen, 2004), and a contemporary policy context where the means and uses of teacher assessments of pupils are being questioned and are in flux.[1]

Month of birth effects, and ability grouping

The third paper of this thesis concentrates on the contribution of early in-class ability grouping to the creation of difference between pupils according to relative age within year group cohort. As noted above, in-year

[1] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/483058/Commission_on_Assessment_Without_Levels_-_report.pdf
https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/358070/NC_assessment_quals_factsheet_Sept_update.pdf

disproportionalities by birth month in academic attainment (and in non-academic experiences) have increasingly been recognised at the research- and policy-levels (Crawford *et al*, 2007; Crawford *et al*, 2011; Crawford *et al*, 2013a; Department for Education, 2010b; Department for Education, 2014c; Sharp *et al*, 2009), and there is growing concern among parents and educators regarding the causes of month of birth effects and possible ways in which the effects can be lessened (Campaign for Flexible School Admissions for Summer Born Children, n.d.; Pre-School Learning Alliance, 2015; The Association for Professional Development in Early Years, 2015).

A 2015 session of the Education Select Committee on Summer Born Children (summer-born children are, in England, the relatively youngest within their school year) elicited over 100 pieces of written evidence from parents, campaigners, practitioners and academics (Commons Select Committee, n.d.). The committee concluded that: ‘There is widespread agreement that a problem exists, on average, for summer born...children’ (Education Committee, 2015) - and that timely investigation of the factors that contribute to this problem is vital.

Chapter 5 therefore focusses on the potential influence of one systematic, institutional practice in the formation of relative age effects: early in-class ability grouping. Ability grouping is identified as a likely candidate in the formation of relative age differences, because, to a significant degree, children are distributed across in-school groupings according to their age (Hallam & Parsons, 2013; Campbell, 2013).

Like Chapter 3, therefore, Chapter 5 unpicks the potential for ability grouping in primary school to create difference between children; and, in common with both previous chapters, it concentrates on the mediating pathway of teacher perceptions and judgements. It contributes to ongoing discussion among policy-makers, researchers, parents and practitioners about the factors that are important in the creation of relative age effects, and suggests an intervention that might alleviate these effects.

Summary

Thus the three papers of this thesis make an original contribution to knowledge and to ongoing policy and practice formulation in areas both of perennial interest and relevance and of current prominence. Each of the chapters make recommendations as a result of their analysis, and the discussion section continues to outline the questions raised, the unfolding policy context, and priorities for future research.

Thesis structure

The remainder of this document is structured as follows. Chapter 2 discusses the data used for analysis, raising and addressing key methodological issues, and stating the approaches to be taken. Chapter 3 contains the first substantive paper: an exploration of the relationship between stream placement and teacher perceptions. Chapter 4 presents the next paper – an examination of patterns of bias and stereotyping in teacher judgements of pupils. Lastly, the third paper, in Chapter 5, investigates the hypothesis that early in-class ability grouping plays a part in month of birth differences between children.

Each empirical Chapter stands alone, and begins with an introduction to the literature preceding and underpinning the analysis presented. The discussion and conclusions summarise the key findings of the thesis, while situating them in the emergent policy context, and making recommendations for further research which will build upon the analyses here.

Chapter 2

Data, sample, and analytical decisions

The Millennium Cohort Study teacher survey

This thesis is based on data from the Millennium Cohort Study (MCS), an ongoing longitudinal investigation of a national UK sample including 11,695 babies born in England around the turn of the century. The children and their families have been interviewed five times to date: within the child's first year (2001), then at ages three (2004), five (2006), seven (2008) and 12 (2012) (Hansen *et al*, 2012).

In 2008, at wave four, when study pupils were in Year Two and aged seven, a subsample of English MCS children's teachers participated in a separate survey (see Johnson *et al*, 2011 for details), and responses to this sub-study form the core of this thesis.

A reduced sample

Inclusion of children in the teacher survey depended on a number of contingencies, as detailed in Table 2.1.

Table 2.1: Contingencies upon which children's inclusion in the MCS wave four teacher sample are conditional

Contingency 1	Families must have lived in the eligible UK population at MCS wave one
Contingency 2	Families must have selected to participate at wave one
Contingency 3	Families must have responded at wave one (or two, when the MCS sample was boosted and new families included)
Contingency 4	Families must have continued to participate at wave four
Contingency 5	Families must have given permission and appropriate details to facilitate teacher contact and participation at wave four
Contingency 6	Teachers must have responded

Inevitably, then, given the multiple opportunities for attrition and non-response over and within waves, the sample eventually included in the teacher survey differ from the MCS's intended English population of:

> All children born between 1 September 2000 and 31 August 2001…alive…at age nine months (when the first wave of MCS interviews was intended to take place), and eligible to receive Child Benefit at that age; and, after nine months: for as long as they remain living in the UK at the time of sampling' (Plewis *et al*, 2007).

At age seven, 8,887 interviews took place in England, of which 5,627 (63 percent) also generated responses to the teacher-completed questionnaire (Johnson *et al*, 2011). Therefore, roughly, just under half of the original English sample remains for analysis using the teacher survey data.[2]

Throughout analyses in this thesis, twins and triplets are removed from the data, because the social and psychological processes investigated might differ for these children – for example, having a twin in the same year group might influence a child's stream placement, or their teachers' perceptions of their ability. This leaves a base total of 5,481 seven-year-old English singleton cases. In Chapters 3 and 4, children whose parents report them as attending a fee-paying school are additionally excluded, because this research is contextualised in policy on non-fee-paying, state education – leaving around 5000 working cases with other key information.

The samples used throughout this thesis are, therefore, not perfectly representative of the population of Year Two children in England in 2008. However, the MCS teacher survey is the only large, recent, UK study of primary school children which provides, for example, information on ability grouping practices and positions. This offers a unique opportunity to explore and to begin to quantify the potential effects of contemporary grouping practices. Similarly, it is the only large, current data collection allowing comparison of children's performance on external cognitive tests to teachers' perceptions of their ability. Findings from work using the MCS teacher survey are useful, therefore, because they can form the basis for theory building and policy discussion, notwithstanding the survey's inevitable imperfections.

[2] There is a slight fuzziness around the edges of this sample, due to cross-border movement over waves. 8,767 (98.6 percent) of the 8,887 families interviewed in England at wave four also lived in England at the first wave during which they participated. It is also possible that a small number of the pupils interviewed in England attend a school across the border in a different country, if, for example, they move house after commencing education.

Sample characteristics

Of course, the more accurately research using the teacher sample can represent its target population, the more usable and relevant it may be. An initial check of topline characteristics among sample families reveals no major divergences from the population of English seven-year-olds in the reporting year most closely related to survey fieldwork, as documented by the (then) Department for Children, Schools and Families (Department for Children, Schools and Families 2009b; Department for Children, Schools and Families 2009c). Table x indicates that the sample and the population are fairly balanced according to gender, ethnicity, language spoken in the home, and special educational needs status (the discrepancy between proportions with a statement of SEN may be due to centrally collated official statistics lagging behind survey-reported local diagnoses).

Table 2.2: Pupil characteristics in the English MCS age seven (2008) sample (n = 5481) with completed teacher questionnaires, and in the English school population in 2008-09

Characteristic	**Measure / definition** **Proportions in age 7 MCS teacher sample (unweighted)**	**Measure / definition** **Proportions in school population according to Department for Children, Schools and Families statistics for pupils in 2008-09**
Gender	Parent-report in survey **50.2%** male	Statistics for pupils who were age 7 in January 2009 **51.2%** male
Ethnicity	Parent report in survey / derived variable **80.7%** White 3.3% Indian 6.9% Pakistani / Bangladeshi 3.9% Black 1.7% ‘Other’ 3.4% Mixed ethnicity	Statistics for all primary pupils **79.2%** White 2.5 % Indian 5.5% Pakistani / Bangladeshi 4.9 % Black 3.8% ‘Other’ 4.1% Mixed ethnicity
English as an additional language	Parent report in survey: response to question on, “language spoken in household” **86.3%** “English only”	“First language is known or believed to be English” – statistic for all primary pupils **84.6%** English first language
Diagnosed / recognised with special educational needs (SEN)	Teacher report in survey: response to question, “Has this child EVER been recognised as having SEN?” **22.6%** “yes” Teacher report in survey: response to question, “Does this child have a full statement of SEN?” **2.6%**	Pupils in Year Two in 2008-09 **21.8%** with any SEN recorded Pupils in Year Two in 2008-09 **1.3%**

Can any bias in the teacher survey sample accurately be accounted for, or corrected?

This high-level correspondence provides some indication that the sample is not obviously biased away from the 2008 population of Year Two school children. However, it should also be noted that this is not the population from whom the sample were selected (see Plewis *et al*, 2007, for sample definition), and that, as noted, over waves of the MCS there is significant drop-out (Plewis, 2007; Johnson *et al*, 2011). Thus Mostapha (2013) recommends that analyses using the teacher survey attenuate estimates according both to the design weights that account for the MCS's original, known, disproportionate sampling strategy at wave one, and the attrition / non-response weights for wave four that attempt to compensate for disproportionalities in propensity to respond according to observed characteristics among participants.

There are at least four key pertinent issues that call into question the extent to which weighting in this way can accurately and reliably render the diminished teacher sample representative of its target population, given particularly that construction of the wave four attrition weights relies upon decisions regarding inclusion and testing of available measured characteristics. The weights are based upon a number of family-level variables: mother's age; main parent's education level; initial sample strata lived in; whether consent is given to data linkage; sweep in which family entered study; residential movement between waves; whether or not the cohort child was breastfed; child's ethnicity; child's gender; working status of main parent; housing tenure; type of accommodation; whether income information is provided; whether a single / dual parent family (Ketende [Ed.], 2010). However, adjustments according to propensity to respond based only on these limited characteristics may potentially produce erroneous findings – firstly, because their use assumes that there is homogeneity of participant response within each sub-cell of modelled factors.

Homogeneity of response

That is, there must be no overall difference in any of the surveyed-collected information that is provided / would be provided among sub-groups of responders compared to non-responders with the given modelled characteristics. Responses / potential responses among all families of mothers of a certain age, who have, for example, no formal qualifications, who were born in a ‘deprived’ area of England, and so on, must even out as equivalent among those who participate and those who do not, if up-weighting the responses of the participants with the modelled characteristics that predict attrition can represent accurately those participants who have failed to respond. This assumption may not hold if families who drop out would have provided information that differs from the data generated by those who remained in the sample.

Extent to which weights predict attrition

Secondly, even if homogeneity of response is assumed, the weights developed for wave four of the MCS can account for only a small proportion of attrition. As well as showing that decisions regarding factors to include influence precision and estimates when using the attrition models, Plewis *et al* (2010) suggest that there are unobserved and unmodelled factors that predict non-response to the MCS wave four survey:

> ...despite using a wide range of explanatory variables, discrimination [between responders / non-responders at wave four of the MCS] is on the low side...[a] possible implication is that the models do not discriminate well because data are missing not at random... it might not be generally possible to predict which cases will become non-respondents with a high degree of accuracy. (ibid, p.14–15)

Lack of teacher sample-specific weights

Thirdly, it should be noted that weights specifically for the teacher sample have not yet been developed. Weights are available only to the level of the main survey at wave four (contingency four in Table 2.1), so even notwithstanding the issues raised above, their use in no way offers an unproblematic or simple means by which this sample may confidently be rendered representative of English children born at the turn of the century.

As described, only 63 percent of the main wave four English sample have teacher participation, so the weights do not account for the non-response of 37 percent of children's teachers.

Mostapha (2013) argues nonetheless that the best overall solution to attrition to the teacher survey is to weight to the last contingency possible using the wave four main survey weights:

> In order to have unbiased results when using the teacher survey, one therefore has to use the sample-design and non-response weights in order to adjust for stratification, clustering and attrition (p 6-7)… it is clear that researchers should use the design and MCS non-response weights when undertaking statistical analyses with the teacher survey in order to avoid biased estimates of cohort member characteristics (ibid, p.8).

However, the examples presented by Mostapha (2013) alongside this contention do not entirely support its argument. For example, in the main wave four (weighted) sample, 87.3 percent of pupils are reported as being of White ethnicity. In the unweighted teacher sample, 88.3 percent are reported as White, but this rises to 89.7 percent when wave four main survey weights are applied to the teacher sample – biasing estimates for this characteristic away from the main sample. Similarly, average weekly income is reported as £565 in the main weighted wave four sample, as £582 in the unweighted teacher sample, and as £590 in the weighted teacher sample (see Mostapha, 2013: Table 3; Table 4).

Reassuringly, however, these examples also indicate that use / omission of weights does not make an enormous difference to proportions. Again, this provides some suggestion that the teacher sample is not excessively skewed from the population – at least, not according to the observed characteristics which make up the weights.

Clustering within schools

The recommendation that MCS weights be used for the teacher sample is problematic for one last reason: it prohibits clustering by school or by teacher, because nesting is one of the survey's design features, and children are clustered according to the local area in which they were born. There is some clustering of pupils within schools at wave four. Though not enough to allow robust comparison of within-school and between-school drivers (the 5,481 core teacher survey children attend 2,700 schools in 154 local authorities) it is desirable to take this factor into account in some analyses, in case school-level factors other than those under investigation are driving results.

A pragmatic, exploratory approach

Given the imperfection of the teacher survey as a population sample, the various considerations and contentions raised above, and the lack of an unproblematic solution, this thesis takes an open pragmatic and exploratory approach. It does not presume that the MCS wave four teacher sample is fully representative of Year Two children in England in 2008, but it recognises the large sample as a highly useful resource within which to test theories and begin to indicate patterns.

Throughout the chapters, the sensitivity of results to different weighting / non-weighting / clustering specifications is therefore tested, as appropriate and practicable, given the research questions and sample used in each section. Some additional sample comparisons and descriptions are also made, in order to situate findings. For the most part, in practice, results are not sensitive to the various specifications; nor do further comparisons yield major differences. This provides some indication that the MCS wave four teacher survey is in fact, as hoped, a reasonably robust sample, and that it is certainly an appropriate one with which to describe patterns, test theories, and make recommendations.

Chapter 3

The influence of stream placement on teacher judgements and assessments of pupils

Introduction

Streaming, the practice of grouping all pupils within a cohort according to a measure or conception of overall 'ability,' was widespread in England in the early 20th century. Having been consigned to a relatively higher or lower position, pupils spent at least the majority of their lessons being taught at the level deemed 'appropriate' to their allocated group. But, over time, alongside the reform to comprehensive education, streaming became gradually less common, and was extremely rare in primary schools by the 1990s (Blatchford *et al*, 2010; Hallam & Parsons, 2013).

Reversing this trend, however, the past two decades have seen a government-prescribed and sanctioned push back towards various forms of ability grouping (Boaler, 1997; Conservative Party, 2007; Department for Children, Schools, and Families, 2008b; Department for Education, 1992; Department for Education, 2010a; Department for Education and Skills, 2005). Underpinned by political and philosophical assumptions of innate and immutable differences in fundamental ability and potential (Department for Education, 1992, p 12; Department for Education and Skills, 2005, p 20), this has corresponded to a resurgence of streaming among primary school pupils as young as seven years old. In the space of a decade, estimates of the prevalence of the practice have grown from less than 2 percent of all primary pupils in 1999 (Hallam *et al*, 2003) to nearly 18 percent of Year Two pupils in 2008 (Campbell, 2013a).[3]

This resurrection of streaming among young children in England appears either to be politically 'accidental,' or to be ideologically driven – or both – given that the majority of available evidence indicates that early grouping

[3] Many more pupils are also ability grouped in-class, or for individual subjects like literacy and numeracy (Campbell, 2013a).

neither raises overall average attainment nor leads to greater parity or equality of opportunity or achievement (Slavin, 1990; Sutton Trust / Educational Endowment Foundation, 2014). International research by the OECD has suggested that '[e]arly student selection has a negative impact on students assigned to lower [streams] and exacerbates inequities, without raising average performance,' and recommends that 'selection should be deferred to upper secondary education while reinforcing comprehensive schooling' (OECD, 2012, p.10). Reviewing a mostly British literature, Kutnick *et al* (2005) conclude that, '[pupil ability groupings] appear to have replicated the achievement spectrum that they were designed to reduce' (p.12), while Dunne *et al* (2007) update previous syntheses of the research, and conclude that grouping is 'disadvantageous for those in lower sets and increases the overall attainment gap' (p.8).

The body of evidence on streaming is reasonably robust and persistent. In their latest (2016) government-funded review, the Educational Endowment Foundation deem their findings 'moderately' reliable, and conclude that:

> The evidence on...streaming is fairly consistent and has accumulated over at least 30 years of research. Although there is some variation depending on methods and research design, conclusions on the impact of ability grouping are relatively consistent (https://educationendowmentfoundation.org.uk/evidence/teaching-learning-toolkit/setting-or-streaming/).

As well as raising questions regarding the theoretical and empirical bases of the assumptions behind streaming, the cannon of evidence has, regularly, demonstrated inequalities in 'ability' grouping placement which only reflect wider educational and societal disparities in opportunity, achievement and outcomes (Ansalone, 2003; Boaler, 1997; Boaler *et al*, 2000; Kutnick *et al*, 2005, Wiliam & Bartholomew, 2004). The most recent evidence on prevalence and patterns within the UK suggests that, even controlling for prior measures of academic aptitude and attainment, low-income primary school pupils are disproportionately often placed in the lowest streams, along with boys, pupils who are relatively younger within their school year (in England, summer-borns), and children with less educated parents. There are also some indications of disproportionality by ethnicity (Hallam & Parsons, 2013). That these inequalities exist even after taking account of manifest educational performance indicates that factors other than any kind of measure of 'ability' are influencing the stream to which each child is allocated, and that the process of streaming may not, therefore, be 'fair.' Studies suggest moreover that teacher perceptions of pupils' behaviour, rather than any indication of their academic aptitude, may at times be influential in determining stream placement (Boaler, 1997; Blatchford *et al*, 2010).

Given disparities in placement according to pupil characteristics, the evidence that streaming can be particularly 'disadvantageous for those in lower sets' is especially troubling. Streaming, it seems, might provide an educational structure which, rather than alleviating between-group differences, could be the very origin of some of these differences – or which might serve at least to embed and over-extrapolate them, and potentially to widen their magnitude.

Research has suggested several mechanisms through which streaming might be instrumental in creating, entrenching or amplifying inequalities. Studies indicate firstly that pupils' own self-concept, perceptions and behaviours can be influenced by the group to which they are assigned (Ansalone, 2003; Boaler, 1997; Croizet & Claire, 1998; Kutnick *et al*, 2005; Raey, 2006; Shih *et al*, 2005; Steele & Aronson, 1995; Yopyk *et al*, 2005). There is evidence that being placed in a higher stream may lead to positive self-expectations and mind-sets, while being placed in a lower group can result in demotivation and 'anti-school attitudes' – and that these processes lead to relatively higher and lower attainment (Kutnick *et al*, 2005).

Secondly, research proposes that educational opportunities and quality of teaching can differ according to stream placement, with the progress of children in upper groups being facilitated to a higher level than those placed at the bottom of the hierarchy (Ansalone, 2003; Boaler, 1997; Kutnick *et al*, 2005). As there is also some evidence that movement between stream placements may be rare once positions have been established (Blatchford *et* *al*, 2010; Hallam & Parsons, 2013), this means that some pupils' trajectory of opportunity may be determined by and strongly premised upon their early allocation to a given stream.

Lastly, studies indicate that stream placement may influence the perceptions and expectations class teachers hold of their pupils. Research suggests that teachers (consciously or unconsciously) label and stereotype children based on a variety of characteristics (Burgess & Greaves, 2009; Campbell, 2013b; Hansen & Jones, 2011; Reaves *et al*, 2001; Thomas *et al*, 1998), and, in particular, there is evidence that teachers formulate and act upon expectations of pupils according to the level of their academic group placement (Ansalone, 2003; Boaler, 1997; Boaler *et al*, 2000; Ireson & Hallam, 1999; Rubie-Davies, 2010). Assigned stream level may therefore affect teacher perceptions of their whole class and of each pupil within the class.

This is crucial not least because there are well-established relationships between teacher perceptions and pupil attainment. From Rosenthal and colleagues' experimental research in the 1960s (Rosenthal & Jacobsen, 1968) to the present, a solid body of evidence has built which suggests that teacher beliefs, expectations and judgements regarding their pupils can influence pupils' achievement: 'when teachers believe… their students [are] very able [they interact] with them in ways which promote…their academic development' (Rubie-Davies, 2010; see also Brophy & Good, 1970; Good, 1987; Miller & Satchwell, 2006). As most academic achievement at the primary level is currently judged and assessed by teachers (Department for Education, 2014d), these processes and their influence on pupils' progress are more important than ever.

The current chapter

Teacher perceptions, judgements and assessments are therefore the focus of this chapter. While some studies have explored the relationships between stream placement and teachers' views of pupils, most have been small-scale and qualitative, and explicit controls for the impact and mediation of the many factors and processes which may confound any direct associations between stream level and teacher judgements have been insufficient (Blatchford *et al*, 2010; Ireson & Hallam, 1999; Kutnick *et al*, 2006). There is a dearth of up-to-date UK research particularly in the primary sector – presumably due, in part, to the fact that the documented resurgence of streaming among young pupils has arisen fairly rapidly since the turn of the century (Hallam & Parsons, 2013), and that, subsequently, discussion of this specific ability grouping practice has returned to the research and public discourse in very recent years. Only lately have studies begun to exploit the potential of emerging data in identifying the possible effects of different ability grouping practices on pupil progress and attainment (Campbell, 2014; Hallam & Parsons, 2014).

The current chapter therefore uses contemporary large-scale survey data for a sample of English pupils in early primary school, and accounts for a broad variety of factors which may explain spurious apparent connections between stream placement and teacher perceptions. Controls include demonstrable pupil performance / aptitude, pupil, family and teacher characteristics, measures of pupil behaviour and teacher perceptions of behaviour, and prior attainment and assessment.

Analysis utilises two discrete groups of measures of teacher judgements – official, teacher-assessed Key Stage One test scores, and survey-reported perceptions – thereby examining whether any effect of streaming on judgment is sensitive to / an artefact of the situation and measure used, or holds steady across contexts and domains. By exploring the data using detailed regression modelling, analysis here hopes more definitely to isolate any true relationships, and to test the hypothesis that teacher assessments and judgements of pupils are influenced by the stream to which the pupil is allocated, thus contributing to attainment disparities.

Methodology

Sample and data

Data are derived from wave four of the Millennium Cohort Study (MCS), which took place in 2008 – please see Chapter 2 for a more detailed discussion of the children included. 914 (17.5 percent of the) seven-year-old English MCS state school pupils with teacher response are reported as being streamed, and data on stream placement itself is available for 882 singleton pupils within this group, who form the core sample for whom analyses are performed in this Chapter (see University of London 2008; 2011a; 2011b; 2012a; 2012b for data source references, and Annex I for questionnaire extract).

Only children who are streamed are included in analysis. No comparisons are made between children who are streamed and not streamed, in order to negate the possibility of differences between schools that chose to stream and not to stream confounding results. The 882 MCS sample pupils for whom stream placement information is available differ only minimally from those English, singleton, state school MCS children who are reported as not being streamed, according to a number of key characteristics (see Annex 3A) – suggesting that the sample of streamed pupils is a reasonable one within which to investigate the relationship between placement level and teacher perceptions.

In this Chapter, unless otherwise stated, all estimates are weighted for the MCS's design features and for attrition to the main wave four survey, as recommended by Mostapha (2013). However, as discussed in Chapter 2, this is not an unproblematic approach, so an alternative specification produces unweighted analyses with clustering of standard errors at the school-level (reported in results section). All analyses use Stata versions 12 and 13.

Outcome variables

Two separate sets of regression analyses are undertaken to examine the relationships between stream placement and teacher judgment, using two respective groups of outcome measures: officially recorded Key Stage One scores, which are entirely teacher-assessed, and perceptions of each pupil's 'ability and attainment' as reported by teachers during MCS surveying.

Outcome group one: Teacher-assessed Key Stage One scores

The first measures of teacher judgment used in analyses are the Key Stage One (KS1) scores allocated to each child. KS1 assessment takes place at age seven, at the end of Year Two. This is the year during which MCS surveying took place, and for which information on the pupils' stream placement is provided. In 2008, KS1 attainment was entirely teacher-assessed, so a pupil's recorded achievement at this stage is entirely dependent on the perceptions, judgements and decisions made by the respondent class teachers.[4] This outcome measure indicates whether stream placement is associated with teacher judgment when that judgment is required for official assessment, and whether stream placement has an influence on a pupil's publicly and permanently recorded 'achievement.'

Overall average point score (APS) at KS1 is used as the first outcome in this set of analyses, and attainment levels in reading and maths form the second and third. A pupil's APS is constructed from their teacher's judgements of performance across reading, writing, maths and science (equally weighted). In the sample used in this Chapter, scores range from 3 to 22.5 (mean = 15.9; SD = 3.4). The mean score for pupils found in the bottom stream is 11.2, for those in the middle stream: 14.5, and for those in the top stream: 18.5.

Children are allocated separate categorical reading / maths attainment levels by their class teachers, and possible levels (from lowest to highest) are: 'working towards level 1' (3.7 percent of sample children fall into this category for reading, and 2.3 percent for maths), 'achieved level 1' (19.9 percent / 12.3 percent), 'achieved level 2c' (15.7 percent / 20.0 percent), 'achieved level 2b' (28.1 percent / 31.2 percent) 'achieved level 2a' (32.6 percent / 34.2 percent).[5]

[4] See http://nationalpupildatabase.wikispaces.com/KS1 and http://www.bristol.ac.uk/cmpo/plug/support-docs/ks1userguide2011.pdf for further detail on KS1 assessment and scoring.

[5] More detailed

Three respective KS1 outcome variables are therefore investigated, using the following regression techniques:

1. Average point score (range: 3-22.5) – modelled using linear regression.
2. Reading attainment level (scale: ‘working towards level 1,’ achieved level 1,’ ‘achieved level 2c,’ ‘achieved level 2b,’ ‘achieved level 2a’) – modelled using ordered probit regression.
3. Maths attainment level (scale: ‘working towards level 1,’ achieved level 1,’ ‘achieved level 2c,’ ‘achieved level 2b,’ ‘achieved level 2a’) – modelled using ordered probit regression.

Outcome group two: Survey-reported teacher judgements

During the MCS teacher survey, respondents were asked to ‘rate…the study child's ability and attainment…in relation to all children of this age’ (see Annex II), and these judgements are used as a test of consistency as well as a measure of the effects of stream placement on more ‘everyday’ perceptions not directly required for official assessment. Teachers could choose to judge that a pupil was: ‘well above average,’ ‘above average,’ ‘average,’ ‘below average,’ or ‘well below average.’ Ratings were recorded for teacher perceptions of the child’s ‘ability and attainment’ across seven domains: speaking and listening / reading / writing / science / maths and numeracy / physical education / information and communication technology / expressive and creative arts.

In some analyses in this Chapter, the seven-sub-responses are each allocated a score of one to five (where one represents ‘well below average’ and five ‘well above average’), and summed to represent one ‘overall’ rating, ranging from 7-35 (mean = 22; SD = 5.3; top stream pupils’ mean = 26; middle stream pupils’ mean = 21; bottom stream pupils’ mean = 16). This is intended to measure each teacher’s general judgment of pupil ability (analysis using this outcome is modelled using linear regression).

descriptive statistics by stream placement are not available for this outcome, because small cell sizes prohibit release of this analysis by the Secure Data Service, within whose secure remote desktop the parts of this Chapter using KS1 scores were performed.

Among the 851 sample pupils with data on both stream placement and survey-reported teacher judgements, responses for each domain are, in the main, highly correlated with this overall summed total (see Table 3.1). Judgements of ability in physical education and in arts are less strongly related to the total and to judgements in each other subject, suggesting some delineation between teacher perceptions of performance in 'academic' and 'non-academic' domains. Therefore, the summed total including all subjects is used for the main analysis, but additional sensitivity checks excluding judgements on physical education and arts are also carried out (using the five remaining domains; scale 5-25).

Table 3.1: Correlations between summed teacher judgment and judgements in each individual domain

	Overall ability	Reading ability	Writing ability	Science ability	Maths ability	PE ability	ICT ability	Arts ability
Overall ability	1.00							
Reading ability	0.90	1.00						
Writing ability	0.91	0.87	1.00					
Science ability	0.90	0.78	0.78	1.00				
Maths ability	0.89	0.80	0.80	0.80	1.00			
PE Ability	0.66	0.42	0.47	0.51	0.48	1.00		
ICT ability	0.84	0.68	0.68	0.73	0.70	0.60	1.00	
Arts ability	0.74	0.56	0.59	0.60	0.52	0.57	0.62	1.00

Ns = 851-871 (unweighted; sample limited to those pupils with complete information on stream placement). All estimates weighted for survey design and attrition to main wave four survey.

Further analyses are performed separately for judgements of reading and of maths ability, respectively (here, the scale is 1-5), using ordered probit modelling. Three main survey-reported teacher judgements of 'ability and attainment' are therefore used as outcomes:

1. Aggregated overall judgment (range: 7-35) – modelled using linear regression.
2. Judgment of reading ability (range: 1-5) – modelled using ordered probit regression.

3. Judgment of maths ability (range: 1-5) – modelled using ordered probit regression.

Key predictor variable: stream placement

The key predictor in modelling against all outcomes is pupil's stream placement (top, middle, or bottom), as reported by their teacher. Streaming is defined in the teacher questionnaire as 'group[ing] children in the same year by general ability and they are taught in these groups for most or all lessons.' In the sample of 851 pupils with data on both teacher survey judgment and stream placement, 41 percent are reported as being in the top stream, 31 percent in the middle stream, and 28 percent in the bottom stream. 17.2 percent of the slightly smaller sample of MCS pupils with data on KS1 scores and with teacher response regarding whether streamed are reported to be subject to the practice (a comparable proportion to that reported for the main survey sample). A subsample of 651 Year Two pupils in mainstream (i.e. non-special) schools have data on both stream placement itself and KS1 scores, and 45 percent are reported to be in the top stream, 31 percent in the middle stream, and 24 percent in the bottom stream.[6]

Key controls: cognitive test scores

Shortly before children's teachers were contacted for their survey, the seven-year-old MCS pupils were visited in their homes by interviewers who administered three separate cognitive ability tests. The mean lag between pupil cognitive tests and teacher survey was 3.8 months, the median 3 months, and the mode 2 months. Performance scores on these tests provide key counterpoint controls in modelling to teacher judgements, allowing analyses of whether children who perform equivalently, but who are placed in different streams, are judged differently by their teachers.

The first of the tests used is the British Ability Scales Word Reading test. This is designed to assess children's English reading ability (see

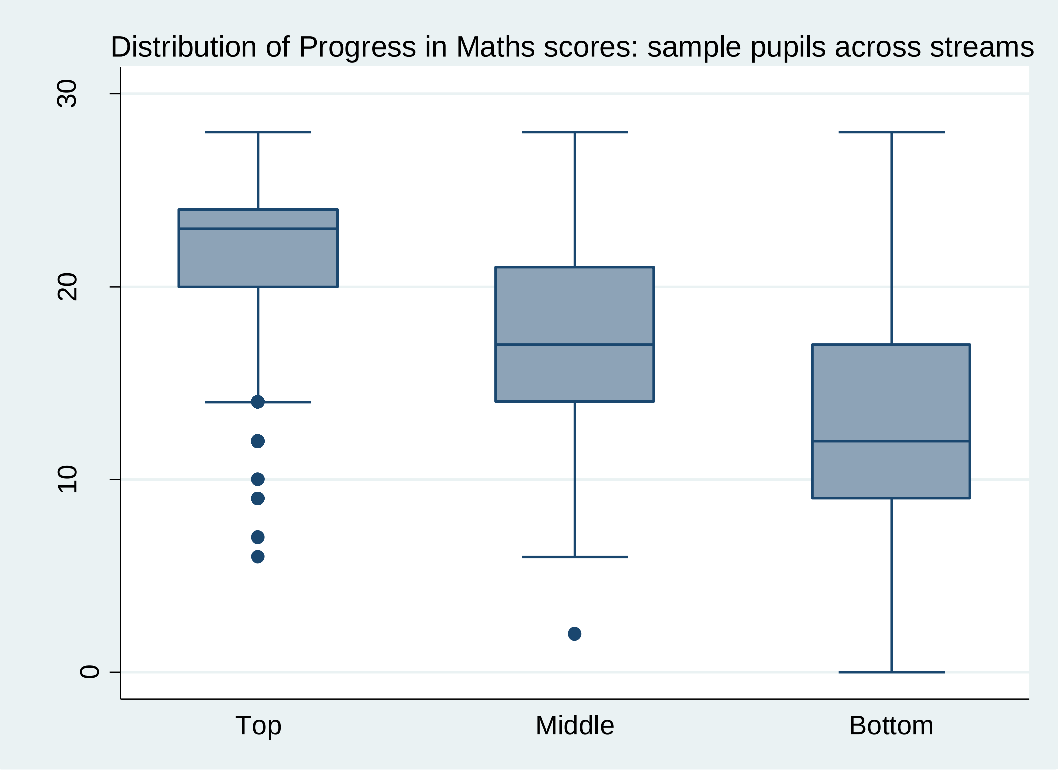
[6] The sample with KS1 scores is smaller than the survey sample due to factors such as lack of parental consent for linkage to educational records, and administrative failure in linkage to these records (see Johnson & Rosenberg, 2013).

http://www.glassessment.co.uk/products/bas3). The ability score (a scaled but not otherwise standardised score) is utilised (see Hansen, 2012). Secondly, performance on the Progress in Mathematics test is included. This test is designed to measure pupils' mathematical ability across use of numbers, shapes, and skill in data handling, and to provide an indication of performance in maths at the given developmental stage (see http://www.gl-assessment.co.uk/products/progress-maths). The shortened version used in the MCS entailed routing to sections of varying difficulty levels, and Rasch scaling was used to convert the raw scores to a count score equivalent to that which would be attained were the full test completed (see Hansen, 2012) and this scaled score is used. Lastly, scores on the British Ability Scales Pattern Construction Test are incorporated. The Pattern Construction Test has been developed to provide an indication of overall cognitive aptitude (http://www.gl-assessment.co.uk/products/bas3) and, as with the Word Reading Test, the ability score is used for modelling.

Scores for all three tests are used in as 'raw' a form as possible (weighted / scaled only for question difficulty / routing / selection), and are not otherwise standardised or modified. This means that each simply represents a child's performance as manifest in completing that particular test on the given day. Notwithstanding this, because children took the tests at slightly different ages within the MCS wave four fieldwork windows, and because the lags between tests and teacher survey / KS1 assessment vary slightly, both pupil age at cognitive tests and pupil age at teacher survey / age at KS1 assessment (here proxied by month of birth) are controlled for in all analyses, to ensure that these factors do not confound results.

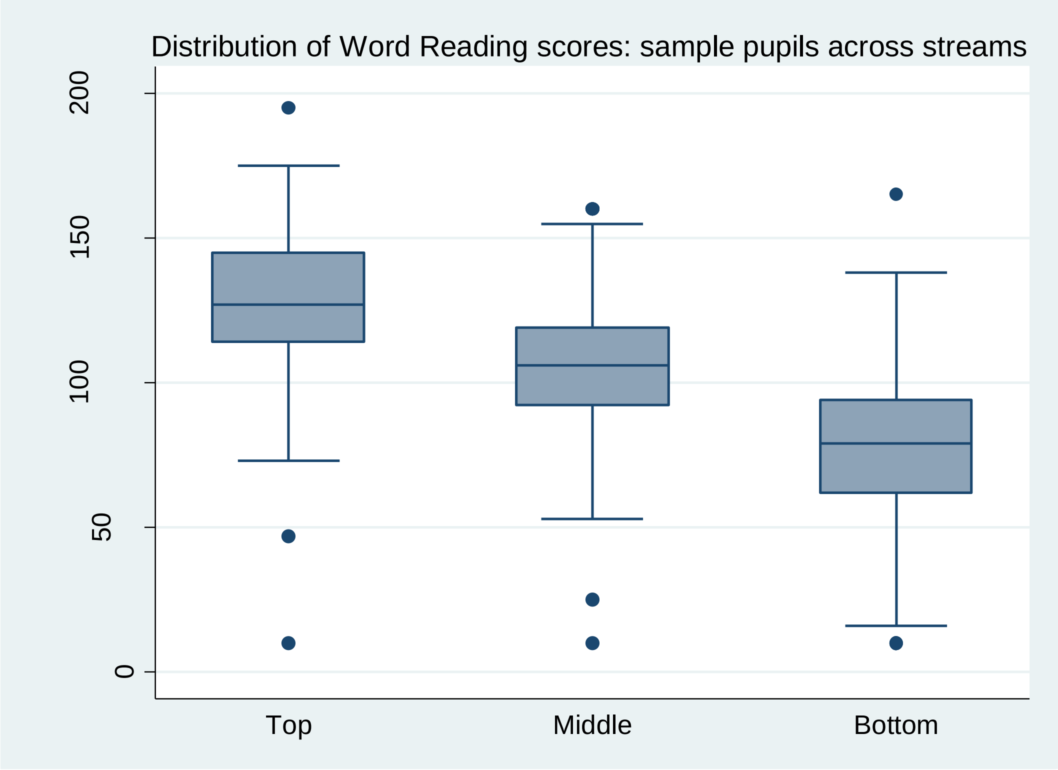
Figures 3.1, 3.2 and 3.3, below, illustrate the distribution of scores on the three cognitive tests for pupils situated in each stream, in the sample with survey-reported teacher judgements.

Figure 3.1:



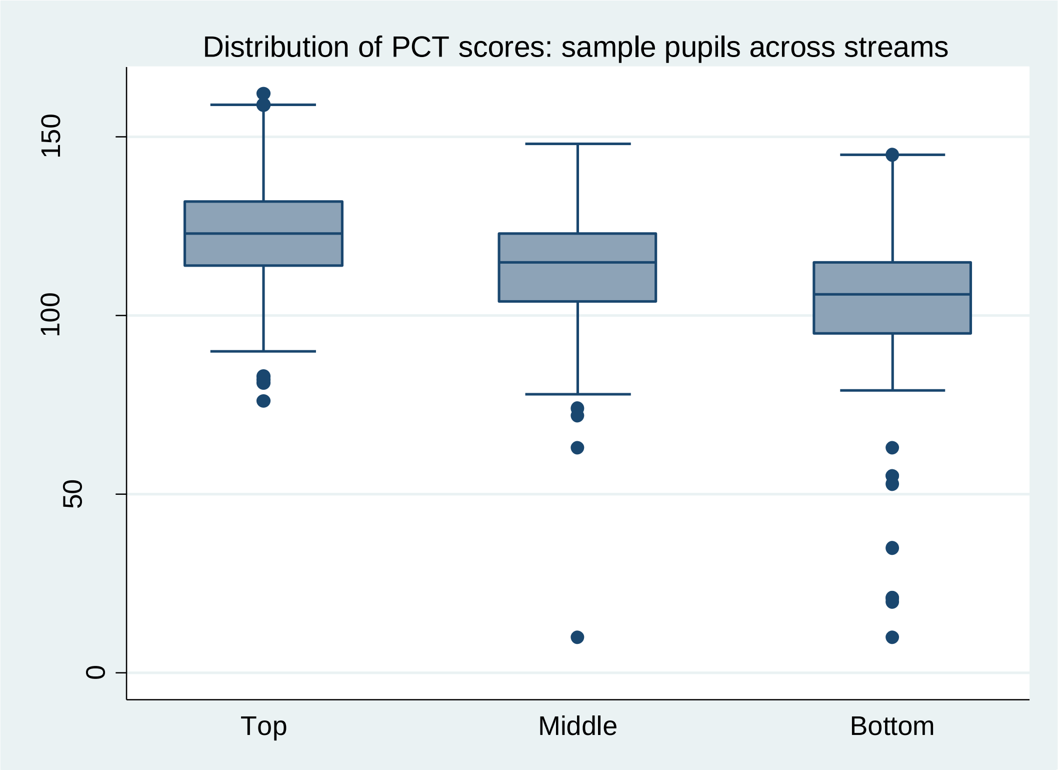
n = 840; Mean for all pupils = 18.2. Line represents median, box represents 25th and 75th percentiles (Q1 and Q3, respectively), whiskers represent Q3+1.5(Q3-Q1) / Q1-1.5*(Q3-Q1).

Figure 3.2:



n = 837; Mean for all pupils = 108.5. Line represents median, box represents 25th and 75th percentiles (Q1 and Q3, respectively), whiskers represent Q3+1.5(Q3-Q1) / Q1-1.5*(Q3-Q1).

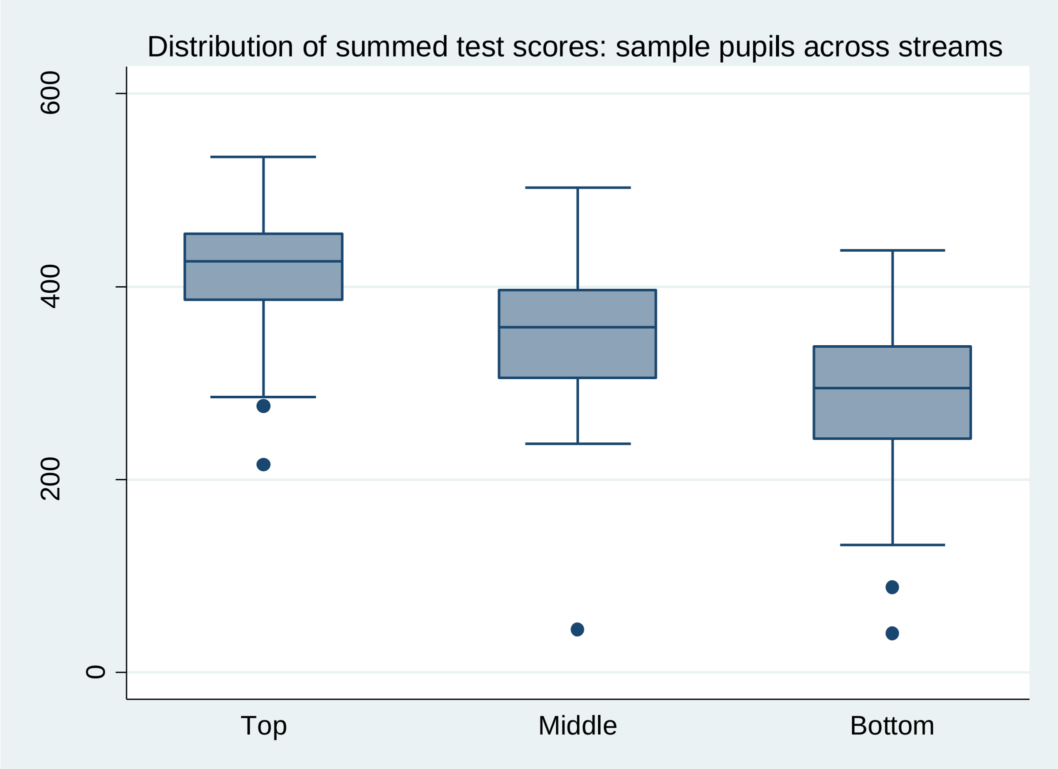
Figure 3.3:



n = 835; Mean for all pupils = 114.6. Line represents median, box represents 25th and 75th percentiles (Q1 and Q3, respectively), whiskers represent Q3+1.5(Q3-Q1) / Q1-1.5*(Q3-Q1).

While there is variation between streams, with pupils in the higher groups scoring better, on average, in all the tests, there is also an overlap between groups: some children who score equivalently on the cognitive tests are situated in different streams. Most overlap is apparent in PCT scores – particularly notable given that the PCT is intended to measure ‘overall’ cognitive ability, just as stream placement is intended to reflect ‘general’ ability across subjects. Figure 3.4, below, shows the distribution of each child’s combined cognitive test score across streams when the three scores are summed together and equally weighted to provide an alternative generalised representation of aptitude and performance. Again, there is an overlap of similarly-scoring children between streams. Annex 3B presents the equivalent information for pupils in the sub-sample with KS1 scores, and the same patterns hold for this group.

Figure 3.4:



n = 829; Mean for all pupils = 366.6. Line represents median, box represents 25th and 75th percentiles (Q1 and Q3, respectively), whiskers represent Q3+1.5(Q3-Q1) / Q1-1.5*(Q3-Q1).

Additional controls

There are inequalities according to pupil and family characteristics in stream placement level, and these characteristics may bias teacher perceptions, and / or stream placement itself. Therefore it is crucial to control for these and other potential confounders in modelling, in order to indicate any independent effect of streaming.

- Pupil and family characteristics

Table 3.2 illustrates distributions across streams according to key individual-level characteristics within the sample with teacher survey judgements, and shows, for example, that girls tend more often to be found in the higher stream, along with pupils relatively older within the school year, children from higher-income families, and those with more educationally qualified parents. Therefore analyses control for pupil gender, pupil birth month, family income-level, main parent's highest qualification level, and also for pupil ethnicity. This eliminates the possibility that children in the top stream are not, for example, judged more favourably by virtue of being in the top stream but because they are relatively more mature and developed, being disproportionally autumn-born.

Table 3.2: Percentage of sample pupils with each characteristic placed in each stream*

	Top stream	Middle stream	Bottom stream
All pupils (n = 882)	41	32	27
Boys (n = 461)	39	26	34
Girls (n = 421)	43	37	20
September-born (n = 79)	68	20	12
October-born (n = 73)	60	10	30
November-born (n = 74)	57	29	14
December-born (n = 92)	45	37	18
January-born (n = 85)	44	20	35
February-born (n = 51)	46	26	29
March-born (n = 76)	30	40	30
April-born (n = 59)	36	29	25
May-born (n = 68)	29	42	30
June-born (n = 95)	23	37	40
July-born (n = 69)	32	31	36
August-born (n =61)	25	46	29
White ethnicity (n = 671)	41	32	28
Mixed or 'other' ethnicity (n = 56)	44	26	30
Indian ethnicity (n = 36)	40	36	24
Pakistani or Bangladeshi ethnicity (n = 89)	43	39	18
Black or Black British ethnicity (n = 30)	41	24	36
Low-income (n = 267)	25	33	42
Higher-income (n = 615)	48	31	21
Parent NVQ level 1 (n = 72)	32	35	33
Parent NVQ level 2 (n = 258)	37	31	32
Parent NVQ level 3 (n = 138)	42	29	30
Parent NVQ level 4 (n = 228)	52	35	13
Parent NVQ level 5 (n = 45)	61	29	10
Parent overseas qualification only (n = 39)	43	32	25
Parent no qualifications (n = 102)	26	28	46

*All estimates weighted for survey design and attrition to main wave four survey. Ns are unweighted

- Behaviour and perceptions of behaviour

As mentioned, research has also suggested that stream placement may be determined by pupil behaviour rather than by ability, performance, or attainment, as well as indicating a correspondence between teacher perceptions of children's behaviour and of their academic ability (Brown & Sherbenou, 1981; Strand, 2007). Table 3.3 shows mean total difficulties scores for the Strengths and Difficulties questionnaire (SDQ; see http://www.sdqinfo.com) as completed by sample children's parents at age five, a time preceding stream placement for the academic year of interest. The SDQ is intended to measure manifest problematic behaviours, so the measure taken at this prior time should pick up on any strong, enduring, non-situation-dependent behavioural tendencies which might have affected the stream in which a pupil was subsequently placed. Correspondingly, Table 3 indicates that children who were eventually situated in the bottom stream at age seven were, on average, rated more highly by their parents at age five for emotional symptoms, conduct problems, hyperactivity, and peer problems – and received a lower score for pro-social behaviour. In order, therefore, to disentangle any resultant association between pupil behaviour, stream placement, and teacher perceptions, scores for each of the sub-scales of this age five parent-assessed SDQ are used as controls in modelling.

Table 3.3: Mean score on each scale of age five parent-completed SDQ test*

	Top stream	**Middle stream**	**Bottom stream**	**All streams**
Emotional symptoms^ (n = 799)	1.3	1.4	1.8	1.5
Conduct problems^ (n = 802)	1.3	1.7	2.3	1.7
Hyperactivity^ (n = 795)	2.9	3.6	5.0	3.6
Peer problems^ (n = 801)	1.0	1.2	1.7	1.3
Pro-social behaviour^^ (n = 802)	8.3	8.4	7.7	8.2

*All estimates weighted for survey design and attrition to main wave four survey. Ns are unweighted

^Range = 1-10. Higher score is 'worse' and represents more problematic behaviours and fewer 'desirable' behaviours. ^^Range 1-10. Higher score is 'better' and represents more pro-social behaviours.

In line with the possibility that teachers' contemporaneous perceptions of pupil behaviour may influence their perceptions of pupil ability, Table 3.4 shows the distribution across streams of teacher-assessed SDQ scorings at age seven, measured during the same survey within which judgements of ability were provided. There is an evident tendency of pupils in the bottom stream to be rated as displaying more problematic and fewer pro-social behaviours (and vice versa for the top stream), so it is possible that these perceptions of behaviour, rather than stream placement itself, are driving any differences in teacher perceptions of ability differentiated by stream. To control for this, modelling adds the five subscale scores of the teacher-assessed SDQ at age seven, as well as responses to a general follow-up question asking teachers: 'Overall, to summarise, do you think that this child has difficulties in one or more of the following areas: emotions, concentration, behaviour or being able to get on with other people?' (Table 3.5 shows a pattern where problems are more likely to be reported for children in the bottom stream.)

Table 3.4: Mean score on each scale of age seven teacher-completed SDQ test*

	Top stream	**Middle stream**	**Bottom stream**	**All streams**
Emotional symptoms^ (n = 882)	1.4	1.8	2.2	1.7
Conduct problems^ (n = 882)	0.6	0.8	1.6	0.9
Hyperactivity^ (n = 882)	1.7	3.3	5.4	3.2
Peer problems^ (n = 882)	1.0	1.3	2.1	1.4
Pro-social behaviour^^ (n = 882)	8.3	7.6	6.4	7.6

*All estimates weighted for survey design and attrition to main wave four survey. Ns are unweighted

^Range = 1-10. Higher score is 'worse' and represents more problematic behaviours and fewer 'desirable' behaviours. ^^Range 1-10. Higher score is 'better' and represents more pro-social behaviours.

Table 3.5: Teacher report of whether pupil has overall difficulties: percentage with each response in each stream*

	Top stream	Middle stream	Bottom stream	All streams
No difficulties	83	68	26	63
Yes – minor difficulties	11	24	39	23
Yes – definite difficulties	4	7	28	12
Yes – severe difficulties	2	1	7	3

*All estimates weighted for survey design and attrition to main wave four survey. N = 875 and is unweighted

- Prior assessment / attainment: Foundation Stage Profile

Teacher perceptions of pupils may also be influenced by what they know about the pupil's prior attainment, and by judgements made and conveyed by other staff within their school. In addition, prior attainment / judgements may have been influential in determining the stream to which a child is allocated. Table 3.6 indicates a correspondence between Foundation Stage Profile (FSP) score, assigned two years previously, by the class teachers who taught the pupils' reception groups when they were five, and stream placement at age seven. Modelling therefore controls for this score. Inclusion of the FSP assessment also picks up, to some extent, on any academic and cognitive skills not already proxied by the three cognitive tests - albeit as assessed and developing two years previously.

Table 3.6: Mean total FSP score at age five*

	Top stream	Middle stream	Bottom stream	All streams
FSP total score (range 0-117)	98.1	83.6	69.1	86.0

*All estimates weighted for survey design and attrition to main wave four survey. N = 774 and is unweighted

- Special educational needs diagnosis

Modelling controls additionally for teacher report of whether each child has ever had any level of recognised special educational need (SEN). Table 3.7 shows a strong relationship in the sample between being reported to have a special need and placement in the bottom stream, so inclusion of this factor accounts for the possibility that SEN status might influence stream placement, teacher judgment (as suggested by Campbell, 2013b), or both. If stream placement remains significantly associated with judgment, having controlled for pupil and family characteristics, for perceptions of pupil behaviour, for prior attainment, and for SEN status, this will strongly support the hypothesis that the stream in which a pupil is placed has an independent effect on their teacher's perceptions and judgements.

Table 3.7: Teacher report of whether pupil has ever been recognised with SEN: percentage with each response in each stream*

	Top stream	Middle stream	Bottom stream	All streams
Yes	8	19	72	29
Don't know	0	1	0	1
No	92	80	27	70

*All estimates weighted for survey design and attrition to main wave four survey. N = 774 and is unweighted

- Teacher characteristics

Lastly, because research suggests that different streams of pupils may tend to be taught by teachers with different characteristics (Kutnick *et al*, 2005), modelling controls for some of these characteristics, so far as the data available allow. Teacher gender, total years teaching, and years spent teaching at current school are included. Table 3.8 indicates some possible disproportionalities across sample pupils. Though, overall, patterns are not easily interpretable, inclusion of these controls accounts for any mediating influence they may have on the relationship between stream placement and teacher judgment.

Table 3.8: Percentage of sample pupils in each stream taught by teachers with each characteristic*[7]

	Top stream	Middle stream	Bottom stream	All streams
Female teachers (n = 496)	91	93	94	93
Male teachers (n = 40)	9	7	6	7
Teacher taught for 24-48 years (60)	12	13	7	11
Teacher taught for 14-23 years (106)	18	22	27	22
Teacher taught for 8-13 years (87)	16	18	20	18
Teacher taught for 4-7 years (133)	29	21	28	26
Teacher taught for 1-3 years (199)	24	25	18	23
Taught at school for 8-48 years (148)	28	27	30	28
Taught at school for 4-7 years (159)	36	26	37	33
Taught at school for 1-3 years (199)	35	47	33	39

*All estimates weighted for survey design and attrition to main wave four survey. Ns are unweighted.

Modelling

All analyses combine the key predictor variable (stream placement) with both the key controls (cognitive test scores) and the additional controls detailed above, and regress these predictors on each of the six measures of teacher judgment (KS1-assessed / survey-reported). Controls are added through cumulative model specifications, and Table 3.9, below, describes each specification for analyses where survey-reported judgements form the outcomes. Table 3.10 describes variables added at each stage when KS1 assessments form the outcomes. Controls differ minimally for this outcome (due to availability in the respective datasets).[8]

[7] There is substantial missing data on this section of the teacher survey – this is accounted for in modelling with the inclusion of a 'missing' category in order not to lose cases. See e.g. Annex 3C.

[8] The KS1 outcomes are accessed through linked MCS – National Pupil Database (NPD) data, and this includes measures of school-type and year in which pupil joined school, which may potentially have some bearing on stream practices and implementation / judgements of pupils – so these factors are included in KS1 analyses.

Table 3.9: Cumulative specifications for models with survey-reported teacher judgements as outcomes

Specification	Predictors	Outcome
One	Stream placement Maths Test score Reading Test score Pattern Construction Test score Age at cognitive tests Age at teacher survey	Survey-reported teacher judgements of ‘ability and attainment,’ summed (range 7-35; linear regression) **or** Survey-reported teacher judgment of maths ‘ability and attainment’ (range 1-5; ordered probit regression) **or** Survey-reported teacher judgment of reading ‘ability and attainment’ (range 1-5; ordered probit regression)
Two adds…	Pupil gender Pupil month of birth Pupil ethnicity Pupil’s family’s income-level Pupil’s main parent’s highest qualification (age 7)	
Three adds…	Age 5 parent SDQ: emotional Age 5 parent SDQ: conduct Age 5 parent SDQ: hyperactivity Age 5 parent SDQ: peer Age 5 parent SDQ: pro-social Age 7 teacher SDQ: emotional Age 7 teacher SDQ: conduct Age 7 teacher SDQ: hyperactivity Age 7 teacher SDQ: peer Age 7 teacher SDQ: pro-social Teacher overall judgment of pupil behaviour	
Four adds…	Foundation Stage Profile total score	
Five adds…	Any diagnosis of special educational need	
Six adds…	Teacher gender Teacher years teaching Teacher years teaching at this school	

Table 3.10: Cumulative specifications for models with Key Stage One assessments as outcomes

Specification	Predictors	Outcome
One	Stream placement Maths Test score Reading Test score Pattern Construction Test score Age at cognitive tests Month of birth	KS1 Average point score (range: 3-22.5; linear regression) **or** Reading attainment level (scale: 'working towards level 1,' achieved level 1,' 'achieved level 2c,' 'achieved level 2b,' 'achieved level 2a'; ordered probit regression) **or** Maths attainment level (scale: 'working towards level 1,' achieved level 1,' 'achieved level 2c,' 'achieved level 2b,' 'achieved level 2a'; ordered probit regression)
Two adds…	Pupil gender Pupil ethnicity Pupil's family's income level Pupil's main parent's highest qualification (age 7) School-type Whether pupil joined in Year Two Whether pupil joined in Year One	
Three adds…	Age 5 parent SDQ: emotional Age 5 parent SDQ: conduct Age 5 parent SDQ: hyperactivity Age 5 parent SDQ: peer Age 5 parent SDQ: pro-social Age 7 teacher SDQ: emotional Age 7 teacher SDQ: conduct Age 7 teacher SDQ: hyperactivity Age 7 teacher SDQ: peer Age 7 teacher SDQ: pro-social Teacher overall judgment of pupil behaviour	
Four adds…	Foundation Stage Profile total score	
Five adds…	Any diagnosis of special educational need	
Six adds…	Teacher gender Teacher years teaching Teacher years teaching at this school	

Chronology and assumptions behind modelling strategy

For modelling truly to reveal any directional relationship from stream placement to teacher judgment, and to rule out the possibility of reverse causality, it is necessary firstly that stream placement should precede teacher judgment, and secondly that the judging teacher should not have been instrumental in determining placement. That the first is the case rests on an assumption that cohort-wide stream placement would have been established at the beginning of Year Two, and altered little in the year that followed, before teacher judgment was provided during the teacher survey (which took place during and mostly towards the end of the academic year [Huang & Gatenby, 2010]) and before KS1 assessments, which took place at the end of that year.

In analyses where the outcome is survey-reported teacher judgment, therefore, teachers participating in the MCS are assumed to provide details of each child's already-established stream placement which, crucially, has preceded their judgment of the child as provided in the same questionnaire. In analysis using KS1 results as the outcome, the minority of cases where fieldwork spilled over into Year Three are removed from the sample, to ensure that information only on stream placements in the year cumulating in KS1 assessments is included.

The second supposition, that the respondent class teacher who provides KS1 assessment / judgment should not have allocated the MCS pupil to their stream placement, is suggested both by the nature of streaming itself and by (admittedly slightly dated) reviews of evidence on school organisational practices. As streaming takes place at the whole-year level, placement may be officially determined by some combination of performance in previous years, formal assessments by previous years' teachers, pre-established placements, and / or school-based test performance (Blatchford *et al*, 2010; Kutnick *et al*, 2005; 2006) – and, as evidenced in the previous sections, drivers other than the officially stated seem also to be tacitly influential. Once streams have been decided upon, each set of pupils may be allocated to one of the year group's assigned class teachers – meaning that this teacher is unlikely to be heavily involved in the allocations themselves. (Note that this contrasts with the probable processes behind other types of ability grouping, such as within-class grouping, where the class teacher is likely to be a key decision-maker – this practice is discussed in more detail in Chapter 5)

Results: Stream placement and KS1 scores

Table 3.11, below, presents key results for each specification of the model where KS1 Average Points Score (range: 3-22.5; SD: 3.4) is the outcome (see Annex 3C for estimates for all modelled covariates). Even controlling for cognitive test scores and the full range of potentially confounding variables, pupils in the top stream are awarded significantly higher and pupils in the bottom stream significantly lower teacher-assessed scores at KS1. At specification six, children in the top stream are awarded scores 1.2 points (35 percent of a standard deviation) higher than those in the middle stream (p < .001), and children in the bottom stream scores 1.3 points (38 percent of a standard deviation) lower (p <.001).

Table 3.11: Difference in teacher-assessed Key Stage One average point score according to pupils' stream placement^ ^^

	Spec 1	Spec 2	Spec 3	Spec 4	Spec 5	Spec 6
Top stream	1.335***	1.371***	1.375***	1.229***	1.230***	1.209***
	(0.208)	(0.210)	(0.193)	(0.198)	(0.199)	(0.198)
(Middle stream)	0	0	0	0	0	0
	(.)	(.)	(.)	(.)	(.)	(.)
Bottom stream	-1.677***	-1.586***	-1.395***	-1.376***	-1.275***	-1.266***
	(0.234)	(0.231)	(0.238)	(0.236)	(0.250)	(0.255)
Maths Test score	0.101***	0.0963***	0.0816***	0.0779***	0.0755***	0.0781***
	(0.018)	(0.019)	(0.017)	(0.016)	(0.016)	(0.016)
Word Reading Test score	0.0520***	0.0498***	0.0488***	0.0470***	0.0462***	0.0458***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Pattern Construction Test score	0.0256***	0.0240***	0.0203***	0.0201***	0.0206***	0.0198***
	(0.006)	(0.006)	(0.005)	(0.005)	(0.005)	(0.005)
Constant	15.99**	16.11**	18.72***	18.23***	18.44***	17.78***
	(5.045)	(5.327)	(5.198)	(5.169)	(5.049)	(5.037)
N	639	639	635	635	635	635
R^2	0.799	0.809	0.825	0.829	0.830	0.833

Standard errors in parentheses. Reference category in brackets. Coefficients from linear regression model.

Ns are unweighted; coefficients are weighted for initial survey design and for attrition to the level of the main wave four survey.

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

^Outcome is KS1 Average Points Score; range: 3-22.5

^^Specification one controls for age at tests and month of birth, specification two adds pupil gender, pupil ethnicity, family income-level, main parent's highest qualification, school type, pupil's length of time attending school; specification three adds age five parent-assessed SDQ, age seven teacher assessed SDQ, age seven teacher judgment of pupil's behaviour; specification four adds Foundation Stage Profile score; specification five adds pupil special educational needs diagnosis; specification six adds teacher gender, teacher years teaching, teacher years teaching at this school. See Annex 3C for all coefficients.

Results continue to hold when children's KS1 reading levels and KS1 maths levels are examined respectively. In these ordered probit models, the appropriate cognitive test (reading / maths) as well as the additional two tests (reading / maths plus PCT) continue to be controlled for, to isolate disparities for children scoring equivalently on both the relevant test and the other assessments. This eliminates the possibility that it is performance in the other domains that is influencing teachers' perceptions of a child's aptitude in the subject of interest – so findings here represent the relationship between stream placement and Key Stage One reading / maths score for children who score equally in that relevant, recently completed cognitive test, and in the other cognitive tests, and who are similar according to other covariates.

Table 3.12 indicates that, at specification six, children are more likely to be assessed at a higher reading level at KS1 if they are in the top stream rather than the middle stream (p <.001), while pupils in the bottom stream are more likely to be rated at a lower level than those in the middle steam (p <.05). Similarly, children scoring equivalently on the maths cognitive test who are otherwise alike but who are in the top rather than middle stream have a higher probability of being assessed at a higher level at maths by their teacher (p <.05), while children in the bottom stream are less likely (p <.001).

Table 3.12: Differences in Key Stage One reading / maths level according to pupils' stream placement (specification six)^

	Reading level	Maths level
Top stream	0.913***	0.540*
	(0.219)	(0.209)
(Middle stream)	0	0
	(.)	(.)
Bottom stream	-0.438*	-0.903***
	(0.205)	(0.208)
Maths Test score	-0.00980	0.0778***
	(0.013)	(0.013)
Word Reading Test score	0.0501***	0.0174***
	(0.005)	(0.003)
Pattern Construction Test score	0.0103*	0.0181***
	(0.005)	(0.005)
Cut 1: Constant	-6.124	-9.501*
	(4.652)	(3.859)
Cut 2: Constant	-3.257	-7.347^{+}
	(4.620)	(3.822)
Cut 3: Constant	-1.908	-5.833
	(4.619)	(3.811)
Cut 4: Constant	-0.107	-4.259
	(4.646)	(3.821)
N	437	460

Standard errors in parentheses. Reference category in brackets. Coefficients from ordered probit models.

Ns are unweighted; coefficients are weighted for initial survey design and for attrition to the level of the main wave four survey.

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

^Outcome is KS1 reading / maths level: 'working towards level 1' / achieved level 1' / 'achieved level 2c' / 'achieved level 2b' / 'achieved level 2a.'

^^Controlled for age at tests, month of birth, pupil gender, pupil ethnicity, family income-level, main parent's highest qualification, school type, pupil's length of time attending school; age five parent-assessed SDQ, age seven teacher assessed SDQ, age seven teacher judgment of pupil's behaviour; Foundation Stage Profile score; pupil special educational needs diagnosis; teacher gender, teacher years teaching, teacher years teaching at this school.

Results: Stream placement and survey-reported teacher judgements

Table 3.13 presents key results for each model specification, for analysis where the outcome is summed survey-reported teacher judgment (see Annex 3D for all model coefficients). Findings are congruent with those using KS1 results. They show an enduring relationship between pupils' stream placements and their teachers' judgements of their 'ability and attainment.' Even at specification 6, being in the top stream is associated with overall teacher judgements of 'ability and attainment' (range: 7-35; SD: 5.3) 2.7 points (51 percent of a standard deviation) higher (p < .001), and being in the bottom stream associated with judgements -1.7 points (32 percent of a standard deviation) lower (p <.001).

Table 3.13: Difference in survey-reported summed teacher judgment of 'ability and attainment' according to pupils' stream placement^ ^^

	Spec 1	**Spec 2**	**Spec 3**	**Spec 4**	**Spec 5**	**Spec 6**
Top stream	3.157***	2.874***	2.661***	2.586***	2.611***	2.569***
	(0.286)	(0.274)	(0.260)	(0.253)	(0.250)	(0.258)
(Middle stream)	0	0	0	0	0	0
	(.)	(.)	(.)	(.)	(.)	(.)
Bottom stream	-2.702***	-2.384***	-1.964***	-1.897***	-1.686***	-1.704***
	(0.327)	(0.328)	(0.318)	(0.299)	(0.289)	(0.280)
Maths Test score	0.0951***	0.0971***	0.0681**	0.0646**	0.0602**	0.0611**
	(0.023)	(0.024)	(0.021)	(0.021)	(0.021)	(0.021)
Word Reading Test score	0.0489***	0.0502***	0.0484***	0.0456***	0.0437***	0.0440***
	(0.005)	(0.005)	(0.004)	(0.004)	(0.004)	(0.004)
Pattern Construction Test score	0.0313***	0.0258***	0.0168*	0.0166*	0.0172*	0.0159*
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Constant	6.932	34.41***	36.48***	36.02***	35.91***	35.84***
	(5.809)	(7.845)	(7.509)	(7.417)	(7.317)	(7.194)
N	829	829	823	823	823	823
R^2	0.703	0.737	0.769	0.773	0.775	0.776

Standard errors in parentheses. Reference category in brackets. Coefficients from linear regression model.

Ns are unweighted; coefficients are weighted for initial survey design and for attrition to the level of the main wave four survey.

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

^Outcome is summed teacher survey-reported judgment; range: 7-35

^^Specification one controls for age at tests and age at teacher survey, specification two adds pupil gender, pupil month of birth, pupil ethnicity, family income-level, main parent's highest qualification; specification three adds age five parent-assessed SDQ, age seven teacher assessed SDQ, age seven teacher judgment of pupil's behaviour; specification four adds Foundation Stage Profile score; specification five adds pupil special educational needs diagnosis; specification six adds teacher gender, teacher years teaching, teacher years teaching at this school. See Annex 3D for all coefficients.

Table 3.14 shows that results also hold when teacher judgment of reading ability is considered in isolation (conditional upon children's reading ability test score, maths and PCT test scores, and all non-cognitive test covariates), as well as when maths ability is considered alone. Judgements of both reading and maths ability, like summed overall teacher judgements, are related to the stream in which a pupil is situated – higher stream placement is associated with higher judgment of both reading and maths ability, even when pupils score equivalently on the relevant cognitive test, the additional cognitive tests, and are otherwise similar.

Table 3.14: Differences in survey-reported teacher judgements of level of reading / maths 'ability and attainment' according to pupils' stream placement (specification six)^ ^^

	Reading judgment	Maths judgment
Top stream	1.193*** (0.158)	1.143*** (0.158)
(Middle stream)	0 (.)	0 (.)
Bottom stream	-0.837*** (0.170)	-1.087*** (0.182)
Maths Test score	0.00523 (0.011)	0.0499*** (0.012)
Word Reading Test score	0.0338*** (0.002)	0.0102*** (0.002)
Pattern Construction Test score	0.00426 (0.003)	0.0111*** (0.003)
Cut 1: Constant	-10.09** (3.022)	-10.67** (3.485)
Cut 2: Constant	-7.912** (3.015)	-8.587* (3.471)
Cut 3: Constant	-5.563^{+} (3.015)	-6.198^{+} (3.507)
Cut 4: Constant	-3.465 (3.027)	-4.219 (3.515)
N	843	839

Standard errors in parentheses. Reference category in brackets. Coefficients from ordered probit models.
Ns are unweighted; coefficients are weighted for initial survey design and for attrition to the level of the main wave four survey.
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
^Outcomes are survey-reported teacher judgements of reading / maths ability; range: 1-5
^^Controlled for age at tests and age at teacher survey, pupil gender, pupil month of birth, pupil ethnicity, family income-level, main parent's highest qualification; age five parent-assessed SDQ, age seven teacher assessed SDQ, age seven teacher judgment of pupil's behaviour; Foundation Stage Profile score; pupil special educational needs diagnosis; teacher gender, teacher years teaching, teacher years teaching at this school.

Robustness checks

A sensitivity check was carried out to examine whether removing teachers' judgements regarding less 'academic' subjects from the overall survey-reported summed judgment of 'ability and attainment' affected findings. Results are entirely consistent using this alternative outcome (Annex 3E). A second check replicates analyses without MCS survey weights but with clustering of standard errors at the school level. Again, findings are consistent and remain significant at the 5 percent level (Annex 3F). Lastly, linear versions where the five levels of response regarding reading / maths ability / assessment were treated as continuous variables yielded equivalent results (Annex 3G.)

Discussion

This research set out to explore whether teacher judgements and assessments of pupils are influenced by the stream to which a child is allocated. Having controlled for children's recent performance on relevant cognitive tests, as well as a range of pupil, family, and teacher characteristics, pupil behaviour, teacher perceptions of pupil behaviour, and prior performance and assessment, it finds a consistent association between stream level and teacher judgements of pupils' academic ability and attainment. This holds both for assessments of performance at KS1 and survey-reported teacher perceptions. The hypothesis that teacher judgements of pupils are influenced by the stream to which a pupil is allocated is therefore supported.

Analysis here has indicated that, on average, children placed in higher streams are judged and assessed disproportionately favourably, and children in lower streams at a disproportionately lower level. That this apparent effect is significant across measures and academic domains suggests that it is strong and pervasive. Findings therefore call into question the general utility and equitability of the practice of streaming. Analyses in this chapter also show that certain groups of pupils (boys, low-income pupils, pupils whose parents have fewer qualifications, summer-born children) are over-represented in lower streams, and under-represented in the highest groupings. Rather than going any way towards promoting parity in academic achievement, there is a danger therefore that the increasing use of streaming among primary school pupils will only perpetuate or widen attainment gaps.

Alternative and additional explanations

Findings in this chapter indicate a cross-domain and pan-situational relationship between stream placement and teacher judgements of pupils. However, as well as supporting the hypothesis that stream placement influences teacher perceptions and assessments, results here may also be interpreted as suggesting additional explanations.

The MCS data are observational, so it is logically feasible that alternative factors could explain the patterns described. All survey instruments and applications are vulnerable to some extent to measurement error, and unobserved factors, unproxied by the factors included, may play some part in the patterns described.

However, in mitigation of this possibility, a rich and comprehensive set of covariates are included in modelling. Findings are congruent with previous studies – which, as noted in the introduction to this chapter, have presented a relatively consistent and durable body of evidence. Therefore, the explanation favoured here – of a direct influence of streaming upon teacher perceptions – seems theoretically coherent and justified.

Yet it remains possible that, in the period between cognitive testing and teacher survey / KS1 assessment, pupils' actual performance (rather than or as well as teacher perceptions of that performance) have followed a course that is in line with their placement level. The trajectory of the manifest development of children in lower streams may be depressed and that of children in higher streams augmented as a result of any effects of stream placement in addition to those on teacher judgements. As discussed in the introduction to this Chapter, previous studies have indicated possible influences of streaming through pupils' own self-perceptions and motivations, and through educational quality and opportunities. These factors may explain some of the apparent association between placement and assessments, and should be explored in future research

Though these possibilities have not been eliminated according to analysis in this chapter, two key points can be noted. Firstly, the time lags between cognitive test completion and teacher judgements are short (particularly for survey-reported assessments, at 2-4 months on average), suggesting that a discrepancy between judgment of attainment (built over the school year) and actual attainment is arguably the more likely explanation than significant change in this brief period in manifest performance. Secondly – and perhaps more crucially – regardless of the hypothesis that is favoured, what is indisputably indicated by findings here is that sample children who are similar according to the observed characteristics and in recent test performance are subsequently differentially assessed in line with their stream placement, and that this relationship is evident in their documented, teacher-assessed ‘achievement.’

In fact, given that the MCS’s cognitive tests were taken mid-year, while stream placement is assumed to have been determined at the beginning of the academic year, and given the possible ongoing, cumulative and iterative influence of this placement though many pathways, it is probable – notwithstanding the caveats regarding observational data referred to above – that findings in this Chapter are in fact merely snapshot underestimates of the overall effects of streaming. Analysis is conditional on scores from tests taken only months before teacher assessments, and these test scores may already have been affected by the child’s placement in this (and possibly previous) academic year(s). That results are consistent and significant when differences have only a limited window within which to manifest indicates the likely immediacy, strength and enduring influence of the practice of streaming.

Conclusions and policy recommendations

Whichever explanation for results in this chapter is preferred, streaming appears to have a durable association with a range of teacher judgements that stretches to official, recorded ‘attainment.’ This is congruent with indications from previous research that streaming is ‘disadvantageous for those in lower sets and increases the overall attainment gap’ (Dunne *et al*, 2007). Given the recent and widespread move back towards ability grouping of primary school children, where the national use of streaming has risen sharply in the past two decades (and, if this trend has continued since last monitored, where it may be ever still more prevalent), these warnings that stream placement can influence both teachers’ perceptions of pupils and permanent decisions regarding ‘attainment’ are particularly pertinent and immediately applicable to current policy and practice.

Of course, indications of probable effect from existing survey data can only go so far in unpicking the processes and complexities behind the averages reported here. It is not possible, for example, fully to explore differences in relationships according to teacher, school, or school constitution using the information collected in the MCS survey. In order to do this, comprehensive, whole school samples are necessary – and in order for these to be nationally meaningful, the overall sample should constitute as many institutions as possible. Collecting information on whether streaming takes place and on the stream placement of each individual pupil, and making this information available for analysis through the National Pupil Database, would address this need and allow proper monitoring and scrutiny of the impacts of streaming. As the practice seems to be becoming rapidly more widespread, and given consistent indications of its effect across research studies, it is imperative that instigation of this data collection be prioritised.

In the meantime – notwithstanding the desirability of more detailed information and analysis – findings here, along with the body of previous research, invite continued and urgent debate by policy-makers and practitioners about the utility and equitability of streaming. Can the recent move towards use of the practice among young children really be justified by anything other than blind ideology, or does the available evidence in fact indicate that it should be ceased altogether?

Chapter 4

Bias and stereotyping in teachers' judgements of seven-year old pupils

Introduction

Teacher assessment and pupil attainment

Since the introduction of the National Curriculum in 1988, the time dedicated to standardised assessment of English pupils has increased considerably, alongside a growing requirement that much of this assessment be performed by class teachers. Teacher judgements currently dominate children's designated attainment levels within primary education. At the time of writing, the Foundation Stage Profile (FSP; covering the years up to age five) is entirely teacher-assessed, along with the newly-introduced phonics screening test (taken at ages six and seven), and Key Stage One (KS1) attainment (age seven). Primary education culminates at age 11 with the awarding of Key Stage Two (KS2) grades, which comprise two components: the results of externally set examinations, and ratings by teachers (Bew, 2011a, 2011b; Department for Education, n.d.3; Wyse *et al.,* 2008).

This approach to assessment, with its reliance on an understanding of each child built over time rather than based simply on a one-off performance in a set test, has several arguable advantages. It avoids the lack of nuance of the one-shot test, and also the test's time-and place-dependency, which might result in an inaccurate picture of a child's abilities should they underperform on a given day, in the given situation, or in response to the limited test stimuli (Harlen, 2007). Some evidence indicates moreover that formalised testing can be stressful and demotivating for pupils (Harlen, 2004; 2007), and it has also been suggested that exams may be counterproductive to meaningful knowledge acquisition insofar as they encourage 'teaching to the test' at the expense of deeper, sustainable learning and wider exploration (Harlen, 2007; Wyse *et al.,* 2008). However, despite its potential advantages over more formalised and 'objective' measures, teacher assessment is not, in itself, entirely unproblematic, and nor is a primary schooling so heavily intertwined with its processes.

The past decade's national statistics on the performance of English pupils have consistently indicated that certain groups achieve at lower levels than others throughout their early education. Low-income pupils in receipt of free school meals (FSM), pupils with any diagnosis of special educational needs (SEN), Pakistani, Black African, and Black Caribbean pupils, as well as pupils speaking English as an additional language are regularly reported as under-attaining in the primary phase. In addition, boys score generally at a lower level than girls at the foundation stage, although they attain higher levels at maths (and girls at English) at KS1 and KS2 (Department for Education 2011; 2012a; 2012b).

Because attainment indicators depend so heavily on teacher assessment, this invites the question of whether these apparent achievement gaps may be, to some extent, an artefact of the measurement method used. There is an enduring body of evidence which indicates that teacher assessments are subject consistently to a large and significant level of error (Brookhart, 2013; Eckert *et al.*, 2006; Harlen, 2005), and, more importantly, research also indicates that some of this error may be systematic (Harlen, 2005; Robinson & Lubienski, 2011), and that there may be regular patterns of inequality in teacher judgements of English primary school pupils (Burgess & Greaves, 2009; Reeves *et al.*, 2001; Thomas *et al.* 1998).

Bias in teacher assessment

For example, examining national KS2 data, Burgess & Greaves (2009) exploit the distinction between the teacher-assessed and externally-examined components of the test, comparing marks awarded to pupils according to the two measures. They demonstrate disparities in teacher assessment which are in line with several of the nationally-reported attainment gaps: seeming under-assessment of pupils in receipt of FSM, of pupils with SEN, and of Black Caribbean and Black African pupils. This suggests that teacher-level bias may influence the KS2 scores allocated to each pupil.

Analysing the English sub-sample of the Millennium Cohort Study (MCS), Hansen & Jones (2011) indicate that teachers may also be biased in their assessments of pupils at the beginning of primary school. They compare children’s FSP scores to self-completed cognitive tests taken outside of school, and find greater disparities according to gender in the teacher-assessed FSP measure than in the child-completed tests. Teacher assessments pronouncedly favour girls to a greater extent than cognitive test performance, indicating that gender disproportionality at the foundation stage may, like inequalities at KS2, be attributable in part to biased judgements.

Qualitative research, some of it government-commissioned, has moreover begun to suggest mechanisms that might underpin these apparent biases in assessment and resultant attainment, particularly with regard to ethnic disparities. Evidence that perceptions and behaviours among teaching staff may play a part in creating variation has been provided by Maylor *et al.*’s (2009) evaluation of the Black Children’s Achievement Programme, which concludes that, ‘Institutional factors / processes including negative teacher attitudes / expectations’ and ‘stereotypical thinking about the ability of Black children serve to undermine teacher ability to raise Black children’s attainment at an individual and group level’.

Similarly, Strand *et al.*’s (2010) investigation into *Drivers and Challenges in Raising the Achievement of Pupils from Bangladeshi, Somali and Turkish Backgrounds* reports that: ‘Racism and structural inequalities may be important influences on the attainment of many Bangladeshi and Somali students’. As also suggested by Burgess and Greaves’ large-scale quantitative work (2009), these studies indicate that stereotyping at the teacher-level may provide some explanation for the ostensible attainment differentials among primary school pupils.

Biased assessments through stereotyping

There are a number of theories of what stereotypes *are*, and of behaviours associated with their presence. Many are grounded in the premises that stereotypes comprise invariant, homogenous, evaluative judgements of a given group (e.g. income, gender or ethnic group), and that stereotypes enable judgements of group members to be made quickly and with cognitive ease (Hilton & von Hipple, 1996; McGarty *et al.*, 2002.) By stereotyping, therefore, teacher judgements of pupils can be made quickly and with cognitive efficiency (though with compromised *accuracy*) based, in part, on a preconceived 'template' of the ability and attainment of low-income pupils, pupils with SEN, White pupils, Black Caribbean pupils, and so on. Stereotyping is not assumed to take place on a conscious or deliberate level: the process's efficiency is thought to be engendered by its automaticity.

Theorists argue furthermore that stereotypes must be held at the group or institutional level: '…stereotypes should be formed in line with the accepted views or norms of social groups that the perceiver belongs to' (McGarty *et al.*, 2002, p.2). The possibility, therefore, is that among the English teaching profession there exist normalised notional templates of pupil attainment, which are premised on pupil characteristics, inform judgements of each child, and skew assessments in line with these characteristics.

Building upon previous evidence to test the stereotype model

To date, relatively little credence or focus appears to have been afforded in the policy arena to the possibility that bias and stereotyping might provide some explanation for systematic variation in children's achievement, particularly in primary school. Despite the growing body of evidence that this may be the case, policy has tended to look instead to the family-level for first causes of inequalities, often citing socio-economic differences as the primary driver, and directing resources accordingly (Department for Children, Schools and Families, 2008b; Department for Education, 2010a; Department for Education 2010c; Department for Education and Skills, 2005). Yet if the process of stereotyping can definitively be implicated as instrumental in biases in teacher assessment (and consequentially as contributing to attainment disparities), this will clearly indicate a point at which intervention to mitigate these inequalities might be deployed.

However, existing research does not yet unequivocally support the theory that pupils are being stereotyped by their teachers, or even that apparent biases are wholly unfounded. For example, though they show clear patterns of disparity, and though they propose and support a stereotype model, Burgess & Greaves (2009, p.12) also acknowledge an alternative explanation for their findings. Because their analysis uses comparators from within the same overall system (the teacher who assesses the pupil at KS2 also teaches them for the externally-marked KS2 test), there is a danger of causal explanatory relationships within the system. Burgess & Greaves suggest, for instance, that the notable difference between the teacher-assessed and externally-marked elements of SEN pupils' results, in particular, may be due to: '…an extreme form of "teaching to the test" for pupils with SEN…the teacher's more in-depth knowledge of the student's ability may result in a lower [teacher assessment]'. That is, teachers might explicitly train and focus on certain pupils, whom they see as less able, so that they learn to attain desirable KS2 levels in the test situation. As a result, these test results may not reflect the teacher's day-to-day perception of the pupil's ability – and this, rather than stereotyping, may be what underpins apparent biases.

Hansen & Jones' (2011) analysis partially circumvents this issue and avoids interrelatedness of measures by utilising tests of pupil 'ability' which are not explicitly associated with their schooling, and not directly influenced or reported by their teacher. Cognitive tasks independently administered in children's homes as part of the MCS are compared to school-based, teacher-assessed FSP scores, arguably providing an enhanced indication that teacher judgements are biased away from manifest pupil performance.

However, while Hansen & Jones' study strengthens the evidence that recorded teacher assessments are systematically skewed, a danger remains that FSP scores do not in fact comprise direct portrayals of the mental representation – the potentially stereotype-based 'evaluative judgement' – that each assessing teacher holds of their (groups of) pupils. Because schools themselves, at the institutional level, are judged by the attainment of their pupils, and because teachers' own performance is assessed according to the attainment of their class, it is highly likely that FSP scores serve not only to describe the teacher-perceived attainment or progress of each individual child, but to inform additional purposes (Bradbury, 2011a; Harlen, 2007).

A recent report by Ofqual (2012, p.82) noted, for example, a tendency within teacher assessment to manipulate ‘marks so that candidates [are] placed within certain perceived grade boundaries’, and recent reporting of national scores for the teacher-assessed phonics screening test clearly illustrates this phenomenon (Department for Education, 2012c, p.4).[9] One response to a 2009 Ofsted consultation stated that: ‘Schools can manipulate…scores in ways that Ofsted would be unlikely to support,’[10] while Bradbury (2011b, p.655) describes findings from case studies where ‘assessment results may be influenced by pressure from external advisors, who only recognise certain patterns of results as intelligible,’ and where this moderation brings about amendments to pupils’ test scores in line with established normative expectations. Recorded FSP results may, therefore, provide a somewhat inaccurate representation of teacher perceptions of a given individual or group, due to their complicity with, and the incentives of, a system where the attainment levels awarded to pupils have implications far beyond measuring and assessing each child’s ability, progress or performance.

The current study

Therefore, in order to investigate less ambiguously and more explicitly whether teacher-level stereotyping of pupils may relate to biased assessment according to pupil characteristics, the analysis presented in this chapter uses a measure of teacher judgement which is not part of, nor required by, the education and assessment system, which is removed from its context, and which will not inform evaluations of performance of a teacher or their school. Confidential responses provided by teachers participating in the Millennium Cohort Study (MCS) to questions about their pupils’ ‘ability and attainment’ (at age seven) provide a proxy for the teachers’ mental representations of each pupil. These survey responses should lack the

[9] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/219208/main_20text_20_20sfr21-2012.pdf (p 4).
[10] http://ofstednews.ofsted.gov.uk/article/346

agenda inherent to the formal in-school assessments used in previous research. In addition, like Hansen and Jones' paper, the current study uses independent MCS-administered cognitive test scores (also collected at age seven) as comparators that indicate each child's contemporaneous manifest performance.

Analysis explores whether there are biases in teacher judgements of pupils which correspond to each of the key pupil characteristics underpinning recorded primary-age attainment gaps (family income-level, gender, SEN, ethnicity, EAL) and which may, as proposed, account to some extent for these gaps. Additionally, it begins to explore which of these characteristics appear to dominate and drive any apparent biases, in order further to inform potential interventions which may tackle stereotyping.

Methodology

Sample

In common with the other chapters in this thesis, analysis here uses data from wave four of the Millennium Cohort Study, when pupils were seven years old, and in Year Two at primary school (data source: University of London 2011; 2012). Analysis is restricted to state school children in England, in order to allow comparison with, and interpretation in the context of, Department for Education (DfE) statistics on pupil attainment. Twins and triplets are excluded, because teacher bias and stereotyping may follow a different process for these pupils.

See Chapter 1 for detailed discussion of use of the sub-sample of pupils for whom teacher survey data is available. Cognitive test scores are missing for a small minority, and there is also some non-response to various individual questions utilised in this chapter. The base samples for analysis here thus comprise those 4997 / 4985 (reading / maths) MCS children who continued to participate at wave four, whose teachers responded, and for whom there are all necessary key data.

As in Chapter 2, main estimates in this Chapter are weighted for the MCS's design features and for attrition to the level of the main wave four sample, as per Mostapha (2013), and additional robustness checks with different weighting and sample specifications are carried out and reported in the results section.

Annex 4A compares key characteristics of the English singleton MCS sample at wave one to three samples at wave four: that with teacher survey response, that with all data necessary for analysis in this chapter, and that without teacher response. It also contrasts estimates with and without attrition weights. It suggests some relatively minor differences between samples: that the sample used in this chapter are from families slightly fewer of whom are low-income than those without teacher response at wave four, who are more likely to speak only English at home; that the pupils are more often of White ethnicity, score marginally higher in the cognitive tests, and are slightly more often girls. Where comparison across waves is possible, estimates weighted for design and attrition are similar for the wave one sample and for the sample used in this chapter.

As noted in Chapter 2, therefore, any relationships found in the current analysis can be attributed with certainty only to the children included – but this large, country-wide sample, which, according to the comparisons in Annex 4A does not seem massively skewed, can be used to theory-build and to explore the hypothesis that stereotyping by teachers takes place.

Teacher judgements

Teacher-reported judgements of whether each pupil is ‘well above average / above average / average / below average / well below average’ at both reading and maths, respectively, form the crux of analysis. As reported in Chapter 3, these evaluations are in response to a survey question asking the teacher to ‘rate [the given] aspect of the study child’s ability and attainment [reading / maths]…in relation to all children of this age…’[11]

Overall, teachers’ ratings of the MCS pupils are positively skewed: for reading, 12% are rated ‘well above average;’ 33% ‘above average;’ 33%

[11] See http://www.esds.ac.uk/doc/6848/mrdoc/pdf/mcs4_teacher_england.pdf for full survey documentation, and Annex ii for extracts.

‘average;’ 17% ‘below average;’ and 5% ‘well below average’ (n=4997; weighted estimates). Similarly, for maths, 9% are rated ‘well above average;’ 31% ‘above average;’ 40% ‘average;’ 16% ‘below average;’ and 4% ‘well below average’ (n=4985; weighted estimates).

For modelling in this chapter, responses are recoded into binary variables representing a rating of ‘above’ or ‘below’ average, which indicate whether each child is judged as relatively more or less able, compared to their peers. Responses of *well above average* and *above average* are combined to form the ‘above average’ category, where all else is categorised ‘not above average;’ similarly, responses of *well below average* and *below average* are combined to one ‘below average’ category. While it necessitates a coarser analysis of biases, this merging of responses allows use of an easily interpretable linear probability model, and ensures robust cell sizes in logistic modelling. Four outcome variables are thereby created:

- teacher judgement of reading ‘above average’ / not;
- teacher judgement of reading ‘below average’ / not;
- teacher judgement of maths ‘above average’ / not;
- teacher judgement of maths ‘below average’ / not.

Pupil characteristics

In addition, the following measures of each of the pupil characteristics identified by DfE statistics as underpinning attainment variation are used (all are taken at wave four):

- a derived variable from parent-reported data which indicates whether the family’s income is above / below an OEDC 60 percent of median UK income poverty indicator;
- parent-reported pupil gender;
- teacher report of any recognised SEN (yes / no);

- a derived variable from parent report denoting pupil ethnic group (White / Indian / Pakistani / Bangladeshi / Black Caribbean / Black African);
- a derived variable from parent-reported information on language(s) spoken in the pupil's household (coded to represent English only / additional languages).

Only sub-sets of breakdowns by ethnicity are reported in this chapter, in order to aid meaningful interpretation and comparison with DfE statistics. The census-based eight-category ethnicity categorisation is used throughout analysis, and includes 'other' and 'mixed' classifications – but results for these groups are not presented. Descriptive statistics according to ethnicity may therefore not sum to 100 percent, while in modelling, noted sample sizes are for the whole sample with ethnicity data – as all are included in analysis – although only results for selected groups are outlined.

Teacher judgements and pupil characteristics

Table 4.1 shows the percentage of MCS pupils with each characteristic who are evaluated by their teacher as relatively more or less able than their peers, according to the definitions described above. It indicates a lower chance of being evaluated as 'above average' at reading for low-income pupils, boys, pupils with SEN, pupils of all ethnicities except White and Indian, and pupils speaking languages in addition to English. The same pattern holds for judgements of maths ability, save for a reversal according to gender, with boys more highly rated in this domain.

Table 4.1: Percentage of pupils with each characteristic judged at each level by their teacher*

	Percentage judged 'above average' at reading	Percentage judged 'below average' at reading	Percentage judged 'above average' at maths	Percentage judged 'below average' at maths
Whole sample (n = 4997 / 4985)	45.3	22.2	39.8	20.5
Above 60% median income (n = 3593 / 3585)	52.3	16.6	45.6	16.1
Below 60% median income (n = 1404 / 1400)	26.6	37.3	24.2	32.1
Boys (n = 2494 / 2491)	40.5	27.1	42.4	21.4
Girls (n = 2503 / 2494)	50.1	17.4	37.1	19.5
No SEN diagnosis (n = 3879 / 3864)	55.7	9.3	48.5	9.2
Any SEN diagnosis (n = 1118 / 1121)	11.1	64.7	11.2	57.1
White (n = 4047 / 4032)	46.2	21.7	40.6	19.8
Indian (n = 150 / 150)	46.9	18.1	46.1	14.6
Pakistani (n = 274 / 274)	30.4	29.4	23.8	30.9
Bangladeshi (n = 85 / 86)	38.5	28.3	36.7	24.2
Black Caribbean (n = 68 / 68)	28.6	37.0	20.7	36.7
Black African (n = 112 / 112)	42.8	26.0	25.0	23.4
Speaks English only (n = 4317 / 4305)	46.0	21.9	40.5	20.1
Speaks additional languages (n = 680 / 680)	38.0	25.3	31.8	23.6

*All estimates weighted for survey design and for attrition to the main wave four survey. Ns are unweighted.

Cognitive test scores

At age seven, the MCS children completed a number of cognitive tests during a home visit from a survey administrator. As described in Chapter 3, they included the British Ability Scale Word Reading test, and a shortened version of the Progress in Mathematics test. The Word Reading test is designed to assess children’s English reading skills (see http://www.glassessment.co.uk/products/bas3). The ability score (a scaled but not otherwise standardised score) is used in analysis (see Hansen, 2012). The Progress in Mathematics test is designed to measure pupils’ mathematical ability across use of numbers, shapes, and proficiency in data handling. It is intended to provide an indication of performance in maths at the given developmental stage (see http://www.gl-assessment.co.uk/products/progress-maths). A shortened version was used in the survey and entailed routing to sections of varying difficulty levels. Rasch scaling was used to convert the raw scores to a count score equivalent to that which would be attained were the full test completed (see Hansen, 2012). This scaled score is used in analysis here.

Performances on the two cognitive tests provide respective points of comparison to the teacher assessments of pupil reading and maths ‘ability and attainment.’ As noted in Chapter 3, completion of the cognitive tests shortly preceded teacher completion of their survey: the mean average time lag between cognitive test and teacher survey was 3.8 months, the median 3 months, and the mode 2 months. Comparisons using the two measures necessitate assumptions: a) that the lag between pupil test completion and teacher survey completion does not vary systematically across the pupil characteristics of interest; and b) that children delineated by each of the characteristics of interest develop at equivalent rates in their reading and maths ability and performance, at age seven (so that any apparent bias in teacher assessments cannot be attributed to slower progress during the time lag from pupil survey to teacher survey in some groups). The second of these assumptions cannot explicitly be tested using the MCS data, so remains a supposition (although as the modal time lag was short, at two months, it seems reasonably unproblematic); the first is accounted for by including a control for test-teacher survey lag in all main analyses.

Test scores and pupil characteristics

Table 4.2, below, shows the mean Word Reading scores and Progress in Maths scores for the samples of pupils who took the tests and who also have responses to the teacher-completed question on reading / maths ability (respectively), according to each characteristic of interest.

Table 4.2: Mean scores by characteristic on Word Reading and Progress in Maths tests*

	Mean Word Reading score (range: 10-214)	**Mean Progress in Maths score (range: 0-28)**
Whole sample with teacher reading / maths judgement (n = 4997 / 4985)	108.54	18.41
Above 60% median income (n = 3593 / 3585)	112.48	19.17
Below 60% median income (n = 1404 / 1400)	98.06	16.40
Boys (n = 2494 / 2491)	105.85	18.43
Girls (n = 2503 / 2494)	111.24	18.40
No SEN diagnosis (n = 3879 / 3864)	116.49	19.65
Any SEN diagnosis (n = 1118 / 1121)	82.50	14.39
White (n = 4047 / 4032)	108.00	18.61
Indian (n = 150 / 150)	117.05	19.61
Pakistani (n = 274 / 274)	108.93	15.32
Bangladeshi (n = 85 / 86)	114.95	15.68
Black Caribbean (n = 68 / 68)	101.43	16.77
Black African (n = 112 / 112)	117.74	16.81
Speaks English only (n = 4317 / 4305)	108.17	18.58
Speaks additional languages (n = 680 / 680)	112.28	16.75

*All estimates weighted for survey design and for attrition to the main wave four survey. Ns are unweighted.

On average, sample girls' scores on the Word Reading test are higher than boys', pupils with SEN have lower scores than those with no recognised SEN, and mean scores for low-income and Black Caribbean pupils are also relatively low. Pupils speaking languages in addition to English have higher reading scores, on average, than pupils speaking only English, and Indian, Bangladeshi and Black African pupils also have comparatively high scores.

Though measured on different scales and not, therefore, directly comparable, these descriptive statistics begin to indicate incongruities between children's cognitive test scores and judgements by their teachers. Sample pupils speaking languages in addition to English appear more likely to score relatively well on the BAS Word Reading test – but are less likely than pupils speaking only English to be rated highly at reading by their teacher. Similarly, Black African and Bangladeshi pupils score relatively highly on the Word Reading test – but are again less likely to be judged ‘above average’ and more likely to be judged ‘below average’ by their teacher

As with Word Reading scores, Table 4.2 indicates that sample pupils with SEN and low-income pupils are more likely to attain relatively low scores on the Progress in Mathematics test. In contrast to Word Reading, however, pupils speaking languages in addition to English score lower, on average, than pupils speaking English only, and pupils of all reported ethnicities except for White and Indian are relatively more likely to attain a lower score on this test. Mean scores for boys and girls are very similar, which again indicates some discrepancy between scores and teacher judgements of pupils’ maths ability, which showed a tendency to favour boys (Table 1).

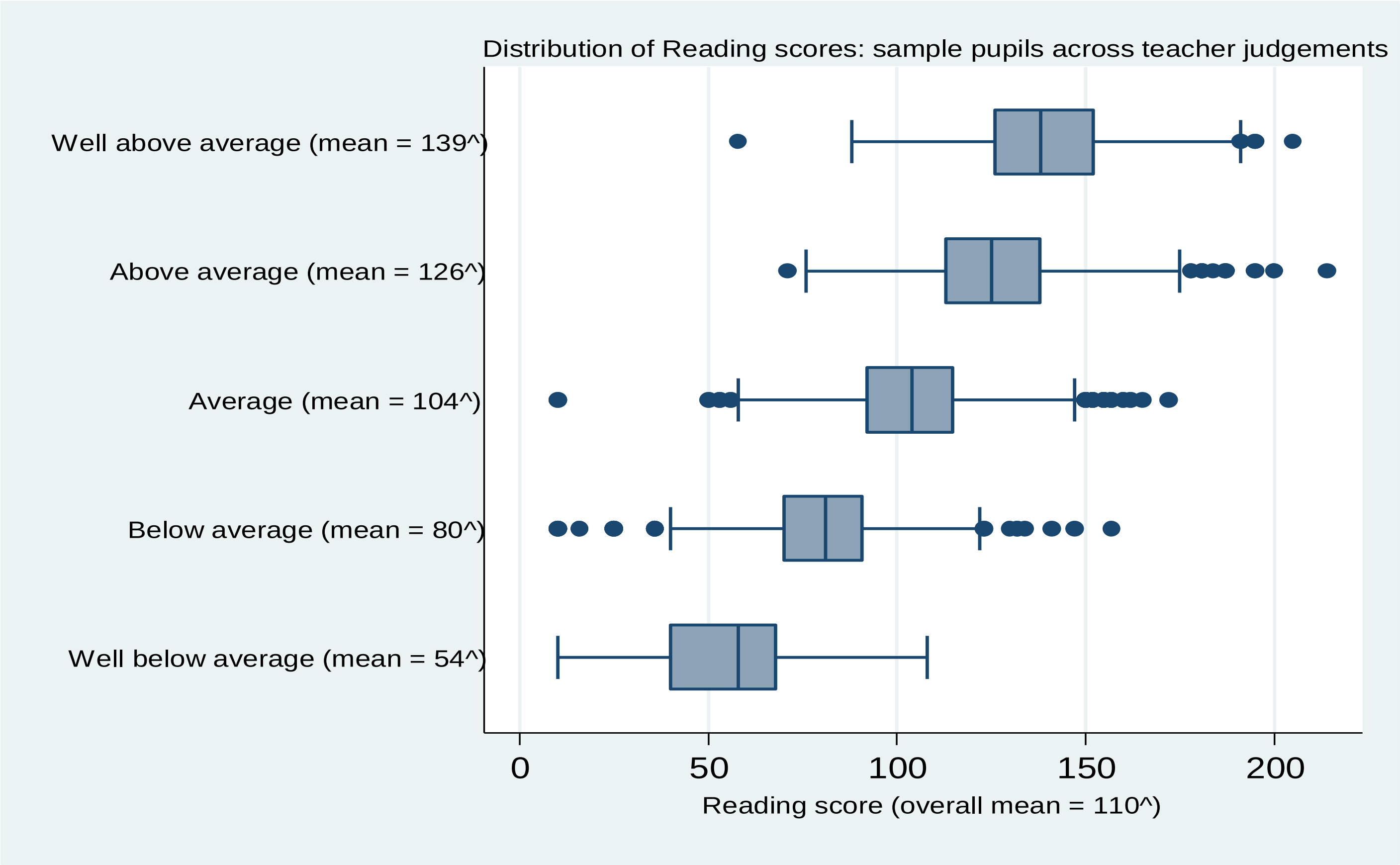
Modelling: Are some groups of pupils systematically rated less favourably by their teachers?

That there are apparent incongruities between average scores of pupils with varying characteristics for the Word Reading test and teacher judgements of reading ‘ability and attainment’ begins to support the possibility that there may be biases in teacher perceptions of pupils according to the pupils’ characteristics. In order explicitly to investigate this, regression modelling compares teacher judgements of pupils who differ according to a given characteristic but who score at the same level on the relevant cognitive test.

The methodology here relies on a general overall relationship, across the sample, between performance on each cognitive test and teacher assessment of pupil ‘ability and attainment’ in the relevant domain. This relationship is strong. Within the whole sample, a naïve regression of BAS Word Reading test score on whether a pupil’s teacher perceives their reading as ‘above average’ indicates that each additional point scored on the Word Reading test (range 10–214) is related to a likelihood of being judged ‘above average’ increased by 1.1 percentage point (p <. 001). For teacher judgements of reading ‘below average’, the relationship is inverted and there is a decrease of –.8 of a percentage point (p < .001). The relationship between point increase in Progress in Maths score (range 0–28) and judgement of ‘above average’ in maths is 4.2 percentage points (p < .001). For judgements below average it is –3.4 percentage points (p < .001).

Figure 4.1 presents the means and distributions of BAS Word Reading test scores for pupils judged to be at each level of reading ‘ability and attainment’ by their teacher, and Figure 2 presents the equivalent information for maths scores and judgements. These figures again illustrate, across all sample pupils, overall linear associations between test scores and teacher judgements. Pupils with a higher cognitive test score tend to be judged to have a higher level of ‘ability and attainment’ by their teacher, though this is not a perfect relationship, and there are also overlaps.

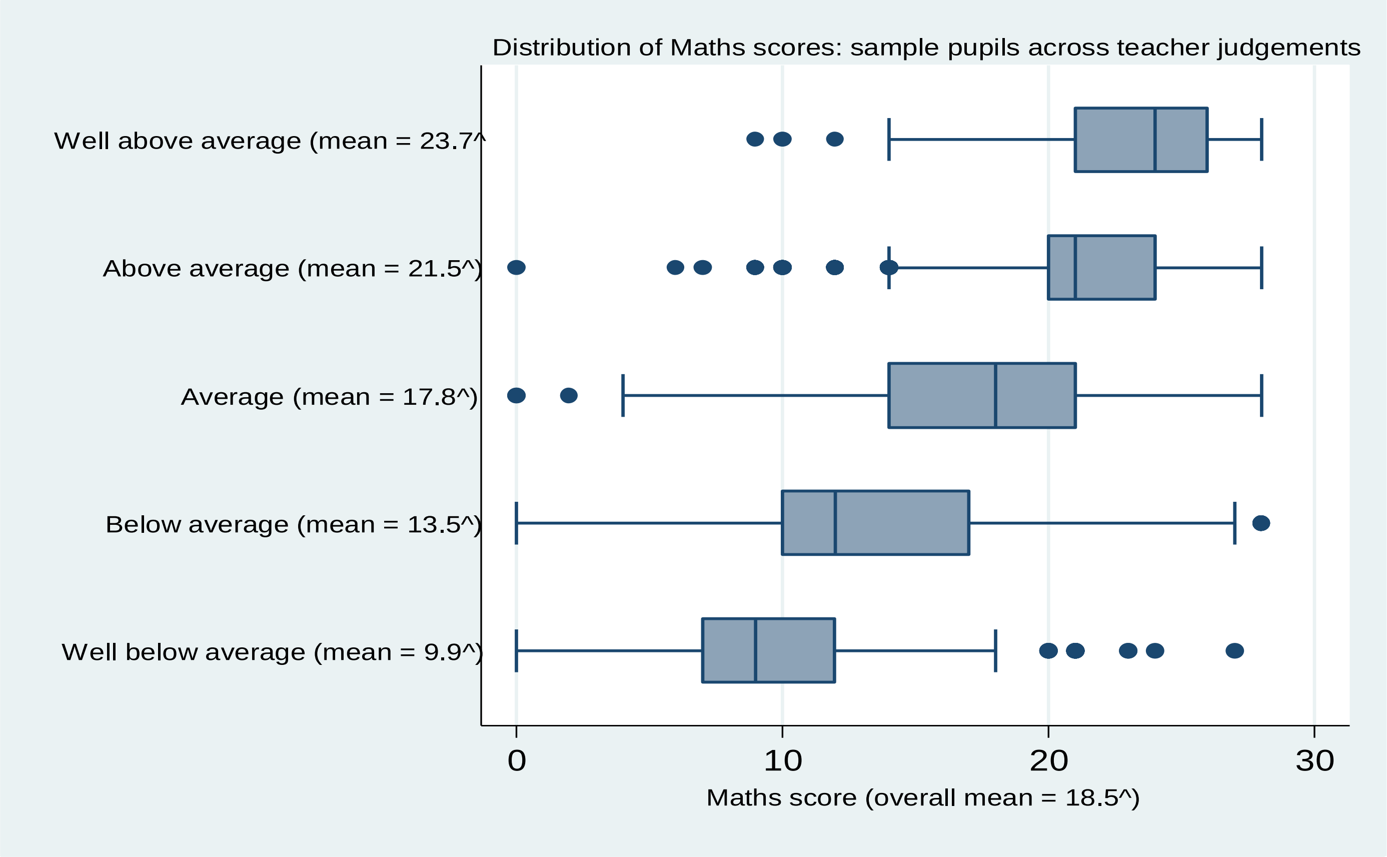
Figure 4.1: Distribution of and mean BAS Word Reading scores of pupils with each teacher judgement of reading 'ability and attainment'



N = 4997 (unweighted). ^Means are unweighted; weighted estimates: overall mean = 109; well above average = 139; above average = 126; average = 104; below average = 79; well below average = 54.

Line represents median, box represents 25th and 75th percentiles (Q1 and Q3, respectively), whiskers represent Q3+1.5(Q3-Q1) / Q1-1.5*(Q3-Q1).

Figure 4.2: Distribution of and mean Progress in Maths scores of pupils with each teacher judgement of maths 'ability and attainment'



N = 4985 (unweighted). ^Means are unweighted; weighted estimates: overall mean = 18.4; well above average = 23.7; above average = 21.4; average = 17.7; below average = 13.6; well below average = 10.3.
Line represents median, box represents 25th and 75th percentiles (Q1 and Q3, respectively), whiskers represent Q3+1.5(Q3-Q1) / Q1-1.5*(Q3-Q1).

If there are no biases in teacher judgements according to the pupil characteristics of interest, these associations should not vary, nor the imperfection of the relationship be explained, by income-level, gender, SEN status, language, or ethnicity. Girls and boys, for example, who score at the same level on the Word Reading test, should have equal probabilities of being judged ‘above average’ at reading by their teacher.

A linear probability model is used to test whether this is the case. The outcome (for example) is whether a child is judged ‘above average’ at reading, and the predictors are: pupil gender, and ability score on the reading test. The likelihood of boys being judged ‘above average’ at reading by their teacher is thereby compared to the likelihood of girls who score at the same level. Analysis takes the following form:

$$\begin{aligned} &Probability\ of\ being\ judged\ `above\ average'\ at\ reading\ by\ teacher_{0-1} \\ &\qquad = Constant + \beta 1 Boy_{0/1} + \beta 2 BAS\ word\ reading\ score + error \end{aligned}$$

The coefficient for boys represents the percentage point difference in likelihood, compared to girls who score equivalently on the Word Reading test, of being judged ‘above average.’ A coefficient of *0* would therefore indicate that there is no bias according to gender in teacher assessments of reading ability. A positive coefficient indicates a positive bias for boys, and a negative coefficient a negative bias.

Analysis is repeated separately for each pupil characteristic and outcome, resulting in the following basic models (Table 4.3). All analyses use Stata (versions 12 and 13).

Table 4.3: Variables used in and structure of linear probability models^

Model	Outcome	Predictors	
1	Teacher judgement of reading *above* average / not	BAS Word Reading test ability score	+ above / below 60% income
2			+ boy / girl
3			+ SEN / not
4			+ White / Indian / Pakistani / Bangladeshi / Black Caribbean / Black African
5			+ English only / additional languages
6	Teacher judgement of reading *below* average / not	BAS Word Reading test ability score	+ above / below 60% income
7			+ boy / girl
8			+ SEN / not
9			+ White / Indian / Pakistani / Bangladeshi / Black Caribbean / Black African
10			+ English only / additional languages
11	Teacher judgement of maths *above* average / not	Progress in Maths score	+ above / below 60% income
12			+ boy / girl
13			+ SEN / not
14			+ White / Indian / Pakistani / Bangladeshi / Black Caribbean / Black African
15			+ English only / additional languages
16	Teacher judgement of maths *below* average / not	Progress in Maths score	+ above / below 60% income
17			+ boy / girl
18			+ SEN / not
19			+ White / Indian / Pakistani / Bangladeshi / Black Caribbean / Black African
20			+ English only / additional languages

^All main models also include a control for age at cognitive test (linear variable) and a control for time lag between cognitive test completion and teacher survey completion (categorical variable).

Results

Biases in teacher judgements of pupils' reading ability

Table 4.4 indicates variation in the average likelihood of MCS pupils who differ according to each characteristic (income-level, gender, SEN status, ethnicity and language) being rated relatively highly at reading, compared to peers who score equivalently on the Word Reading test. As described in Table 3, separate models were estimated for each characteristic, and findings from each discrete model are presented.

Children from low-income families, boys, pupils with any recognised diagnosis of SEN, and children who speak other languages in addition to English appear less likely to be judged ‘above average’ at reading by their teacher – despite scoring equivalently to their comparison counterparts in the reading test. All these differences are significant at p < .05 at a minimum. MCS pupils of all non-White ethnicities also appear less likely to be judged ‘above average’ at reading (compared to White pupils), and differences from the White reference group are, again, highly significant for most.

Table 4.4: Difference in percentage point likelihood of pupils with each respective characteristic being judged ‘above average’ at reading by their teacher, compared to pupils with the reference characteristic, and controlling for reading cognitive test score^

Income model	
Low-income (ref = higher-income)	-.111 (.014)***
Word Reading score	.010 (.000)***
Intercept	-.1.030 (.189)***
Gender model	
Boy (ref = girl)	-.041 (.013)**
Word Reading score	.011 (.000)***
Intercept	-1.115 (.192)***
SEN model	
SEN (ref = no SEN)	-.112 (.017)***
Word Reading score	.010 (.000)***
Intercept	-.931 (.190)***
Ethnicity model	
Indian (ref = White)	-.088 (.045)*
Pakistani (ref = White)	-.174 (.026)***
Bangladeshi (ref = White)	-.147 (.059)**
Black Caribbean (ref= White)	-.110 (.038)**
Black African (ref = White)	-.134 (.055)**
Word Reading score	.011 (.000)***
Intercept	-1.074 (.018)***
Language model	
Other languages (ref = English only)	-.123 (.021)***
Word Reading score	.011 (.000)***
Intercept	-1.061 (.018)***

N for each model = 4997 (unweighted). *** = p < .001; ** = p < .05; * = p < .10. Standard errors in brackets. ^All estimates weighted for survey design and for attrition to the main wave four survey, and controlled for age at cognitive test and time lag between test and teacher survey

Separate models estimate the likelihood of each pupil group being judged ‘below average’ at reading, and these result are presented in Table 4.5. They are entirely in line with findings ‘above average,’ inverting the direction of effect. As well as being 11 percentage points less likely to be rated ’above average’ by their teachers, for example, low-income pupils are 8.3 percentage points more likely to be judged ‘below average,’ and again, this is highly significant.

Table 4.5: Difference in percentage point likelihood of pupils with each respective characteristic being judged ‘below average’ at reading by their teacher, compared to pupils with the reference characteristic, and controlling for reading cognitive test score^

Income model	
Low-income (ref = higher-income)	.083 (.012)***
Word Reading score	-.009 (.000)***
Intercept	1.148 (.147)***
Gender model	
Boy (ref = girl)	.051 (.009)***
Word Reading score	-.009 (.000)***
Intercept	1.253 (.150)***
SEN model	
SEN (ref = no SEN)	.328 (.017)***
Word Reading score	-.007 (.000)***
Intercept	.817 (.137)***
Ethnicity model	
Indian (ref = White)	.044 (.040)
Pakistani (ref = White)	.089 (.029)**
Bangladeshi (ref = White)	.129 (.041)**
Black Caribbean (ref= White)	.096 (.038)**
Black African (ref = White)	.127 (.029)***
Word Reading score	-.009 (.000)***
Intercept	1.175 (.151)***
Language model	
Other languages (ref = English only)	.070 (.017)***
Word Reading score	-.009 (.000)***
Intercept	1.168 (.150)***

N for each model = 4997 (unweighted). *** = p < .001; ** = p < .05; * = p < .10. Standard errors in brackets. ^All estimates weighted for survey design and for attrition to the main wave four survey, and controlled for age at cognitive test and time lag between test and teacher survey

Biases in teacher judgements of pupils’ maths ability

In line with the lesser incongruity within the descriptive statistics, slightly fewer disparities emerge for maths (Table 4.6). No significant difference in teacher perceptions is found between MCS pupils speaking only English / speaking an additional language, and pupils of most ethnicities are as likely as White pupils scoring at the same level on the Progress in Maths test to be evaluated as ‘above average’.

However, inverting the relationship indicated for judgements of reading ‘above average,’ boys are *more* likely than girls to be judged relatively highly at maths. Sample Black Caribbean pupils are significantly less likely than their equivalently performing White counterparts to be judged ‘above average’ – along with children from low-income families, and those with any recognised SEN.

Table 4.6: Difference in percentage point likelihood of pupils with each respective characteristic being judged ‘above average’ at maths by their teacher, compared to pupils with the reference characteristic, and controlling for maths cognitive test score^

Income model	
Low-income (ref = higher-income)	-.106 (.016)***
Maths score	.039 (.001)***
Intercept	-1.160 (.018)***
Gender model	
Boy (ref = girl)	.050 (.012)***
Maths score	.041 (.001)***
Intercept	-1.191 (.018)***
SEN model	
SEN (ref = no SEN)	-.176 (.019)***
Maths score	.036 (.001)***
Intercept	-1.005 (.213)***
Ethnicity model	
Indian (ref = White)	.009 (.039)
Pakistani (ref = White)	-.045 (.027)
Bangladeshi (ref = White)	.083 (.043)
Black Caribbean (ref= White)	-.130 (.037)**
Black African (ref = White)	-.080 (.053)
Maths score	.041 (.001)***
Intercept	-1.178 (.017)***
Language model	
Other languages (ref = English only)	-.014 (.020)
Maths score	.041 (.001)***
Intercept	-1.171 (.017)***

N for each model = 4985 (unweighted). *** = p < .001; ** = p < .05; * = p < .10. Standard errors in brackets. ^All estimates weighted for survey design and for attrition to the main wave four survey, and controlled for age at cognitive test and time lag between test and teacher survey

Again, separate models estimate the likelihood of each pupil group of being judged ‘below average’ at maths (Table 4.7), and though more results again are non-significant here, those significant at p < .05 are entirely in line with findings ‘above average.’ Pupils with any diagnosis of SEN are *more* likely to be judged as ‘below average’ at maths compared to those without a diagnosis, low-income pupils are more likely than higher-income pupils, and Black Caribbean pupils more likely than White pupils.

Table 4.7: Difference in percentage point likelihood of pupils with each respective characteristic being judged ‘below average’ at maths by their teacher, compared to pupils with the reference characteristic, and controlling for maths cognitive test score^

Income model	
Low-income (ref = higher-income)	.071 (.014)***
Maths score	-.032 (.001)***
Intercept	1.153 (.173)***
Gender model	
Boy (ref = girl)	.022 (.013)*
Maths score	-.033 (.001)***
Intercept	1.150 (.175)***
SEN model	
SEN (ref = no SEN)	.353 (.020)***
Maths score	-.023 (.001)***
Intercept	.830 (.157)***
Ethnicity model	
Indian (ref = White)	-.018 (.027)
Pakistani (ref = White)	.008 (.035)
Bangladeshi (ref = White)	-.054 (.046)
Black Caribbean (ref= White)	.108 (.055)**
Black African (ref = White)	-.028 (.061)
Maths score	-.033 (.001)***
Intercept	1.157 (.026)***
Language model	
Other languages (ref = English only)	-.026 (.020)
Maths score	-.033 (.001)***
Intercept	1.157 (.175)***

N for each model = 4985 (unweighted). *** = p < .001; ** = p < .05; * = p < .10. Standard errors in brackets. ^All estimates weighted for survey design and for attrition to the main wave four survey, and controlled for age at cognitive test and time lag between test and teacher survey

Which characteristics underpin biases in judgements of reading and maths ability?

In order to begin to assess which characteristics might be important in driving these apparent biases and which stereotypes might be implicated, analysis now incorporates each predictor variable simultaneously in a comprehensive model, and is repeated separately for teacher judgements of ‘above average’ reading and maths. The sample is then split between boys and girls to investigate any variation in patterns according to gender. Table 4.8 presents reading results for the whole sample, followed by findings for boys and girls, respectively.

Table 4.8: Difference in percentage point likelihood of pupils with each respective characteristic being judged ‘above average’ at reading by their teacher, controlling for each other factor and reading cognitive test score^

	All (n = 4997)	**Boys (n = 2494)**	**Girls (n = 2503)**
Low-income (ref = higher-income)	-.086 (.015)***	-.056 (.018)**	-.115 (.021)***
Boy (ref = girl)	-.036 (.013)**		
SEN (ref = no SEN)	-.100 (.017)***	-.102 (.022)***	-.102 (.021)***
Indian (ref = White)	-.050 (.046)	.027 (.041)	-.146 (.071)**
Pakistani (ref = White)	-.095 (.035)**	.031 (.053)	-.195 (.060)**
Bangladeshi (ref = White)	-.068 (.061)	-.081 (.091)	-.059 (.074)
Black Caribbean (ref= White)	-.060 (.040)	.022 (.051)	-.165 (.062)**
Black African (ref = White)	-.071 (.056)	-.022 (.074)	-.126 (.076)*
Other languages (ref = English only)	-.038 (.028)	-.096 (.041)**	.007 (.048)
Word Reading score	.010 (.000)***	.009 (.000)***	.011 (.000)***
Intercept	-.934 (.189)***	-1.321 (.249)***	-.605 (.037)***

*** = p < .001; ** = p < .05; * = p < .10. Standard errors in brackets. ^All estimates weighted for survey design and for attrition to the main wave four survey, and controlled for age at cognitive test and time lag between test and teacher survey Ns are unweighted.

Though there is a general lessening in the magnitude of biases for each characteristic, all remain significantly related at the 5 percent level to teacher judgements of sample children’s reading, even when covariates are accounted for – though disparities by ethnic group appear to be moderated by the other factors, and language spoken is significant only for boys. Biases according to income-level and ethnicity appear generally to be stronger for girls, while, overall, boys remain assessed at a relatively lower level.

Table 4.9 presents results for teacher judgements of maths ‘above average.’ It suggests that gender may be key to teacher judgements of the maths ability and attainment of sample pupils (given the larger significant coefficient here than when gender is considered alone, without covariates [Table 6]). SEN status and income-level also remain significant predictors here, but biases for Black Caribbean boys and Black African pupils seem to be moderated by the covariates, and are non-significant. Accounting for confounders also renders the relationship between spoken language and teacher ratings non-significant and, in contrast to analysis for reading, there is some suggestion that biases in judgements for maths according to SEN status may be stronger for boys – though, overall, boys are more likely to be judged ‘above average’ at maths.

Across these analyses for reading and for maths there therefore appears to be some degree of bias according to each of four factors: income-level, gender, SEN status, and ethnicity – even accounting for every other factor, and for language spoken. Some differences in magnitude and significance are revealed according to gender among the MCS children, and relationships vary by academic domain. It seems, therefore, that stereotyping according to each of these four characteristics might underpin biases in teacher judgements of pupils, but that it may follow different trends according to subject area and gender.

Table 4.9: Difference in percentage point likelihood of pupils with each respective characteristic being judged 'above average' at maths by their teacher, controlling for each other factor and maths cognitive test score^

	All (n = 4985)	Boys (n = 2491)	Girls (n = 2493)
Low-income (ref = higher-income)	-.093 (.017)***	-.088 (.022)***	-.103 (.023)***
Boy (ref = girl)	.072 (.012)***		
SEN (ref = no SEN)	-.181 (.020)***	-.223 (.025)***	-.121 (.024)***
Indian (ref = White)	.008 (.036)	.013 (.069)	.023 (.059)
Pakistani (ref = White)	-.007 (.038)	.018 (.065)	-.021 (.053)
Bangladeshi (ref = White)	.106 (.047)**	.177 (.073)**	.045 (.084)
Black Caribbean (ref= White)	-.074 (.041)*	-.013 (.052)	-.143 (.055)**
Black African (ref = White)	-.064 (.050)	-.016 (.069)	-.105 (.079)
Other languages (ref = English only)	-.002 (.027)	-.046 (.046)	.031 (.045)
Maths score	.034 (.027)***	.034 (.002)***	.035 (.002)***
Intercept	-1.027 (.211)***	-1.138 (.268)***	-.794 (.308)**

*** = p < .001; ** = p < .05; * = p < .10. Standard errors in brackets. ^All estimates weighted for survey design and for attrition to the main wave four survey, and controlled for age at cognitive test and time lag between test and teacher survey Ns are unweighted.

Robustness checks

Four discreet robustness checks have been carried out to ensure that choices in modelling, weighting and sample selection have not influenced the overall findings presented in this chapter. Firstly, analyses are repeated using binary logistic rather than linear probability models. Results are equivalent (1). Secondly, as there is some missing data on the variable accounting for time lag between cognitive test and teacher survey (cases are incorporated in the main analysis using a missing category), analyses are repeated without those pupils for whom information is missing here. This makes little difference to the direction, significance or magnitude of findings (2).

As the pupils in the teacher sample are unevenly distributed across schools, and because some schools have several pupils and others only a single child, an additional check is carried out to examine whether extreme groups of teachers in more populous schools may be driving results. The pool for analysis is restricted to one pupil per teacher, and few differences are made to overall findings (3). Lastly, analysis is carried out *without* the wave four main sample weights, but *with* clustering at the school level. Again, the overall findings hold (4).

As an example, Table 4.10 shows key coefficients across these different analyses (1-4) for teacher judgements of reading ‘above average,’ according to the first model (Table 4.4).

Table 4.10: Difference in likelihood of pupils with each respective characteristic being judged 'above average' at reading by their teacher, compared to pupils with the reference characteristic, controlling for reading cognitive test score: robustness checks^

	Original results (*B*)	**Check 1: Logistic model (*Difference and p value for difference in model-predicted probabilities*^^)**	**Check 2: Excluding cases with data missing on time lag (*B*)^^^**	**Check 3: One pupil per teacher (*B*)**	**Check 4: Unweighted; clustered by school (*B*)^^^^**
Low-income (ref = higher)	-.111 (.014)***	-.129 (.014)***	-.111 (.015)***	-.101 (.018)***	-.116 (.014)***
Boy (ref = girl)	-.042 (.013)**	-.048 (.012)***	-.043 (.013)**	-.027 (.015)*	-.044 (.011)***
SEN (ref = no SEN)	-.112 (.017)***	-.178 (.019)***	-.107 (.017)***	-.118 (.020)***	-.116 (.014)***
Indian (ref = White)	-.088 (.045)*	-.074 (.038)*	-.106 (.046)**	-.062 (.035)*	-.083 (.031)**
Pakistani (ref = White)	-.173 (.026)***	-.168 (.025)***	-.180 (.027)**	-.156 (.033)***	-.189 (.023)***
Bangladeshi (ref = White)	-.147 (.059)**	-.171 (.068)**	-.171 (.057)**	-.175 (.072)**	-.146 (.051)**
Black Caribbean (ref = White)	-.110 (.038)**	-.121 (.040)**	-.120 (.043)**	-.047 (.044)	-.102 (.041)**
Black African (ref = White)	-.135 (.055)**	-.128 (.050)**	-.132 (.059)**	-.160 (.063)**	-.183 (.039)***
Other languages (ref = English only)	-.123 (.021)***	-.122 (.019)***	-.131 (.020)***	-.103 (.020)***	-.131 (.017)***
N	4997	4997	4641	2995	4997

*** = p < .001; ** = p < .05; * = p < .10. Standard errors in brackets. ^All estimates bar Check 4 weighted for survey design and for attrition to the main wave four survey. All bar Check 2 controlled for age at cognitive test and time lag between test and teacher survey. Ns are unweighted.

^^Calculated using "margins, pwcomp (eff)" in Stata 13. ^^^See Annex 4B for coefficients of age and timing controls ^^^^Robust SEs estimated using "vce (cluster)" in Stata 13.

Additional analyses

An earlier version of this chapter has been published as a working paper (Campbell, 2013b) and includes a number of further explorations and checks. They are not reported in detail here, because they are peripheral to (and congruent with) the overall message of this chapter – but, briefly, they include the following.

In order both to allow the cognitive test score predictors to be non-linear and to investigate whether disparities are particularly pronounced for pupils scoring at a certain level, categorical quintile versions of the test score are interacted with each characteristic of interest, in respective single-characteristic models. Many of the biases seem to be strongest for children scoring around the mean on each test, perhaps because there is less discrimination in manifest ‘ability’ for these pupils, evoking a tendency among teachers to draw more strongly on other information such as stereotypes in these average cases. However, there are some exceptions to this general trend – biases for pupils with a diagnosis of SEN are significant at all test quintiles, and biases for higher-scoring Black African and low-income sample children appear to be particularly strong (see Campbell, 2013b, p.39–43).

The MCS was sampled according to a stratified, disproportionate cluster design (Plewis, 2007) which oversampled areas with high numbers of minority ethnic families and areas with high levels of deprivation. Sample pupils with particular characteristics are concentrated in certain areas. For example, children from some minority ethnic groups, and those who speak languages in addition to English, are clustered. Lower-income pupils are also disproportionately represented in particular regions. It is possible, therefore, that the seeming biases may arise from variation across local practices, tendencies and perceptions – rather than from stereotypes relating to key pupil characteristics at the level of the teaching profession. To test whether this is the case, analysis by each characteristic is repeated with controls for Government Office Region – and results hold according to this specification (see Campbell, 2013b, p.44–53).

To explore whether biases may largely be attributable to teachers of pupils in relatively homogeneous, more wealthy areas, where there are – for example – fewer minority ethnic pupils, rather than to those in more diverse areas, where teachers have a wider experience of pupils with a variety of characteristics, a further specification restricts the analytical sample to those children born in ‘disadvantaged’ or ‘ethnic minority’ wards at MCS wave one (see Plewis, 2007, for further details of ward make-up). Analysis here controls also for GOR, and though some biases are slightly reduced for pupils in these more heterogeneous strata, most remain – offering support to the hypothesis of consistent, pan-area stereotyping of pupils (see Campbell, 2013b, p.44–53).

Summary and discussion

Analysis set out to explore the possibility that biases in teacher judgements of pupils may result from systematic stereotyping and that these biases might contribute to variation in recorded attainment among primary school children. It finds that, in this sample of English seven-year-olds, there are inequalities in teacher perceptions of pupils’ reading and maths ‘ability and attainment’ which correspond to each of the key pupil-level characteristics delineating these achievement gaps. On average, low-income pupils seem to be rated less favourably by their teachers, along with pupils with any SEN diagnosis, non-White pupils, pupils speaking languages in addition to English, and boys (reading) / girls (maths). Because both independent measures of pupil test performance and indicators of teacher perceptions of pupils which are not required by or implicit with formal in-school assessments are used in this chapter, findings support the possibility that the socio-cognitive process of stereotyping may indeed be instrumental in constructing attainment differentials.

Limitations and alternative explanations

Though results here are congruent with previous research indicating relative over- and under-assessment of pupils according to their characteristics, it remains feasible that there are supplementary or alternative explanations for results, and that patterns may to some extent be an artefact of measurement error

The latter possibility is mitigated to a degree by the robustness checks and analyses mentioned earlier: particularly by modelling which interacts test score level with characteristic of interest. Because the direction (though not the magnitude) of effect holds across levels, and tends often to be strongest around the mean, this lessens the possibility that it is error at the outskirts that might drive patterns.

It is, however, logically viable that teachers are not, in fact, biased: that assessments by teachers participating in the MCS are actually more 'accurate' compared to cognitive test performance, or that the two measure different things – if cognitive tests favour the groups that seems to be judged less favourably by their teachers. Potentially, for example, low-income children may underperform relative to their capacity when at school, compared to higher-income children. If this explanation is to any degree legitimate, it raises a host of additional questions and begs further research into the processes that could lead to systematically depressed in-school performance among all those groups regularly reported as under-attaining.

Previous research belies the likely primacy of this explanation however: particularly Burgess and Greaves' (2009) work, which finds biases in the same direction as analyses here, according to similar characteristics. Even in their study, when comparative measures are both situated within the education system, and KS2 written tests are compared to KS2 teacher assessments, discrepancies that complement those described in this chapter are reported. Therefore, as certain groups of children seem to be perceived less favourably relative both to their performance on in-school and on independent tests, it seems likely that bias lies within the perceptions of their teachers, rather than in the children's capacities as manifest across these different test situations.

Stereotyping as explanation

Notwithstanding the possibility of alternative or additional explanations, therefore, findings in this chapter support the prospect that mechanisms of stereotyping, beyond the level of the individual pupil and their family, and outside of the control of the child or their parents, appear to be at work determining assessment levels awarded and recorded pupil attainment. Unless these tendencies are addressed, they may continue to play some part in creating and perpetuating inequalities.

Analysis here also began tentatively to unpick the constitution of the stereotypes proposed to explain biases. It finds that income-level, gender, SEN status, ethnicity (and, to a lesser extent, language spoken) all appear to play a part in accounting for disparities in judgement of sample pupils, and that there is some variation by gender and by subject domain. This suggests that any intervention aimed at alleviating stereotyping and its effects on teacher perceptions and assessments may need to take account of the complex nature of the process and of its components, rather than simply targeting biases associated with one characteristic in isolation.

It should be noted that findings and conclusions in this chapter do not serve as any condemnation of teachers – as a profession or as individuals – as enacting the process of stereotyping to any unusual (or to any deliberate) degree. Stereotyping is conceived to be a universal, non-conscious, automatic cognitive function which enables speed and efficiency in thought and behaviour. According to theory, all individuals have a propensity to enact this function to some degree: there is no reason that teachers should be exempt, nor unusually prone. Bias in judgements of pupils is just one manifestation of the human tendency to stereotype.

Where might stereotypes of pupils originate?

Analyses using the MCS cannot indicate what may be creating and forming the stereotypes that seem to provide a normative template for skewed teacher perceptions, and there are a number of possible explanations. Firstly, it is feasible that the expectations of different groups of children that are made pertinent to teachers through explicit characteristic-based regulation of pupil, teacher and school performance levels (Bradbury, 2011b) might reify and reinforce differentiated notions of potential and ability which become embedded and self-fulfilling.

Secondly, the messages conveyed by the various policy initiatives which require schools and teachers to focus on selected pupil groups might perpetuate an assumption that these groups are fundamentally lacking. For example, the current concentration on low-income families through the pupil premium may inadvertently imply and contribute to a stereotype that poorer pupils across the board are deficient in ability and potential. Similarly, recent initiatives targeting certain ethnic groups (Maylor *et al*., 2009; Tikley *et al.*, 2008) might build a sense that these groups are essentially less capable, and feed into differentiated expectations.

Thirdly, as suggested by Burgess & Greaves (2009), direct personal experience might inform the process of stereotyping. Teachers may form generalised templates through their everyday experiences and interactions with pupils, and if a proportion of children from a given group are observed to perform in a certain way, a teacher may form a stereotype and over-generalise to all children in this group.

Lastly, of course, teachers function not only within schools and the education system but also within wider society. Media and other discourses regarding the societal positioning and features of different social groups may create stereotypes of these groups, potentially seeping into and influencing teachers' perceptions of the children in their classroom.

Unfortunately, the data used in this chapter do not offer the possibility of testing the extent to which any or all of these potential mechanisms play a part in developing the stereotypes which appear to be held by teachers, and the interrelationships between teachers and the systems and structures within which they function cannot be established here. There may conceivably be a number of points and means of intervention through which stereotyping of pupils could be mitigated, but findings from this chapter initially support one in particular: addressing and confronting the process at the teacher-level.

Tackling stereotyping

It has long been argued that self-awareness of perceptions and expectations, and self-reflectiveness, are crucial to effective teaching:

> …for teachers to optimise learning they need to have a greater awareness of the complexities of individual differences [and] the importance of perceptions and expectations of pupils on learning outcomes…(Hallam & Ireson, 1999)

Earp (2010) reviews the cognitive-psychological literature on stereotype activation and consequential behaviours and also argues (here, in relation to stereotyping according to ethnicity) for mindfulness:

> A teacher who is unaware of the basis for her judgements may conclude that they stem from the realities of her student's performance, rather than (directly or indirectly) from the activation of stereotypes about that student's [ethnic] group.

Discussing the research on ways in which teachers may thwart the stereotyping process, Earp suggests that, 'Teachers are just the sort of people who are in a position to automate egalitarian motives,' and describes how recent cross-disciplinary studies have indicated that it is feasible that teachers may, with time and effort, 'train' and tame the stereotyping mechanism. Potentially this might involve actively learning to draw on alternative stereotypes of pupils, to presume motivation and ability in each student, and / or consciously and deliberately to be balanced and constructive in feedback to and interactions with all pupils. Earp concludes that,

> ...it is essential that schools of education include in their curricula state-of-the-science resources on the unconscious nature of prejudice and the corresponding implications for [the] classroom.'

Though it provides the beginnings of suggestions for change, this existing literature is limited regarding the exact means by which teachers, managers and policy-makers may effectively intervene to alleviate the stereotyping process. The current chapter suggests, however, that this is an area very much worthy of further investigation and trial. Increased credibility and importance should be given to the accumulating evidence that biased judgements and stereotyping might be impacting upon and shaping pupil experiences and attainment, and resources and efforts should be concentrated upon addressing this possibility. Extending the current analysis to further explore, unpick and test the drivers of the patterns it has found should play a part in this. The investigation in this chapter uses just a sample of children (albeit a relatively large one), so tendencies found particularly in the results regarding the various characteristics appearing to underpin stereotypes should be explored further, in enhanced and alternative datasets. The data used in this chapter are moreover extremely limited in the extent to which they can examine any role of differential school-level tendencies in creating or mitigating the biases suggested; this should also be an area for further research.

At the policy level, consideration should be given and examination instigated into the ways in which initiatives and communications might create or reinforce overgeneralised normative templates and result in unintended consequences. If, as speculated and as beginning to be evidenced characteristics-based monitoring inadvertently perpetuates attainment differentials based upon these characteristics (Bradbury, 2011a; 2011b), this would be a point for intervention and reformulation. Similarly, if ostentatious implementation of targeted policies, such as the pupil premium, proves detrimental to the treatment of its recipients, this again suggests reconsideration and revision of methods. Finally, the recent encouragement of work-based initial teacher training (in contrast to the university-based model) (Allen *et al*, 2014) may be considered in light of the findings in this chapter. If a trainee learns predominantly from the practices and norms in their placement school, with less time devoted to critical pedagogical theory, might this serve only to reinforce active stereotypes and expectations, with less scope for new ideas and the challenging of norms and preconceptions?

Conclusion

This chapter finds evidence for unfounded biases in teacher judgements of pupils, and that efforts to ensure parity, equality and meritocracy in the education system have not yet resulted in parity of perception and judgement. Resources might usefully be directed as suggested here: towards building the evidence base on stereotyping; towards developing relevant interventions and strategies within teacher training and professional development; and towards avoiding the inadvertent reinforcement of stereotypes and the worsening of their effects during policy intervention and associated publicity. By recognising and challenging the existence and effects of stereotyping in these ways, it is possible that some of the long-standing and widespread inequalities among primary school children may come to be alleviated.

Chapter 5

In-class ability grouping and the relative age effect

Introduction

Month of birth and academic attainment

In England, as in many other countries, the vast majority of pupils are educated within class groups formed according to the structure of the school academic year. Annually, pupils born over the period beginning in September and ending in August will, with a very few exceptions, comprise a distinct cohort (Riggall & Sharp, 2008).

There is a mounting body of international evidence which indicates a relationship between month of birth, school year structure, and a variety of academic and extra-academic outcomes. Pupils who are younger in the school year (in England, those born during the summer months) tend consistently, throughout compulsory education, to score lower on tests of academic ability than their relatively older peers (Bedard & Dhuey, 2006; Boardman, 2006; Crawford *et al*, 2007; Crawford *et al*, 2011; Daniels *et al*, 2000; Department for Education, 2010b; Lawlor *et al*, 2006; Martin *et al*, 2004; McEwan & Shapiro, 2008; Menet *et al*, 2000; Oshima & Domaleski, 2006; Strom, 2004, Sykes *et al*, 2009). They are more often diagnosed with special educational needs (Crawford *et al*, 2007; Department for Children, Schools and Families 2009d; Department for Education, 2010b; Gledhill *et al*, 2002; Goodman *et al*, 2003; Martin *et al*, 2004; Polizzi *et al*, 2007; Sykes *et al*, 2009 ; Wallingford & Prout, 2000; Wilson, 2000), and progress less frequently into further education (Bedard & Dhuey, 2006; Crawford *et al* 2011; Sampaio *et al*, 2011; Sykes *et al*, 2009). Relatively younger children are also disproportionately likely to report bullying victimhood, to demonstrate lower levels of confidence and self-efficacy, and to report lesser enjoyment of school (Crawford *et al*, 2011; Department for Education, 2010b; Mühlenweg, 2010).

Theories to date on possible causes of relative age disparities have spanned the biological and social sciences. At a biological level, it has been suggested that pre-natal seasonal variations may influence the development of infants in the womb and subsequent post-natal progress (Foster & Roenneberg, 2008; Polizzi *et al*, 2007; Sharp *et al*, 2009). However, international evidence from countries whose school entry cut-off points fall in different seasons, but whose relatively youngest pupils are equivalently disadvantaged, precludes any possibility that seasonally-related biology is key to explaining month of birth gradation across each school year group. Regardless of the relationship between school year cut-off points and the cycle of the seasons, internationally, it is the relatively youngest within each year group cohort who do worst across a range of academic and extra-academic outcomes and experiences (Sharp *et al*, 2009).

Much recent UK research has therefore focused upon exploring and isolating the potential contributions to birth-month attainment variation of drivers related to the high-level structure and shape of the education system. Absolute age differentials have been proposed as influential (given that a system based around annual year groups means that August-born pupils are up to a year younger than September-borns on national assessments), and length of schooling has also been mooted (given that, in the past, local authorities have differed substantially in their policies on exact point of admission after a child's fourth birthday, which means that some relatively younger children received fewer terms of formal education than their older counterparts) (Crawford *et al*, 2013b). This research has tended to conclude that absolute age appears to play the greatest part in explaining variation, so suggests that, given a consistent, linear relationship between age at test and test score, age adjustment of tests scores could play a part in mitigating the month of birth effect (Crawford *et al*, 2014).

However, research has not yet completely accounted for the effect in its entirety (Crawford *et al*, 2014), and the practicalities and effects of any application of test-score adjustment remain largely to be trialled at scale and in context. Additionally, as most studies to date are based upon interpretation of observational data, exploration of this data for additional and complementary explanations for the formation of relative age differentials continues to be useful. It is possible that there are other factors contributing to the apparent, direct relationship between birth month and academic performance, and it is possible that the creation of disparities may most effectively be tackled using a combination of means premised on different understandings of their formation.

Moreover, while nuanced application of age-adjustment of test scores in both formative and summative contexts seems to offer the beginnings of an eradication of the relationship between birth month and attainment, it may not compensate fully for the many non-academic experiences and outcomes which are also associated with being relatively younger within the school year (Crawford *et al*, 2011): experiences and outcomes such as tendency to be bullied, and wellbeing, which are inherently as important to children's lives as their eventual attainment. A more detailed consideration of the construction of the relative age effect, and of the points at which and routes through which it might manifest, will therefore add to an overall understanding of its entirety, and of potential solutions.

Therefore, by examining early factors that contribute to the birth month differences, it may be possible to suggest additional interventions that provide some prospective alleviation of the current disadvantages experienced by pupils born later in the school year. Accumulating support for a pervasive, multi-faceted effect (Crawford *et al*, 2011; Department for Education, 2010) suggests that adjustment of various aspects of its manifestation can best hope to compensate all of its long-run, many-dimensional influences.

An additional motivation for investigation of the causes of month of birth effects is the consideration that differences by birth month may not, in fact, be the most important element in the story conveyed by their stark variation. Month of birth disparities are a useful frame within which to view the English educational system, because they allow a relatively clear lens through which any socio-structural and psychological forces that create gaps among children can be examined. Month of birth is not yet loaded with historical, societal and psychological assumptions and preconceptions in the way that other characteristics (such as pupil gender or ethnicity) may be. Investigation of causes of differences by birth month can therefore help articulate and shed light on key systems and structures that create difference or disadvantage among all children – and practices that make a difference to the relatively young might also have an effect on other between-group attainment differentials.

Early maturational inequalities as potential cause

A number of studies have suggested that the relative social, emotional, behavioural and / or cognitive immaturity of summer-born pupils in early primary school may be key to laying the foundations for inequalities (Boardman, 2006; Sharp *et al*, 2009). Most pupils in England enter primary school at some point during the year following their fourth birthday (Riggall & Sharp, 2008). At this stage, and throughout their early education, the in-cohort age difference of up to a year between relatively younger and relatively older pupils comprises a significant fraction of life lived, and of development.

The possibility, therefore, is that these early maturational inequalities (necessitated both by the structure of the annual cohort-based educational system and the young age at which children first enter schooling) are instrumental in creating the relative age effect. This theory is supported by research which indicates that younger pupils may disproportionately frequently be diagnosed with special educational needs on the basis of relative developmental immaturity, rather than any inherent trait difference (Dhuey & Lipscomb, 2010; Elder & Lubotsky, 2009; Gledhill *et al*, 2002; Wallingford *et al*, 2000). In addition, analysis of international evidence by Sprietsma (2010) begins to suggest that ability grouping (where groups are constructed on the basis of performance / perceived ability relative to cohort peers) may account for some of the attainment variation associated with month of birth.

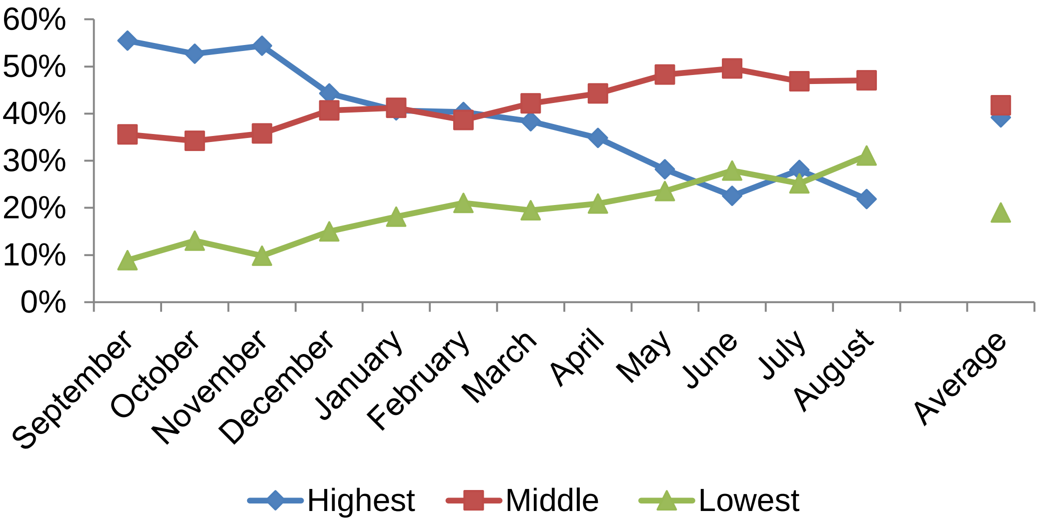
In-class ability grouping and month of birth

Yet, until very recently, a dearth of large-scale national-level data on in-school ability grouping practices has meant that investigation of their potential contribution to the month of birth effect has been constrained. The National Pupil Database does not contain information on whether a pupil is ability grouped, and previous representative surveys have not collected information on these practices. Likewise, with the exception of Sprietsma's (2007) work, there is scant international evidence in the area (Sharp *et al*'s 2009 international literature review presents no studies specifically examining this issue, nor does Sykes *et al*'s [2009] *English-evidence-based birthdate effects: A review of the literature from 1990-on*). Some very dated studies exist (for example, Jinks' 1964 analysis of a single borough's 11-year-olds suggested that pupils relatively younger in the school year tended to be found in lower streams) - but exploration of whether in-class ability grouping may contribute to recent birth month attainment differentials, in England, has only lately become possible, using MCS data.

Analysis of 2008 data for British seven-year-olds who are participating in the MCS shows that, across both whole-year and in-class grouping practices, relatively younger pupils are disproportionately frequently placed in lower groups, while their relatively older peers are more often found in the highest placements. This tendency is consistent across all practices, and steadily, linearly-incrementally related to birth month (Campbell, 2013a; Hallam & Parsons, 2012).

The working paper that informs this chapter reports that 78.8 percent of the 5,374 English MCS children are subject to an overriding, high-level within-class ability grouping (Campbell, 2013a). It shows that among these pupils, September-born children are more than twice as likely as August-born children to be placed in the highest group, with the inverse being the case for the lowest grouping (Figure 5.1) There is strong evidence, therefore, that a large proportion of pupils are in-class ability grouped at a very early age, and there are indications of major disparities in placement according to relative age within cohort. This lends initial support to a theory that early in-class ability grouping, at a stage where absolute age differentials are highly pronounced, may be influential in the creation of the month of birth effect.

Figure 5.1: Percentage of pupils born in each month who are reported as being in each within-class ability group, among those pupils who are reported as being within-class grouped*



(n = 4140) *From Campbell (2013a)

In-class ability grouping and academic attainment

As detailed in Chapter 3, the wider research on the associations between ability grouping and pupil attainment have generally suggested that grouping entrenches between-pupil difference and may have a detrimental effect on pupils placed at lower levels, while advantaging children who are in higher groups (Blatchford *et al*, 2008; Hallam & Parsons, 2012; Kutnick *et al*, 2005; Dunne *et al*, 2007; Blatchford *et al*, 2008). There is also evidence that pupils' positions within in-school hierarchies have tended largely to be stable over time (Blatchford *et al*, 2008). In-class ability grouping in early primary school may, therefore, establish a structured hierarchy which is predicated on birth month and which embeds differentiated trajectories of academic achievement.

In the working paper which predates this thesis, a theoretical model is proposed and explained (see Annex 5A), where the initial birth month disparity in within-class group position may play out in as a disparity in eventual attainment via three possible routes: through pupils' self-perceptions, as engendered by their in-class position; through the educational and assessment opportunities offered to pupils placed at different in-class levels; and through teacher perceptions, expectations of, and behaviours towards pupils situated in different groups (Campbell, 2013). The investigation presented in the current chapter begins to explore this third hypothetical channel.

Teacher perceptions and academic attainment

Teacher perceptions and judgements play a crucial part in pupils' progress and achievement. There is a solid body of evidence which indicates that teacher perceptions of, expectations of, and beliefs about their pupils can influence attainment, and lead to self-fulfilling prophesies (e.g. Rosenthal & Jacobsen, 1968; Rubie-Davies, 2010).

Research has also indicated that teacher judgements of their pupils can relate to the groups of which children are members - groups which may bear little or no necessary relationship to a child's capability or potential (Harlen, 2004). Most importantly, there is also evidence that teacher perceptions of pupil ability and attainment are gradated according to birth month, with August-born pupils tending to be judged as less able by their teachers, and September-borns as more able.

Crawford *et al* (2011) indicate that, at age seven, relatively younger pupils are more likely to be judged by their teacher as of 'below average' ability in reading, writing and maths, while Crawford *et al* (2013a) use national data to show a steady downward September-August trend in the grades allocated through the teacher assessed component of Key Stage Two tests. Unless there truly is a difference in pupil ability which corresponds, expediently, to the structure of the cohort-based educational system, this indicates a fundamental bias in teacher assessments of children according to their birth month – a bias which may further be confounded by the unequal distribution of pupils born in different months across in-class ability groups.

The current study

To investigate whether in-class ability grouping is, as hypothesised, instrumental in the construction of the relative age effect, the current chapter therefore focusses on the mediating pathway of teacher perceptions of pupil ability, and examines whether birth month gradation in these perceptions is greater where there *is* in-class ability grouping than where there *is not*. If there is *no* difference in magnitude of variation, then in-class grouping will not be indicated as a mechanism in the creation and proliferation of the effect. If variation in teacher perceptions according to birth month *is* more pronounced where in-class ability grouping takes place, and given indications that teacher perceptions may affect pupil attainment, then in-class ability grouping will begin to be implicated as playing a part in the formation of the relative age effect.

Therefore, the hypothesis being tested is that: birth month gradation in teacher perceptions of pupil ability will be more pronounced among pupils who are in-class ability grouped than among pupils who are not in-class grouped.

Methodology

Sample

As in the previous chapters, analyses here use 2008 data on seven-year-old, English Millennium Cohort Study (MCS) children, and the children's teachers. Only MCS children surveyed in England are included, so that, in line with the assumption that the structure of a school system underpins associations between month of birth and child outcomes, findings apply within a single educational framework with consistent school year cut-off points.

Twins and triplets are removed from analyses, because in-class groupings and teacher judgements for these pupils may be subject to different tendencies compared to singleton children. This leaves a base total of 5,481 English seven-year-old pupils with returned teacher surveys. There are some variations in sample sizes across analyses due to missing data; exact numbers are stated throughout reporting.

Unweighted data are used for the main analyses in this chapter, but additional, weighted alternative specifications are also reported in the results section. All MCS data used for analyses here are publically available and can be downloaded at http://www.esds.ac.uk/.

Key measures

The two key predictor variables used in analyses are pupil season of birth and teacher report of whether the pupil is in-class ability grouped, or not. The outcome variable is teacher assessment of whether the pupil is of *above average* 'ability and attainment' at a given subject.

The *season of birth* predictor combines month of birth into four categories (autumn, winter, spring, summer), in order to ensure robust sample sizes for modelling. Autumn comprises those born in September, October, or November (27.3 percent of the sample); Winter: December, January, February (25.2 percent); Spring: March, April, May (24.3 percent); Summer: June, July, August (23.2 percent). As detailed in the results section, and in line with the linear incremental associations demonstrated throughout relative age research, this amalgamation of months into seasons does not affect the direction of findings.

The ability grouping predictor variable derives from a question in the wave four teacher survey which asks whether, at age seven, 'In this child's class, is there within-class ability grouping?' – having defined within-class ability grouping as follows:

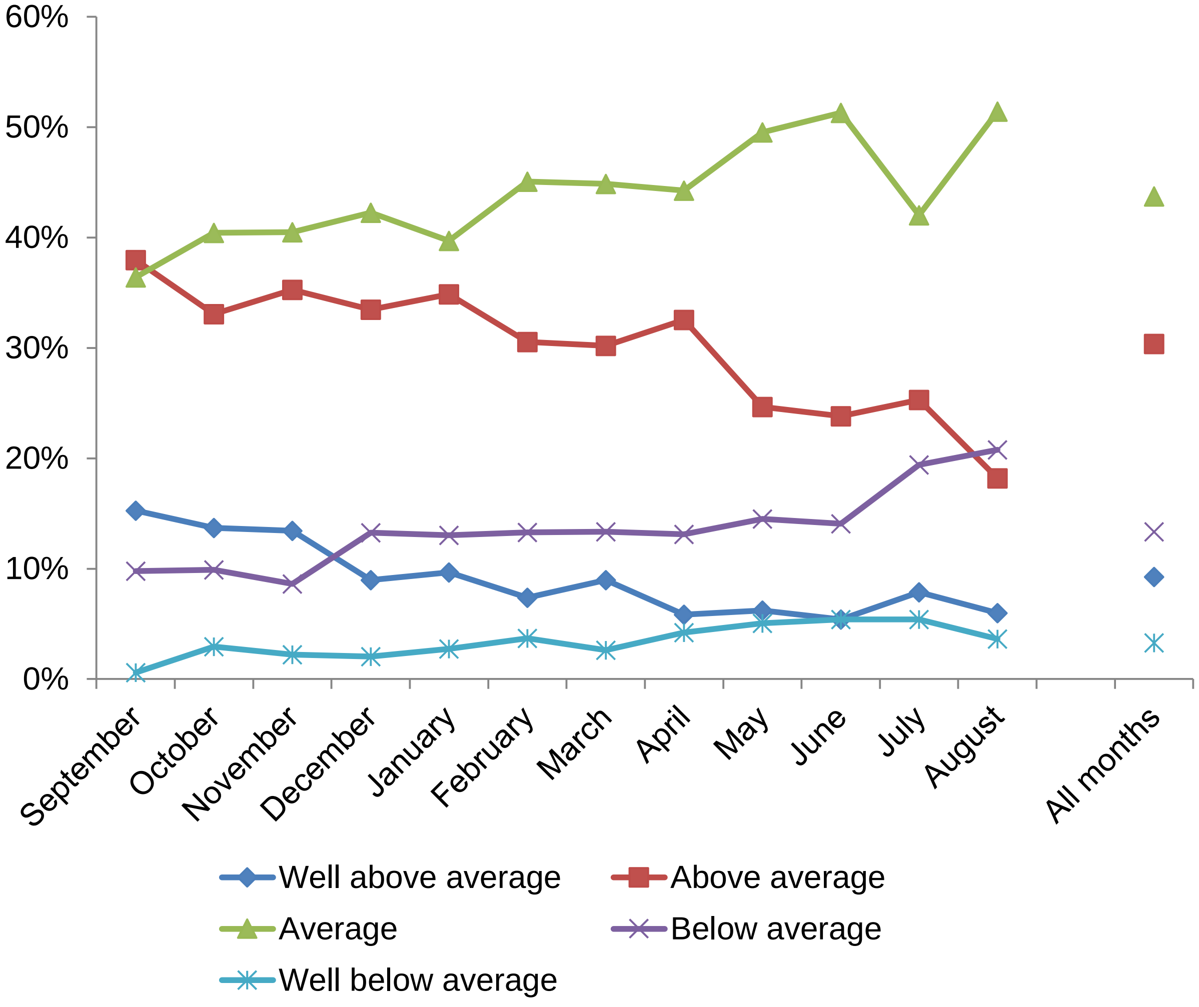
> Some schools group children within the same class by general ability and they are taught in these ability groups for most or all lessons.

Respondents provided a *yes / no* answer to this question, and this is used as a binary 1 / 0 variable in analyses. (See Annex I for further details of questionnaire wording.) 79 percent of the base sample pupils are reported as being in-class grouped.

The outcome variable derives from a question in the teacher survey asking the respondent to 'rate some aspect of the study child's ability and attainment…in relation to all children of this age…' As described in Chapters 3 and 4, and Annex II, teachers could rate children as *well above average, above average, average, below average,* or *well below average.* Teachers were asked their opinion on children's *ability and attainment* in the following domains: speaking and listening; reading; writing; science; maths and numeracy; physical education; information and communication technology; and expressive and creative arts. See Annex 5B for a breakdown of teacher responses in each domain for all sample pupils. The results presented in this chapter are for the first four domains on which teachers were questioned: speaking and listening, reading, writing, and science.

In each subject domain, there is an overriding month of birth gradient in teachers' ratings of pupils' *ability and attainment,* where relatively older children are more likely to be judged *well above average* or *above average*, and relatively younger children are more likely to be judged *average, below average,* or *well below average.* Figure 5.2 illustrates this for judgements of speaking and listening *ability and attainment.*

Figure 5.2: Percentage of sample pupils born in each month judged to be at each level of *speaking and listening* ability and attainment by their respective respondent teacher



(n = 5429)

This five-category teacher judgement outcome variable is recoded to be binary, so that 1, ‘above average,’ combines teacher responses of *well above average* and *above average*, and 0, ‘average or below,’ combines responses of *average, below average,* or *well below average.* This focusses analysis on disproportionalities and patterns in positive, favourable judgements of pupils.

Analytical approach

Linear probability regression is used to model the relationships between birth season, ability grouping, and whether teacher judgement is ‘above average.’ All main analyses in this chapter use the *Generalised Linear Modelling* option in PASW (SPSS) 18. Linear probability regression has been used in some of the most recent research into relative age effects (Crawford *et al*, 2013b) and is chosen for analyses here because the model-predicted probabilities offered are more straightforwardly interpretable than the odds ratios produced by a logistic regression. However, as a check, equivalent analyses have also been performed using the latter technique, and do not affect results (an example is described in the results section).

Because analyses investigate whether the relationship between season of birth and teacher perceptions varies *according to* whether pupils are ability grouped or not, an interaction between these two predictors is key to each model (SoB x Gr), and included along with the season of birth (SoB) and ability grouping (Gr) predictors. The basic equation underpinning all analyses is therefore:

$$y = a + \beta_{123}SoB + \beta_4 Gr + \beta_{567}SoB \times Gr + e$$

The reference categories in each analysis are set as *summer* and *not grouped*. Therefore, given the inclusion of the interaction, the first three coefficients in the equation describe the relationships between likelihood of being judged ‘above average’ and birth season (autumn, winter, or spring – in comparison to the summer reference) among pupils who are *not grouped*. The fourth coefficient describes the relationship between being grouped and probability of being judged ‘above average’ for summer pupils. The fifth coefficient, for the interaction, isolates the association between being grouped and being judged ‘above average’ for autumn-born pupils (and the sixth and seventh for winter and spring-born pupils). Key coefficients are described throughout the results section, alongside graphs which illustrate model predicted probabilities (estimated marginal means) of being judged ‘above average’ for pupils born in each season who are grouped and not grouped.

Stages two, three and four: addition of controls

Because any difference in the relationships between being born in the autumn / summer, being grouped or not, and teacher perceptions may be due to selection of pupils with different family backgrounds and individual characteristics into schools which group / do not group, a second stage of analysis adds controls for a range of pupil- and family-level factors. Table 5.1 describes the variables included at this second stage.

Table 5.1: Controls added cumulatively to each model at stages two, three and four

Stage two: pupil and family controls	Stage three: school and teacher controls	Stage four: previous in-school assessments of pupil
Pupil gender	Whether school at wave four same school child attended two years previously	Total (teacher-assessed, age five) Foundation Stage Profile score
Pupil ethnicity	Whether family pays fees for schooling	Teacher report of any identified special educational need
Pupil age five British Ability Scale (age-) standardised T-scores (Pattern Construction, Picture Similarity, Naming Vocabulary)[12]	Whether family displayed religiosity for school admission	
Family income at age seven	Whether there are mixed year groups in child's class	
Family housing tenure at age seven	Number of classes in child's year group	
Whether languages other than English are spoken in pupil's home at age seven	Number of pupils in child's class	
Main parent's highest academic qualification at child's birth	Respondent teacher's gender	
Main parent's highest vocational qualification at child's birth	Number of years respondent teacher has taught	
Whether a single parent when child was born	Number of years respondent teacher has taught at this school	
Whether internet is available in family home at age seven		
Whether, and length of time for which, pupil was breastfed		

[12] These tests are postulated by their developers and by some users as providing (respective) indications of spatial ability, pictorial reasoning ability, and verbal ability, which, together, measure a latent, absolute, stable trait of 'general conceptual ability' (see Hill, 2005). In these analyses, they are simply assumed to indicate prior performance on cognitive tests, at age five, around the time a pupil entered primary school.

Even controlling for the factors included at this second stage, it is still possible that there are other, systematic, school- or teacher-level differences between grouping / non-grouping establishments which influence teacher perceptions. Stage three therefore attempts to account for this, by adding further controls available in the MCS (see Table 5.1, and Annex 5C, which details each variable, its origin in the MCS surveys, and its distribution in the sample, in greater depth).

Lastly, a fourth stage adds additional controls for previous in-school assessments of pupils, which serve two potential purposes, each premised on a separate assumption.

The MCS data contain no information on the point at which ability grouping commenced for sample pupils, so, firstly, based on an assumption that pupils are grouped at school entry, stage four provides an indication of any continuing, pervasive, additional effect of grouping, *after* Foundation Stage Profile (FSP) teacher assessment at age five, and *after* any special educational needs (SEN) diagnoses prior to surveying at age seven.

Alternatively, if the assumption that grouping placement commences immediately on school entry does not hold, inclusion of the FSP and SEN variables should account for additional school decisions and evaluations which may be entangled with relative age and with grouping practice and placements as initiated, at some point, between entry and age seven. Pupils may be placed in a lower in-class group because they have a SEN diagnosis, or vice versa; because they have a low FSP score, or vice versa; these decisions may take place sequentially, or concurrently.

If associations between ability grouping and teacher perceptions remain, even taking into account the potential confounding effects of these final factors (on top of the variables added at previous stages), stage four will therefore strengthen indications that grouping has a strong, independent effect.

Results

Table 5.2 indicates, for each subject domain, whether and the extent to which non-grouped autumn pupils are more likely to be judged as of 'above average' *ability and attainment* by their teachers, compared to summer-born, non-grouped pupils ('autumn'; see previous equation - this is coefficient 1). In each subject domain, at stages one, two, and three, there is a positive, significant relationship between being born in the autumn and being judged 'above average.' For example, according to stage one analysis, autumn-born pupils are 11.9 percentage points more likely to be judged 'above average' than summer-borns at speaking and listening. At stage four, however, upon addition of controls for previous in-school judgements, this difference is no longer significant. Having controlled for pupil, family, school and teacher characteristics, and previous in-school judgements, non-grouped autumn pupils and non-grouped summer pupils do not significantly differ in their chances of being judged 'above average' by their teacher.

Table 5.2 also indicates any association, for summer pupils, between being ability grouped and being judged 'above average' ('Ability grouped'; coefficient 4 from the equation). At each stage of analysis, in each subject domain, this relationship is negative – being grouped appears to lessen the chances of summer pupils of being judged 'above average' – but it is not statistically significant at the 5 percent level, in any subject, upon addition of controls beyond stage one.

However, the relationship indicated in Table 5.2 between being grouped and being judged 'above average' for autumn pupils ('Autumn x ability grouped'; coefficient 5) is positive and statistically significant at the 5 percent level or above at all stages of analysis, across all subject domains. For example, autumn-born children who are grouped have chances 11 percentage points higher than autumn-born pupils who are not grouped of being judged 'above average' at speaking and listening by their teacher at stage one, and this difference is barely altered at stage four, where it remains significant, at 10.9 percentage points higher.

Table 5.2: Key coefficients at each stage of analysis for relationships between month of birth / ability grouping and probability of being judged ‘above average’ by teacher

	Spec 1	Spec 2	Spec 3	Spec 4
Speaking and listening				
Autumn (ref: summer)	.119** (.041)	.159*** (.038)	.158*** (.038)	.031 (.040)
Winter (ref: summer)	.072 (.042)	.111** (.039)	.117** (.039)	.019 (.040)
Spring (ref: summer)	-.009 (.042)	-.003 (.039)	-.001 (.039)	-.058 (.039)
Ability grouped (ref: not grouped)	-.050 (.034)	-.019 (.032)	-.010 (.032)	-.005 (.032)
Autumn x ability grouped	.110** (.046)	.092* (.043)	.093* (.043)	.109* (.044)
N.	5325	5036	5036	4531
Reading				
Autumn (ref: summer)	.124** (.042)	.171*** (.039)	.158*** (.039)	.024 (.040)
Winter (ref: summer)	.051 (.043)	.100* (.039)	.098* (.040)	.013 (.040)
Spring (ref: summer)	.025 (.043)	.039 (.039)	.036 (.039)	-.008 (.040)
Ability grouped (ref: not grouped)	-.069* (.035)	-.039 (.032)	-.039 (.032)	-.026 (.032)
Autumn x ability grouped	.127** (.047)	.111* (.043)	.119** (.043)	.118** (.044)
N.	5322	5033	5033	4530
Writing				
Autumn (ref: summer)	.115** (.039)	.163*** (.037)	.158*** (.037)	.036 (.039)
Winter (ref: summer)	.032 (.040)	.073 (.038)	.077** (.039)	-.010 (.039)
Spring (ref: summer)	.008 (.040)	.024 (.038)	.027 (.038)	-.009 (.039)
Ability grouped (ref: not grouped)	-.071* (.033)	-.035 (.031)	-.027 (.031)	-.020 (.031)
Autumn x ability grouped	.107* (.044)	.082* (.041)	.081* (.041)	.094* (.094)
N.	5233	5032	5032	4530
Science				
Autumn (ref: summer)	.160*** (.040)	.191*** (.037)	.183*** (.037)	.058 (.039)
Winter (ref: summer)	.045 (.040)	.081* (.038)	.084* (.038)	-.004 (.039)
Spring (ref: summer)	.008 (.041)	.016 (.038)	.016 (.038)	-.030 (.039)
Ability grouped (ref: not grouped)	-.038 (.033)	-.018 (.031)	-.014 (.031)	-.016 (.031)
Autumn x ability grouped	.076 (.045)	.078 (.042)	.083* (.042)	.099* (.043)
N.	5319	5029	5029	4526

*** = p < .001; ** = p < .01; * = p < .05. Standard errors in brackets.

Each coefficient indicates percentage change in predicted probability of being judged ‘above average.’

Controlled at stage two for pupil and family characteristics; stage three adds school and teacher factors; stage four adds pupil FSP score / presence of SEN diagnosis – see Table 1.

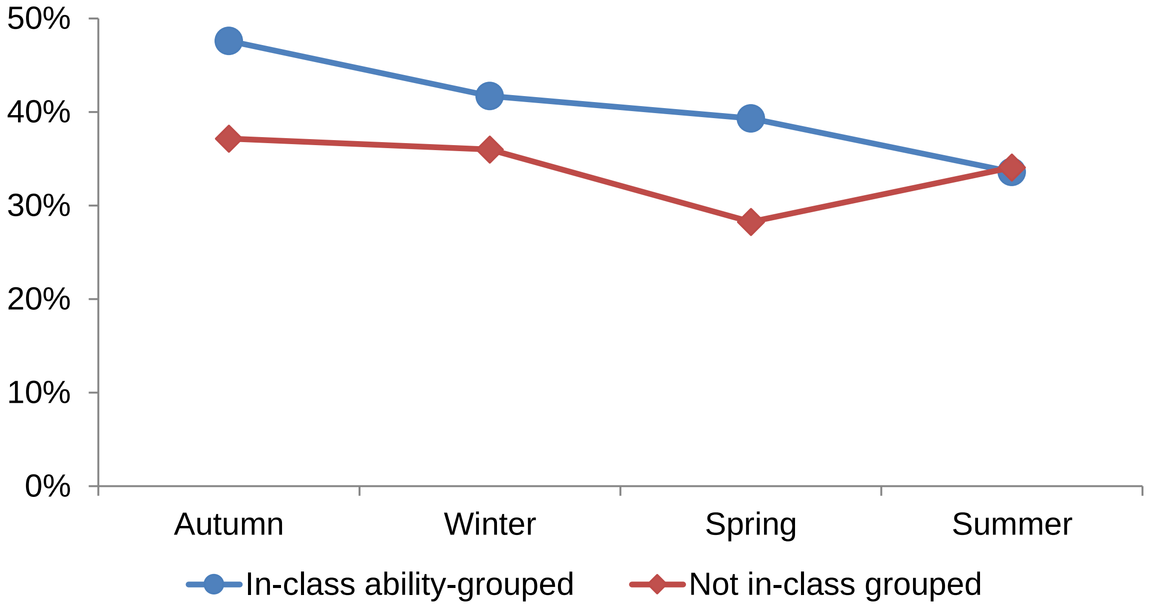
There are therefore three initial findings. Firstly, ungrouped autumn-born pupils are more likely than ungrouped summer-born pupils to be judged as of ‘above average’ ability and attainment by their teachers. This tendency holds upon addition of controls for pupil, family, school and teacher characteristics – but is negated upon addition of controls for previous in-school judgements.

Secondly, the difference made to summer pupils by being grouped appears minimal, though negative. Grouping appears slightly to lower teacher judgements of summer pupils – but these apparent effects are largely non-significant.

Thirdly, however, and in contrast, the practice of in-class grouping is indicated as strongly, positively related to teacher judgements of autumn pupils, even upon addition of all controls, including previous in-school evaluations and decisions.

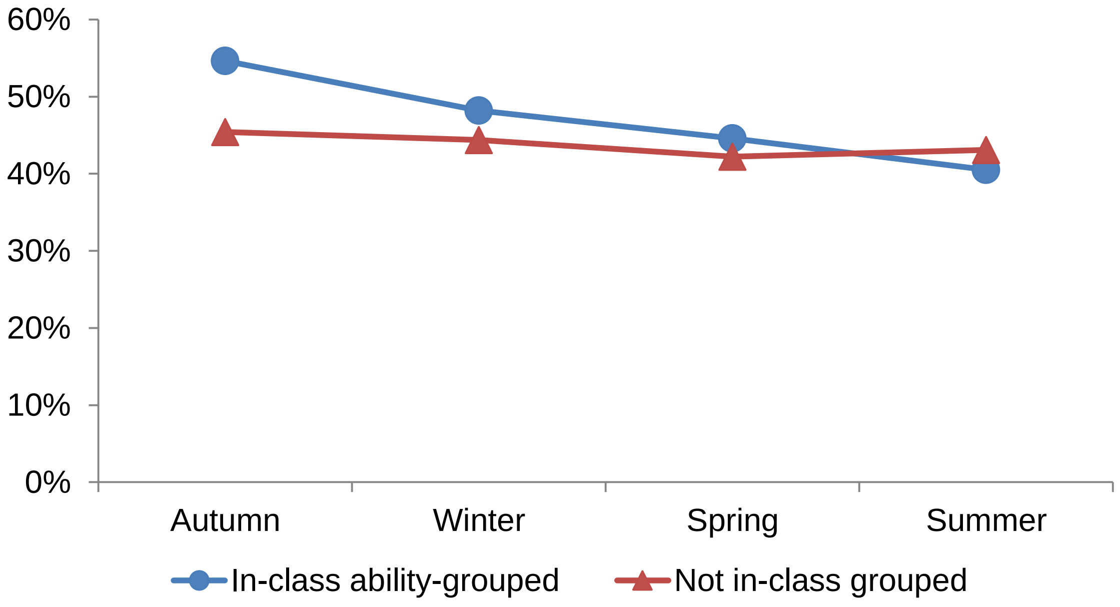
Crucially, these associations result in a much wider autumn-summer gap in teacher perceptions among pupils in schools which in-class group than among pupils in schools that do not in-class group. Figures 3 to 6 illustrate this finding for judgements in each subject domain, at stage four of analysis, with all controls.

Figure 5.3: Mean average percentage predicted probabilities for pupils in each season of birth / in-class grouped or not category, produced by stage four regression – probability of being judged 'above average' in <u>speaking and listening</u> *ability and attainment* by teacher



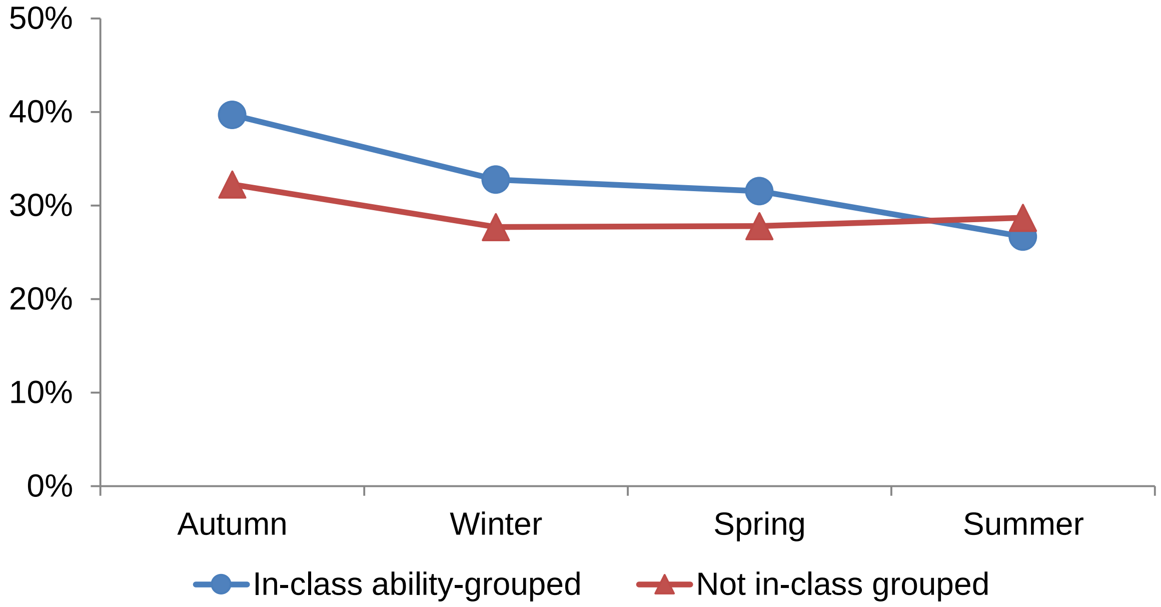
n = 4,531; controlled for pupil and family characteristics *and* school and teacher factors *and* pupil FSP score / presence of SEN diagnosis

Figure 5.4: Mean average percentage predicted probabilities for pupils in each season of birth / in-class grouped or not category, produced by stage four regression – probability of being judged 'above average' in <u>reading</u> *ability and attainment* by teacher



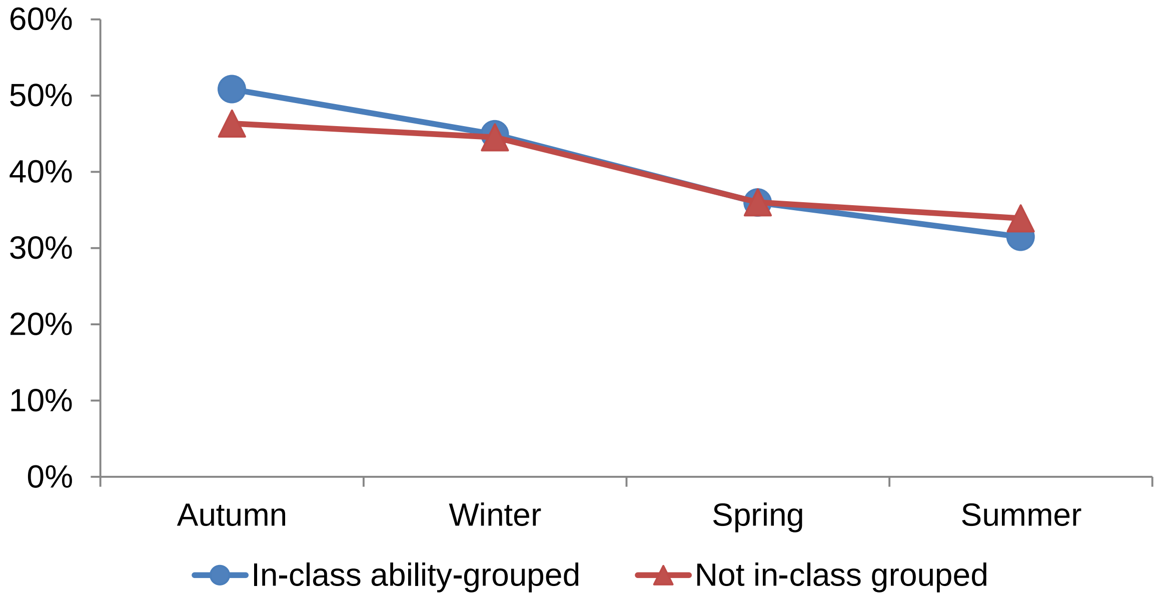
n = 4,530; controlled for pupil and family characteristics *and* school and teacher factors *and* pupil FSP score / presence of SEN diagnosis

Figure 5.5: Mean average percentage predicted probabilities for pupils in each season of birth / in-class grouped or not category, produced by stage four regression – probability of being judged 'above average' in <u>writing</u> *ability and attainment* by teacher



n = 4,530; controlled for pupil and family characteristics *and* school and teacher factors *and* pupil FSP score / presence of SEN diagnosis

Figure 5.6: Mean average percentage predicted probabilities for pupils in each season of birth / in-class grouped or not category, produced by stage four regression – probability of being judged 'above average' in <u>science</u> *ability and attainment* by teacher



n = 4,526; controlled for pupil and family characteristics *and* school and teacher factors *and* pupil FSP score / presence of SEN diagnosis

Figure 5.3 shows an autumn-summer difference in mean percentage predicted probability of being judged ‘above average’ in speaking and listening of 14 percentage points among pupils in schools which group (p < .001). Among pupils in schools which do not group, this difference is much smaller (3.1 percentage points), and non-significant (p = .435). For judgements of reading (Figure 4), the difference is 14.2 percentage points for grouped pupils (p < .001) and 2.3 for non-grouped (p = .553); for writing (Figure 5) it is 13 percentage points for grouped pupils (p < .001) and 3.5 percentage points for non-grouped (p = .355); and for science (Figure 6) it is 15.6 percentage points for grouped pupils (p < .001) and 5.7 for non-grouped (p = .138).

Sensitivity checks

Several alternative analyses were carried out in order to check whether methodological choices may have influenced the direction of results. Firstly, as mentioned, repeating analyses using a binary logistic regression rather than a linear regression produces equivalent findings. For example, in the logistic model, at stage four of analysis investigating teacher judgements of pupils’ speaking and listening *ability and attainment*, autumn-born pupils have odds 13 percent higher than summer-born pupils of being judged ‘above average,’ but as with the linear model, this is not significant (p = .62); grouped summer-born pupils have 4 percent lower odds than non-grouped summer-borns of being judged ‘above average,’ but, again like the linear model, this difference is non-significant (p = .84); while, true to the linear model, the difference between ability grouped and non-ability grouped autumn-born pupils is large and significant: grouped autumn-borns have odds 88 percent higher than non-grouped of being judged ‘above average’ (p = .016).

Secondly, using month rather than season of birth in modelling results in larger standard errors for some estimates due to reduced sample sizes, but does not influence the direction or significance of key results. Indeed, given the linear incremental pattern of associations with birth month, coefficients for the September-August difference are larger than those for the autumn-summer difference. For example, at stage four of analysis using teachers' judgements of speaking and listening, being in-class ability grouped results in a predicted probability of being judged 'above average' 16 percentage points higher for September-born pupils who are grouped compared to those who are not grouped (p = .038). The September-August difference among grouped pupils according to this specification is 22 percentage points (p < .001), while the difference among non-grouped pupils is smaller and non-significant at 6 percentage points (p = .405).

This chapter focusses upon tendencies in teacher judgements of pupils as being 'above average,' and highlights differences caused by a pattern where ability grouping seems disproportionately to favour relatively older pupils in teachers' positive judgements. In order to check the robustness of this finding, analyses were repeated using teacher assessments of whether pupils are 'below average,'[13] rather than 'above average,' and are congruent (see Annex 5D) – there is little association between ability grouping and teachers' judgements that children are 'below average.'

Lastly, as in previous chapters, results were checked for sensitivity to different weighting and clustering specifications (Stata version 12 was used for these checks). Table 5.3 shows that, at specification four, key findings are little changed by use / non-use of design and wave four weights, nor by clustering by school, nor when restricting the sample to a single teacher. As in the original results, no significant association with ability grouping is found for summer-born children, and across the four subjects, the three alternative specifications continue to indicate a positive relationship to being grouped for autumn-borns (bar one exception to this across the twelve analyses).

[13] Comprising a binary outcome variable where *below average* and *well below average* are combined to form 1 = 'below average,' and 0 = 'average or above' (average, above average, well above average).

Table 5.3: Key coefficients at stage four of analysis for relationships between month of birth / ability grouping and probability of being judged 'above average' by teacher: sensitivity checks

Speaking and listening	Original results	Check 1: weighted for survey design and for attrition to the main wave four survey	Check 2: Unweighted; clustered by school	Check 3: Unweighted; one pupil per teacher
Speaking and listening				
Autumn (ref: summer)	.031 (.040)	.056 (.043)	.031 (.040)	-.020 (.050)
Winter (ref: summer)	.019 (.040)	.031 (.046)	.019 (.042)	-.022 (.051)
Spring (ref: summer)	-.058 (.039)	-.038 (.043)	.058 (.040)	-.127** (.050)
Ability grouped (ref: not grouped)	-.005 (.032)	.009 (.035)	.005 (.032)	-.056 (.057)
Autumn x ability grouped	.109* (.044)	.088* (.047)	.109** (.043)	.161** (.056)
N.	4531	4531	4531	2685
Reading				
Autumn (ref: summer)	.024 (.040)	-.009 (.052)	.024 (.040)	.043 (.050)
Winter (ref: summer)	.013 (.040)	-.034 (.052)	.013 (.043)	.022 (.051)
Spring (ref: summer)	-.008 (.040)	-.030 (.046)	-.008 (.041)	.005 (.049)
Ability grouped (ref: not grouped)	-.026 (.032)	-.050 (.041)	-.026 (.034)	-.032 (.041)
Autumn x ability grouped	.118** (.044)	.152** (.055)	.118** (.044)	.099 (.055)
N.	4530	4530	4530	2688
Writing				
Autumn (ref: summer)	.036 (.039)	.031 (.039)	.036 (.039)	.013 (.049)
Winter (ref: summer)	-.010 (.039)	-.028 (.038)	-.010 (.039)	.010 (.049)
Spring (ref: summer)	-.009 (.039)	-.013 (.039)	-.009 (.039)	.032 (.048)
Ability grouped (ref: not grouped)	-.020 (.031)	-.035 (.032)	-.020 (.031)	-.047 (.040)
Autumn x ability grouped	.094* (.094)	.108** (.044)	.094** (.043)	.126** (.053)
N.	4530	4530	4530	2685
Science				
Autumn (ref: summer)	.058 (.039)	.057 (.043)	.058 (.040)	.042 (.049)

Winter (ref: summer)	-.004 (.039)	-.031 (.040)	-.004 (.038)	-.004 (.049)
Spring (ref: summer)	-.030 (.039)	-.024 (.039)	-.030 (.038)	-.054 (.049)
Ability grouped (ref: not grouped)	-.016 (.031)	-.024 (.030)	-.016 (.031)	-.041 (.041)
Autumn x ability grouped	.099* (.043)	.104** (.047)	.099** (.043)	.126** (.054)
N.	4526	4526	4526	2683

*** = p < .001; ** = p < .01; * = p < .05. Standard errors in brackets.
Each coefficient indicates percentage change in predicted probability of being judged ‘above average.’
Controlled for pupil and family characteristics; school and teacher factors; pupil FSP score / presence of SEN diagnosis.

Discussion

Analyses set out to investigate whether there is evidence for the possibility that in-class ability grouping early in primary school may contribute to the creation of systematic birth month differentials in pupil attainment. Findings provide support for the hypothesis proposed. Among children who *are* in-class ability grouped, autumn-summer variation in teacher perceptions of *ability and attainment* is greater than among pupils who are *not* grouped. The already disproportionate tendency of autumn-borns to be judged ‘above average’ is amplified among grouped children. This finding holds upon addition of a range of potentially confounding family, pupil, school and teacher factors.

Results here are consistent both with previous research which indicates that teacher perceptions of pupils are gradated according to birth month (Crawford *et al,* 2011; 2013a) and with studies which suggest that ability grouping may create or embed difference by providing an advantage to pupils placed at higher levels (Blatchford *et al*, 2008; Hallam & Parsons, 2012; Kutnick *et al*, 2005). Findings in this chapter suggest that because they are often placed in the top group when in-class ability grouping takes place, autumn-born pupils may be advantaged through a heightening of teachers’ judgements of their *ability and attainment* which is related to this group placement.

Research indicates that teacher opinions and expectations can influence the academic trajectory of their pupils. Therefore, analyses here indicate that in-class ability grouping may provide a significant ‘boost’ to the development of autumn-born children which raises their progress above their relatively younger peers. Findings begin to support a model where grouping is instrumental in the relative age effect – and where cessation of in-class ability grouping may go some way towards alleviating the effect.

Alternative explanations, and implications of these

The data available in the MCS do not contain information on the exact decision-making and administrative processes that led to each of the study children being grouped, or not grouped. Therefore it is not possible to know whether the presence or absence of in-class grouping is due to school policy, choice on the part of individual teachers, or some combination of these factors. The exact chain of events and pattern of effects is, therefore, uncertain. The main hypothesis proposed in this Chapter is that in-class grouping affects teacher perceptions – but it is possible that, in some cases (and as has, for example, been suggested by Kuklinski & Weinstein, 2000), teachers with a propensity to notions of fixed ability, and a tendency to more extreme discrimination and differentiation between students (including that, potentially, according to birth month), enact these tendencies in a decision to ability group their pupils.

However, this possibility, if it is, in fact, the case for some of the MCS respondent teachers, does not negate the suggestion that ending in-class grouping during early primary school may assuage the month of birth effect. If a policy of *no* early in-class groupings were prescribed, it would disallow a practice which legitimises and reifies assumptions of intrinsic differences in ability and potential (which, as discussed in this and the previous chapter, appear invalidly biased by pupil characteristics, including month of birth); a practice which embeds these assumptions, providing a deterministic conduit through which they may play out. Disallowing ability grouping may therefore, in itself, evoke some reassessment of teachers’ own practices and beliefs –

or, at least, provide some restraint to the application of premature and divisive categorisations and delineations between pupils.

Moreover, research suggests a number of additional channels alongside that of teacher expectations through which ability grouping might affect pupil attainment - including pupil self-perceptions, and differentiated educational and assessment opportunities (Kutnick *et al*, 2005; Blatchford *et al*, 2008). Given the disproportionate distribution of pupils born in different months across the in-class hierarchy, whether the presence of grouping affects teacher perceptions, or vice versa, or both, an *absence* of in-class grouping may, theoretically, prevent its effects from manifesting by blocking a variety of subsequent pathways.

However, as analyses in this chapter are essentially descriptive manipulations of observational data, and notwithstanding the theoretical bases for the explanation favoured, threats and alternatives to this explanation remain. It is possible, for example, that in-class grouping is used more often by schools who are less efficient in teaching relatively younger children – and that the depression in teacher perceptions of these children in fact reflects lowered pupil performance that results from these other in-school factors. Even if this is the case, however, it seems likely that the use of grouping may only compound age-based differentials through the channels discussed above, rather than alleviating them. Longitudinal explorations of the trajectories of the MCS pupils will allow further investigation of these hypotheses.

Policy implications

Recent UK governments have consistently enabled ability grouping (see Department for Education and Skills, 2005; Conservative Party, 2007; Department for Children, Schools, and Families, 2008b; Department for Education, 2010a) - while, at the same time, stating a desire for an educational system which engenders parity of access and opportunity:

> Our schools should be engines of social mobility, helping children to overcome the accidents of birth and background to achieve much more than they may ever have imagined. But, at the moment, our schools system does not close gaps, it widens them (*Department for Education, 2010a, p 6*).

Findings in this chapter, from a large, recent, national sample of seven-year-olds, suggest that the policy and practice of in-class ability grouping pupils early in primary school may, in fact, be detrimental to mobility. If systematic month of birth variation in attainment is to be 'overcome' through changes to policy and practice – and few 'accident[s] of birth' are more arbitrarily foisted upon an individual than their birth *date* – then the evidence here indicates that reversal of the policy of in-class ability grouping in early primary school may contribute to 'closing the gap' between relatively younger and relatively older pupils.

Chapter 6

Discussion, implications, and conclusions

Summary

As stated in the introduction, the papers in this thesis are united by their intention to add to a developing understanding of the factors that contribute to and construct inequalities during early education. They have provided evidence across three interrelated areas of investigation: that concerned with the influence and impact of the practice of streaming on young children; that exploring the existence and effects of stereotyping and bias in assessments of primary school pupils; and that unpicking the factors and processes that contribute to the formation of the 'month of birth effect.'

Chapter 3 presented evidence which contributes to a growing UK and international research-base indicating that streaming widens attainment gaps and disadvantages children placed at lower levels. It indicated that pupils who score at the same level on relevant cognitive tests and who are equivalent according to a wide range of other characteristics are assessed comparatively more or less favourably depending on the stream in which they are situated. Streaming, and stream placement, therefore appear to play a part in shaping, delineating and differentiating children's educational progression.

Chapter 4 went on to describe findings of bias in teachers' assessments of seven-year-olds, and to suggest that these consistent patterns support the hypothesis that stereotyping is instrumental in constructing children's measured attainment during early primary school. It found disparities between children's cognitive test performance and teachers' perceptions of those same children's 'ability and attainment' which correspond to all the major pupil-level characteristics according to which achievement has been measured over the past decade. It therefore calls for a consideration of the part played by these processes in the formation of formally recorded 'attainment' and in classroom learning.

Finally, Chapter 5 provided indications that early in-class ability grouping may play a part in producing difference among children according to their birth month. Particularly given the wider research base which implicates ability grouping as influencing pupils' progression and achievement, this suggests that cessation of early in-class grouping may provide some leverage for a flattening of relative age discrepancies.

Policy context, recent history, and implications of thesis findings

What, then, is the relevance and what are the implications of these findings for educational policy-making?

Since the reform to comprehensive schooling of the mid-20th century, successive UK governments have explicitly declared a commitment to and focus upon parity in education (Central Advisory Council for Education, 1967; Department for Education and Employment, 1997; Department for Education and Skills, 2005). This commitment has continued to feature prominently within the political rhetoric and public promises of the past two decades.

On coming to power in 1997, the Labour government's inaugural Education White Paper proposed that:

> Excellence at the top is not matched by high standards for the majority of children...achievement by the average student is just not good enough…[there is] an unacceptable and growing gap in performance (Department for Education and Employment, 1997, p.10; p.34).

2001's subsequent Paper held similarly that: '…there is still a huge gap, based too often on a child's social or economic background, on their ethnic group…' (Department for Education and Skills, 2001, p.3-4), and, in 2005, the following Paper included a regret that, 'the attainment gap for pupils has not yet narrowed' (2005, p.19). Though Labour's last Education White Paper, in 2009, argued that, 'the gaps have narrowed' (Department for Children, Schools and Families, 2009a, p.14), it went on to stress:

> …there continue to be significant differences between the achievements of different groups of children and young people – most significantly between the disadvantaged and others. The gap is wider in this country than in many others (ibid, p.14).

After the change of government in 2010, the Conservative – Liberal Democrat Coalition began their tenure with a position largely in line with Labour's concluding assessment: 'our schools system does not close gaps, it widens them' (Department for Education, 2010a, p.6).

> Our highest performing students do well but the wide attainment gap between them and our lowest achievers highlights the inequity in our system (ibid, p.47).

Across governments, therefore, this fundamental issue has prevailed. There has been a persistent focus on the presence of 'gaps,' and on 'equity' for all – and the political consensus has been that the education system can, and should, be instrumental in the creation of parity of attainment. Recent governments have, accordingly, formulated a multitude of policy changes.

At a high level, the Labour administration's initiatives and priorities have included: increased investment in early years provision, and in 'early diagnosis and intervention for pupils who face particular challenges' (ibid, 2001, p.9); increased monitoring of pupil 'attainment' against prescribed 'standards;' explicit and differentiated 'target-setting' for various pupil groups; increased inspection and monitoring by Ofsted;[14] increased central government involvement and prescription of practice at the school-level; additional resources and targeted interventions for particular school-types and pupil groups; and advocacy and recommendation of differentiated teaching and of 'ability' grouping. (Department for Education and Employment, 1997; Department for Education and Skills, 2001; Department for Education and Skills, 2005; Department for Children, Schools and Families, 2009a). The Conservative-Liberal Democrat Coalition sustained the emphasis on characteristics-based monitoring and accountability, while also

[14] Office for Standards in Education, Children's Services and Skills (https://www.gov.uk/government/organisations/ofsted/about)

introducing an explicit ‘pupil premium’ of funding for children from low-income families, and reforming the primary curriculum to include an emphasis on early phonics teaching and a corresponding additional stage of judgement and assessment of pupils through a ‘reading check.’ (Department for Education, 2010a).

Have these many initiatives and interventions worked, however, or may some of them have even compounded inequalities? Reviewing the evidence on fluctuations in disparities, Whitty & Anders (2014) contend that, ‘progress has been so limited to date’ (p 3), and, with regard to the 1997-2010 Labour administration, that:

> …although by most measures there was a small reduction in the attainment gap under the New Labour government…[this is] a disappointing achievement when compared with the aspirations of successive Prime Ministers and Secretaries of State for Education. (p.18)

Similarly, appraising progress under the subsequent 2010-2015 Coalition, Lupton & Thompson (2015) conclude that development has been minimal: ‘The next government will inherit a school system in flux and key issues of equity and achievement still unresolved’ (p.5).

The evidence suggests, therefore, that despite the efforts of consecutive policy-makers, inequities are still far from being alleviated. Instead, it seems that, instead of disparities meaningfully being lessened by recent administrations, new ‘gaps’ are only waiting to be discovered or constructed – as indicated by the recent beginnings of documentation of stark differences by birth month. Far from nurturing parity and equity, the past two governments appear to have presided over an education system that has, on the whole, sustained – rather than mitigated – difference. This begs re-examination of the system and its parts, and formation of a deeper, revised understanding and consideration of the processes that may generate inequalities between pupils – and it is to this deeper understanding that the papers in this thesis contribute.

Often missing from discussion at the policy level is explicit recognition and detailed analysis of the logical antithesis of the contention that schools can be ‘engines of social mobility’ (Department for Education, 2010a, p.6): the possibility that, if the education system feasibly can engineer parity and opportunity, it can also, as evidenced in this thesis, create inequalities, produce barriers, and impede progress. When this is acknowledged, it is to date (as in the quotes above) often as a throw-away line (‘the inequity in our system’ [ibid, p.47]), with little scrutiny of the specific factors that may breed difference. This omission is important not least because any interventions and policies genuinely intended to bring positive advancement towards equity may fail if they are working against enduring systematic or human factors that have the opposite effect, or that interact with policies to warp their anticipated outcomes.

Acknowledging and addressing the evidence presented here that streaming, stereotyping, and in-class ability grouping may contribute to the creation of inequalities could therefore play a part in tackling disparities within the education system.

Streaming

Recent policies on streaming have tended to support the practice, either explicitly or implicitly. The most recent Labour government was heavily in favour of in-school ability grouping, despite a stated commitment to comprehensive schooling. The party’s overriding agenda was a:

> ...need to hold on to the values and principles that underpin our commitment to comprehensive education – that every child is special and that all children should have the opportunity and support to develop their skills and ability to achieve their full potential – but apply them in a way that is appropriate to a 21st century world (ibid, 2001, p.6)

Within (or, arguably, in contrast to) this high-level framework, Labour held a more explicit, innatist assumption: that children are of different ‘types,’ and can be categorised, hierarchically, into the ‘gifted and talented,’ the ‘struggling,’ and the ‘just average’ (ibid, 2005, p.20). While simultaneously lauding a non-selective system, their policy documents bemoaned a (too) comprehensive schooling, offering ‘all-ability classes, which made setting by subject ability too rare’ (ibid, p.1).

Labour therefore called for, ‘more grouping and setting by subject ability’ (ibid, p. 10), and their term ended with a re-emergence not only of ability grouping for specific subjects, but with the national normalisation within primary schools of overall, non-subject-based ‘ability’ groupings. As noted, in 2008, the evidence indicated that the majority of seven-year-old children were grouped in-class for all teaching, and that a notable minority (nearly a fifth) were placed in overriding, streamed bands within their year group (Campbell, 2013a; Hallam & Parsons, 2013).

This increasing prevalence of early structural ability groupings appears somewhat incongruent with, or at least potentially to problematize, Labour’s assessment that, ‘Comprehensive schools overcame the ill effects of rigid selection and have done a great deal to improve opportunity’ (2001, ibid). If ‘rigid selection’ at the between-school level can cause ‘ill effects’ and hinder opportunity, it seems feasible that, as evidenced in this thesis, within-school selection might have similar consequences.

Political support for and encouragement of grouping continued unquestioned until the end of Labour’s term in 2010 (the consultation paper for the 2009 *21st Century Schools* White Paper, for example, continued to endorse ‘carefully planned pupil groupings’ [Department for Children, Schools and Families, 2008b, para 3.5]). In contrast, the Coalition Government has been less open and, on the surface, circumspect regarding its strategies in this area.

Possibly this is because its tacit policies are in direct contradiction of the recommendations presented by the evidence-reviewers established under its governance, and to whom it directs funding. The Education Endowment

Foundation (EEF), set up to ‘extend the evidence-base on what works to raise the attainment of disadvantaged pupils in schools in England’ (Education Endowment Foundation, n.d.a) describe an appraisal of research congruent with analyses in this thesis, indicating that, ‘the average impact of setting or streaming on low attaining pupils is negative,’ and that ‘Flexible within-class grouping is preferable to tracking or streaming’ (Education Endowment Foundation n.d.b). Explicitly addressing the matter of ability grouping, in any form, would therefore highlight a policy area where the Coalition, should it endorse grouping, may be open to accusations of going deliberately against the evidence base for which it has paid.

Even on election, the 2010(a) Education White Paper made no reference at all to streaming, setting, or in-class grouping. However, prior to coming to power, the Conservatives were avowedly pro-stratification and pro-selection: their 2007 Green Paper, *Raising the Bar, Closing the Gap,* asserted a belief in ‘[delivering] more teaching by ability which stretches the strongest and nurtures the weakest’ [p.9]. Given disincentives to discussion, and an overall, ongoing Conservative commitment to the social Darwinism of free market competition (the 2015 manifesto talks of ‘doing all we can to help the next generation get on in life and succeed in the global race’ [Conservative Party, n.d.]), it seems unlikely, therefore, that recent silence should be read as any kind of reversal of ideology. Indeed, in response to a (2012) OECD review which argued that ‘student selection – and in particular early tracking (setting and streaming) – exacerbates differences in learning between students,’ the Department for Education issued a response essentially confirming support for these practices:

> It is for schools to decide how and when to group and set pupils by ability as they are best placed to know and meet the learning needs of their pupils. Research shows that when setting is done well it can be an effective way to personalise teaching and learning to the different needs of groups of pupils. (Guardian, 2012, online)

Streaming has therefore been supported through explicit or surreptitious policy, as well as by omission from debate, by all recent governments –

despite a growing evidence-base that it is both inappropriately implemented (with children being allocated to streams not simply on the basis of indications of their capabilities and potential, but according to their other characteristics) and that it is detrimental to equitable progress. The analysis in Chapter 2 of this thesis only strengthens this evidence, while contributing to an understanding of the processes associated with streaming by implicating the psychological impact of stream placement on teachers' perceptions, and consequentially their assessments, of pupils. Findings here make imperative the case for a thorough and transparent review of the use and effects of streaming, and for proper deliberation regarding whether the practice should be allowed to continue in early primary schooling.

It is currently possible only to guess at the motivations for the disconnect and contradiction between the existing, publicly available research (which advises against streaming) and the current policy response (which supports streaming). Chapter 2 of this thesis suggests that being placed in a higher stream can enhance teachers' judgements and assessments of pupils. Perhaps this nods towards a tacit underpinning to policy-making that favours the advantaged while publicly stating a commitment to raising the attainment of the disadvantaged. If this is the case, these underpinnings should be made explicit, so that they can be examined, and their evidence and ideology thoroughly scrutinised.

In line with this recommendation, then, it is worth noting particularly the finding in this thesis of advantage for top-stream pupils, which differ from the picture generally conveyed by reviews of the previous literature, which emphasise penalisation of bottom-stream pupils (e.g. Education Endowment Foundation, n.d.b). The implications of this result are not straightforward, and careful interpretation raises a number of initial questions.

Firstly, and fundamentally, taken alongside findings that lower-group placement is detrimental to assessments, what level of veracity should be attributed to attainment measured in this way? To what extent does this 'attainment' actually represent performance and capability on the part of the child? If recorded achievement is to some degree merely a construct and artefact of situational influences such as stream placement, any ‘advantage’ to top-stream pupils is not an unproblematic or an entirely desirable outcome.

Secondly, assuming a level of validity to these assessments, is the raising of attainment of the children placed in the top stream a fair price to pay for the depression of the trajectories of lower stream children? Particularly given biases and disproportionalities in placement which distribute children unevenly across streams according to their characteristics, and regardless of manifest competence, is this manipulation of trajectories appropriate?

Lastly, potentially null effects in terms of overall efficiency given these two opposed and corresponding impacts provide a final challenge to any argument for streaming based on any apparent advantage to those at the top of the hierarchy. If overall efficiency is not improved due to a symmetrical redistribution, what merit is there to this redistribution?

In order that these implications can accurately be disentangled, policy-makers must firstly acknowledge openly the growing use of streaming among young children. Secondly, they must consider fully the evidence on the implementation and effects of streaming. Thirdly, they must make clear the interaction between their interpretation of this evidence and their chosen ideological framework, when allowing, or legislating against, the use of the practice in early primary school.

Stereotyping

To some degree, and in various ways, the potential for bias and stereotyping to influence achievement has been acknowledged by recent governments. However, fleeting recognitions of possible processes have not yet been meaningfully translated into concrete policy-making – so, in that sense, this is an area still in its infancy.

Labour’s inaugural Education White Paper, for example, nodded towards systematic biases when discussing pupils from non-White ethnic groups – stating, for instance, that ‘Pupils from some groups are disproportionately excluded from school’ (para 49) and that ‘Racial harassment and stereotyping continue' (para 49). It recommended that guidance be provided on:

> …best practice…in tackling racial harassment and stereotyping, in promoting attendance and reducing exclusion of ethnic minority pupils, and in creating a harmonious environment in which learning can flourish…(Department for Education and Employment, 1997, para 50).

However, throughout this paper itself, and in contradiction to the above, inconsistent tone and content also suggest a wider context including stereotyped assumptions within which these recommendations are positioned. The paper states, for instance, that minority ethnic children bring, 'cultural richness and diversity, but some are particularly at risk of under-achievement' (para 49) – which seems to indicate that this (over-)generalised group of non-White children, with their 'diversity,' difference, and implied deficit, are, at least to some extent, the origin of their own failings.

Similarly, Labour's 2005 White Paper echoed insinuations of pupil-level deficiency, and (perhaps non-consciously) manifested an institutional bias, by suggesting that the source of disproportionate under-attainment of certain pupil groups had a basis largely within, rather than outside of, or in the system surrounding, these groups. It stated that, 'Whilst many black and minority ethnic (BME) young people achieve well, a significant number fail to realise their potential,' and it emphasised the importance of:

> …ensuring that schools have expert advice on how to support pupils facing particular challenges – including those from black and minority groups, disabled children…and children with Special Educational Needs (Department for Education and Skills, 2005).

Against a backdrop which at best assumed that causes of disparities in attainment 'are complex' (Department for Education and Employment, 1997), therefore, Labour instigated a series of policy interventions and programmes intended to raise the attainment of pupil groups deemed to be underachieving, including children designated as having SEN, and Black pupils of all backgrounds (Dockrell *et al*, 2007; Maylor *et al*, 2009). Alongside these targeted initiatives, and increased monitoring of attainment by pupil characteristic, the government commissioned a series of research evaluations and reports exploring the reasons for disproportionalities according, for example, to ethnicity (Strand, 2007; Strand *et al*, 2010).

However, subsequent policy-making essentially ignored the evidence generated by these studies, which suggested that individual or institutional perceptions, judgements, and related behaviours might play a part in creating difference between pupil groups. Concrete policies continued to assume that the origins of inequalities in attainment resided almost entirely at the pupil and family-level, and that socio-economic status (SES), often according to the proxy of FSM, was the key explanatory factor underpinning attainment gaps. Thus the Labour administration did not explore proposed mechanisms through which SES variation might play out in differentiated pupil attainment beyond the level of the pupil and their family, and the potential for associated explanation at the teacher, school, or system-level remained largely unacknowledged (Department for Education and Skills, 2005; Department for Children, Schools and Families, 2008b, Department for Children, Schools and Families, 2009a).

On election, the Conservative-Liberal Democrat focus continued to emphasise material family-level poverty as the fundamental driver of inequalities:

> For far too long we have tolerated the moral outrage of an accepted correlation between wealth and achievement at school…Children on free school meals do significantly worse than their peers at every stage of their education (Department for Education, 2010a).
>
> At the heart of our Coalition’s Programme for Government is a commitment to spend more money on the education of our poorest children (Department for Education, 2010c).

Any need to address differences in achievement according to other characteristics, let alone the complexities of the parts played by factors at the child, family, teacher or system level, was largely negated under an assumption of co-relationships between characteristics and a fundamental primacy of FSM-status not just as descriptor but as key driver and origin. The Equalities Impact Assessment (DfE, 2010c) which accompanied the 2010(a)

Education White Paper indicated, for example, that SEN and ethnicity are categories – but not sources – of between-pupil variation, and that consideration of these factors could therefore be set aside. Correspondingly, the Coalition government introduced the heavily publicised Pupil Premium, which channels funds to less wealthy pupils, and presented this as a solution which, by targeting poverty and reallocating resources, would alleviate a variety of documented inequities, including those according to ethnicity and SEN:

> As many deprived [children] also have Special Educational Needs or are members of underachieving ethnic groups…significant numbers of pupils from these groups will also benefit from the extra resources and tailored support the Pupil Premium will provide (Department for Education, 2010c, p.9).

That this overriding policy and its overt implementation may not in fact prove a panacea which closes all gaps and engenders equality has been little acknowledged. Though the potential effects of biased perceptions have been nodded towards during Conservative-Liberal Democrat policy espousal ('the soft bigotry of low expectations' has occasionally been denounced), 'communities' have been blamed openly for a 'deeply embedded culture of low aspiration' (Department for Education, 2010a), and directions for tackling the processes behind 'low expectations' have not proceeded beyond a general edict to develop 'a strong sense of aspiration for all children, whatever their background' (ibid, 2010a). This minimises the complexities of the system, and largely ignores the potential, evidenced in this thesis, for bias and stereotyping to continue to be instrumental in sustaining disparity.

Therefore, it seems that, though there have been momentary acknowledgements of the potential for perceptions and judgements at the system, school or teacher level to influence assessments of pupils, and for bias and stereotyping to affect children's educational experiences and outcomes, these processes have not significantly or meaningfully been addressed, to date, by recent governments.

Chapter 4 challenges this situation, suggesting that the universal human process of stereotyping should explicitly be tackled within educational policy-making and practice. Findings call into question assumptions regarding the processes shaping and influencing primary ‘attainment,’ and, like the evidence in Chapter 3, challenge the veracity of the attainment measures themselves. Like Chapter 3, analysis in Chapter 4 illustrates the contingency of teacher judgements upon factors other than that they seek to measure, and emphasises their manipulability and fallibility as means by which to gauge pupil performance.

This does not lead necessarily to any suggestion that alternative measures of pupil performance should be employed – as discussed in Chapter 3, formal standardised tests, for example, come with their own drawbacks and limitations. However, it problematizes the necessity and desirability of measuring itself. It has long been recognised that the act of observing influences the observed (Landsberger, 1958), and that prescribed and rigid testing, monitoring and reporting force patterns to be imposed that may not previously have been present (Campbell, 1976).[15]

Possibly, then, reduced and less formalised assessment and recording of children’s attainment (particularly according to characteristic) might remove a link from the vicious circle of measurement, denotation and publication that seems potentially to reify, embed and reinforce stereotypes of different pupil groups. There is a need to research and weigh up the utility of data collections against the unintended outcomes of collection, and to challenge the current uses and descriptions of pupil data with attention to and appraisal of all potential effects. The idea that pupils should be ‘measured from the earliest possible point in school’ (Department for Education, 2014d) is problematic, and should comprehensively be analysed with reference to the prospective consequences of this procedure.

This is not, however, to negate the importance of the fundamental human psychological process of unconscious cognitive bias demonstrated by Chapter 4. Like Chapter 3, which illustrated disproportionalities in procedures

[15] “…achievement tests may well be valuable indicators of general school achievement under conditions of normal teaching aimed at general competence. But when test scores become the goal of the teaching process, they both lose their value as indicators of educational status and distort the educational process in undesirable ways” (p. 51-52).

regarding children’s stream placement, Chapter 4 described disparities in judgement according to pupil characteristic which may influence everyday classroom interactions and educational opportunities – regardless of whether these biases are reinforced by formal recording and reporting. So stereotyping within education remains to be addressed, and given the lack of policy response to date, the first step towards this should be a thorough literature review and formulation of promising approaches to be trialled and tested.

In-class ability grouping and birth month effects

At the time of finalising this thesis, interest at the policy-level in factors which may contribute to the month of birth effect, and means by which it may be assuaged, remains high. The Secretary of State for Education has written to the Education Select Committee with a commitment to continued consideration and investigation of the education of summer-born children (Gibb, 2015), though this promise makes no mention of the potential contribution of early ability grouping to the formation of birth month differences.

The evidence presented in Chapter 5 as well as Chapter 3 of this thesis indicates the apparent instrumentality of early ability grouping in shaping pupils’ trajectories and delineating progress. Therefore, if the month of birth effect is genuinely to be addressed, the use of in-class ability grouping among young children (as well as other types of ability grouping) should transparently be monitored and recognised, and the potential for cessation of grouping to contribute not just to a reduction of relative age disparities but to a diminishing of inequalities according to other characteristics considered. Otherwise, given the wide disparities in age, maturity and readiness of children on entry to primary education at four or five years old, early allocation to rigid groupings may continue to exacerbate disparities, as suggested here.

Future research

Streaming

The evidence produced in Chapter 3 may be built upon and expanded in several ways as the longitudinal data of the MCS continue to emerge. Firstly, the apparent effects of early stream placement on later schooling should be examined, in order to map its influence through Key Stage Two and beyond. Secondly, non-academic correlates of stream placement should be investigated – are children's reported experiences of bullying or enjoyment of school related to their placement level, for example? Thirdly, at least one of the other channels proposed to create the association between stream placement and differentiated outcomes may be explored using this data – children's academic self-efficacy, attitudes and motivations.

As noted within the empirical chapter, and given the sample limitations described in Chapter 2, national data on the use of streaming and on individual children's stream placement would help to develop a more definitive sense of the current and unfolding use and consequences of the practice. If this data were available within the National Pupil Database, this would not only make clear the extent of use, it would also enable more detailed analysis of between- and within-school differences in patterns and relationships, and examination of interactions with factors such as school constitution, which may motivate streaming.

Stereotyping

A priority for future research on stereotyping in schools should be exploration of ways to tackle the process. A full literature and practice review of previous interventions and approaches should inform this, along with further data-driven exploration of whether there are schools where biases are lesser or non-existent, and of what is different about these schools. Once this evidence-base has been established, it can be used to generate suggested interventions which might be trialled to tackle and alleviate stereotyping, situated within a psychological theory of behaviour change. An iterative process of testing and monitoring should follow, in order to begin to discover what approaches may be effective in assuaging the stereotyping process and its apparent effects.

In-class ability grouping and birth month effects

Lastly, in order to investigate further the potential impacts of early in-class ability grouping on pupil attainment, future analysis should examine whether presence of the practice at age seven is associated with greater birth month disparity at age 10/11, in both teacher-assessed and externally tested Key Stage Two examinations. In addition, the non-academic correlates of in-class grouping should be explored: is grouping, or group position, related to differences by birth month in wellbeing measures, for example?

Research on ability grouping suggests a number of means by which its effects may manifest, including children’s own academic self-efficacy and attitudes to school. Whether this relationship seems to account for relative age inequalities among the Millennium Cohort pupils should also be considered in future work.

As already noted above with regard to the practice of streaming, national data collection and availability on the presence of ability groupings within schools and on pupils’ relative placement levels would enable a strengthening of the evidence-base on the results of these practices, illuminating further their association with birth month disparities among children.

Conclusion

By presenting three interrelated empirical chapters of original research on psychological and structural processes that may contribute to the construction of inequalities among primary school children, this thesis has raised suggestions for change and intervention that may help to diminish inequities, and begin to bring about parity in education. If streaming, in-class ability grouping, and stereotyping are addressed, this may contribute to a school system ‘in which opportunity is equal for children and young people, no matter what their background or family circumstances.’[16]

[16] Department for Education, n.d.

References

Allen, R., Belfield, C., Greaves, E., Sharp, C., & Walker, M. (2014). 'The Costs and Benefits of Different Initial Teacher Training Routes.' [Online]. London: The Institute for Fiscal Studies. Available at: http://www.ifs.org.uk/uploads/publications/comms/r100.pdf [Last accessed 5th May 2015.]

Ansalone, G. (2003). 'Poverty, tracking, and the social construction of failure: International perspectives on tracking'. *Journal of Children and Poverty*, 9 (1), 3-20.

Bedard, K. & Dhuey, E. (2006). 'The Persistence of Early Childhood Maturity: International Evidence of Long-Run Age Effects'. *The Quarterly Journal of Economics,* 121(4), 1437-1472.

Bew, P. (2011a). 'Review of Key Stage 2 testing, assessment and accountability: Progress Report'. (Government document). [Online.] Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/180401/DFE-00035-2011.pdf [Last accessed 5th May 2015.]

Bew, P. (2011b). 'Review of Key Stage 2 testing, assessment and accountability: Final Report'. (Government document). [Online.] Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/176180/Review-KS2-Testing_final-report.pdf [Last accessed 5th May 2015.]

Blatchford, P., Hallam, S., Ireson, J. Kutnick, P., & Creech, A. (2008). 'Classes, Groups and Transitions: structures for teaching and learning – Primary Review Research Survey, interim report.' [Online]. Available at: http://core.ac.uk/download/pdf/309501.pdf [Last accessed 5th May 2015.]

Blatchford, P., Hallam, S., Ireson, J. Kutnick, P., & Creech, A. (2010). 'Classes, Groups and Transitions: structures for teaching and learning.' In Alexander, R. (Ed.), *The Cambridge primary Review Research Surveys.* England: The University of Cambridge.

Boaler, J. (1997). 'Setting, social class and survival of the quickest'. *British Educational Research Journal*, 23, 575-595.

Boaler, J. Wiliam, D. & Brown, M. (2000). 'Students' experience of ability grouping - disaffection, polarisation and the construction of failure'. *British Educational Research Journal*, 26(5), 631-48.

Boardman, M. (2006). 'The impact of age and gender on Prep children's academic achievements'. *Australian Journal of Early Childhood,* 31(4), 1-6.

Bradbury, A. (2011a). 'Equity, ethnicity and the hidden dangers of 'contextual' measures of school performance'. *Race Ethnicity and Education*, 14(3), 277-291.

Bradbury, A. (2011b). 'Rethinking assessment and inequality: The production of disparities in attainment in early years education'. *Journal of Education Policy*, 26(5), 655-676.

Brookhart, S. M. (2013). 'The use of teacher judgement for summative assessment in the USA'. *Assessment in Education: Principles, Policy & Practice*, 20(1), 69-90.

Brophy, J.E. & Good, T. L. (1970). 'Teachers' communication of differential expectations for children's classroom performance: Some behavioral data'. *Journal of Educational Psychology*, 61(5), 365-374.

Brown, L. L. & Sherbenou, R. J. (1981). 'A Comparison of Teacher Perceptions of Student Reading Ability, Reading Performance, and Classroom Behavior'. *The Reading Teacher,* 34(5), 557-560.

Burgess, S. & Greaves, E. (2009). ‘Test Scores, Subjective Assessment and Stereotyping of Ethnic Minorities.’ [Online]. Available at: http://www.bris.ac.uk/cmpo/publications/papers/2009/wp221.pdf [Last accessed 5th May 2015.]

Campaign for Flexible School Admissions for Summer Born Children (n.d.). [Online]. Available at: http://summerbornchildren.org/home-2/ [Last accessed 5th May 2015.]

Campbell, D. T. (1976). ‘Assessing the Impact of Planned Social Change.’ [Online]. Available at: https://www.globalhivmeinfo.org/CapacityBuilding/Occasional%20Papers/08%20Assessing%20the%20Impact%20of%20Planned%20Social%20Change.pdf [Last accessed 22nd July 2015.]

Campbell, T. (2013a). ‘In-school ability grouping and the month of birth effect: Preliminary evidence from the Millennium Cohort Study.’ [Online]. Available at: http://www.cls.ioe.ac.uk/shared/get-file.ashx?itemtype=document&id=1618 [Last accessed 5th May 2015.]

Campbell, T. (2013b). ‘Stereotyped at seven? Biases in teacher judgements of pupils' ability and attainment.’ [Online]. Available at: http://www.cls.ioe.ac.uk/shared/get-file.ashx?itemtype=document&id=1715 [Last accessed 5th May 2015.]

Campbell, T. (2014). ‘Stratified at seven: in-class ability grouping and the relative age effect’. *British Educational Research Journal*, 40(5), 749-771.

Central Advisory Council for Education (1967). ‘Children and their Primary Schools’. (Government document). [Online.] Available at: http://www.educationengland.org.uk/documents/plowden/plowden1967-1.html [Last accessed 5th May 2015.]

Chaplin Grey, J., Gatenby, R., Simmons, N. & Huang, Y (2010). 'Millennium Cohort Study Sweep 4 Technical Report (Second Edition).' [Online]. London: Centre for Longitudinal Studies. Available at: http://www.cls.ioe.ac.uk/shared/get-file.ashx?id=844&itemtype=document [Last accessed 5th May 2015.]

Commons Select Committee (n.d.). [Online]. Education Committee web forum: Summer Born Children. Available at: http://www.parliament.uk/business/committees/committees-a-z/commons-select/education-committee/dfe-evidence-check-forum/summer-born-children/ [Last accessed 5th May 2015.]

Conservative Party (2007). 'Raising the bar, closing the gap: An action plan for schools to raise standards, create more good school places, and make opportunity more equal.' [Online]. Available at: http://image.guardian.co.uk/sys-files/Education/documents/2007/11/20/newopps.pdf [Last accessed 5th May 2015.]

Conservative Party (n.d.) [Online]. Our long-term economic plan: The best schools and skills for young people. Available at: https://www.conservatives.com/Plan/BestSchoolsAndSkills.aspx [Last accessed 5th May 2015.]

Crawford, C., Dearden, L., & Meghir, C. (2007). 'When You Are Born Matters: The Impact of Date of Birth on Child Cognitive Outcomes in England'. [Online]. London: Institute of Fiscal Studies. Available at: http://www.ifs.org.uk/docs/born_matters_report.pdf [Last accessed 5th May 2015.]

Crawford, C., Dearden, L., & Greaves, E. (2011). ‘Does when you are born matter? The impact of month of birth on children’s cognitive and non-cognitive skills in England’. [Online]. London: Institute of Fiscal Studies. Available at http://www.ifs.org.uk/bns/bn122.pdf [Last accessed 5th May 2015.]

Crawford, C., Dearden, L., & Greaves, E. (2013a). ‘When you are born matters: evidence for England.’ [Online]. London: Institute for Fiscal Studies. Available at: http://www.ifs.org.uk/comms/r80.pdf [Last accessed 5th May 2015.]

Crawford, C., Dearden, L., Greaves, E. (2013b). ‘Identifying the drivers of month of birth differences in educational attainment’. [Online]. London: Institute for Fiscal Studies. Available at: http://www.ifs.org.uk/wps/wp201309.pdf [Last accessed 5th May 2015.]

Crawford, C., Dearden, L., & Greaves, E. (2014). ‘The drivers of month-of-birth differences in children’s cognitive and non-cognitive skills’. *Journal of the Royal Statistical Society A*, 177(4), 829-860.

Croizet, J. C., and Claire, T. (1998). ‘Extending the Concept of Stereotype Threat to Social Class: The Intellectual Underperformance of Students from Low Socioeconomic Backgrounds*’. Personality and Social Psychology Bulletin,* 24, 588-594.

Daniels, S., Shorrocks-Taylor, D., & Redfern, E. (2000). ‘Can starting summer-born children earlier at infant school improve their national curriculum results?’ *Oxford Review of Education,* 26(2), 207-220.

Department for Children, Schools and Families (2007). *National Curriculum Assessment, GCSE and Equivalent Attainment and Post-16 Attainment by Pupil Characteristics, in England 2006/07* (Government document). [Online.] Available at: http://www.erpho.org.uk/Download/Public/17242/1/Schools.pdf [Last accessed 5th May 2015.]

Department for Children, Schools and Families (2008a). *Attainment by Pupil Characteristics, in England 2007/08* (Government document). [Online.] Available at: http://webarchive.nationalarchives.gov.uk/20120504203418/http://education.gov.uk/rsgateway/DB/SFR/s000822/sfr32-2008v2.pdf [Last accessed 5th May 2015.]

Department for Children, Schools and Families (2008b). *21st Century Schools: A World Class Education for Every Child* (Government document). [Online.] Available at: http://webarchive.nationalarchives.gov.uk/20130401151715/http://www.education.gov.uk/publications/eOrderingDownload/DCSF-01044-2008.pdf [Last accessed 5th May 2015.]

Department for Children, Schools and Families (2009a). *Your child, your schools, our future: building a 21st century schools system*. (Government document). [Online.] Available at: http://www.educationengland.org.uk/documents/pdfs/2009-white-paper-your-child.pdf [Last accessed 5th May 2015.]

Department for Children, Schools and Families (2009b). *Schools, Pupils and Their Characteristics: January 2009.* (Government document). [Online.] Available at: http://webarchive.nationalarchives.gov.uk/20120504203418/http://education.gov.uk/rsgateway/DB/SFR/s000843/sfr08-2009.pdf [Last accessed 5th May 2015.]

Department for Children, Schools and Families (2009c). *Special Educational Needs in England: January 2009.* (Government document). [Online.] Available at: http://webarchive.nationalarchives.gov.uk/20130401151655/http://media.education.gov.uk/assets/files/pdf/sfr142009pdf.pdf [Last accessed 5th May 2015.]

Department for Children, Schools and Families (2009d). *Children with special educational needs 2009: An analysis.* (Government document). [Online.]   Available at: http://dera.ioe.ac.uk/9446/1/Main.pdf [Last accessed 5th May 2015.]

Department for Education (n.d.1) [Online]. How to interpret school/college performance measures. Available at: http://www.education.gov.uk/schools/performance/about/a3.html [Last accessed 5th May 2015.]

Department for Education (n.d.2) [Online]. School and college performance tables: School and pupil characteristics. Available at: http://www.education.gov.uk/schools/performance/primary_14/p11.html [Last accessed 5th May 2015.]

Department for Education (n.d.3) [Online]. The national curriculum. Available at: https://www.gov.uk/national-curriculum/key-stage-1-and-2 [Last accessed 5th May 2015.]

Department for Education (1992). *White Paper: Choice and Diversity.* (Government document). [Online.] Available at: http://www.educationengland.org.uk/documents/wp1992/choice-and-diversity.html [Last accessed 5th May 2015.]

Department for Education (n.d.) About us: What we do. [Online.] Available at: https://www.gov.uk/government/organisations/department-for-education/about [Last accessed 22nd July 2015.]

Department for Education (2010a). *The Importance of Teaching - The Schools White Paper 2010.* (Government document). [Online.] Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/175429/CM-7980.pdf [Last accessed 5th May 2015.]

Department for Education (2010b). *Month of Birth and Education. (*Government document). [Online.] Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/182664/DFE-RR017.pdf [Last accessed 5th May 2015.]

Department for Education (2010c). *The Importance of Teaching: White Paper Equalities Impact Assessment.* (Government document). [Online.] Available at: https://www.education.gov.uk/publications/eOrderingDownload/CM-7980-Impact_equalities.pdf [Last accessed 5th May 2015.]

Department for Education (2011). *National Curriculum Assessments at Key Stage 2 in England, 2010/2011 (revised).* (Government document). [Online.] Available at: http://www.education.gov.uk/rsgateway/DB/SFR/s001047/sfr31-2011.pdf [Last accessed 5th May 2015.]

Department for Education (2012a). *Early Years Foundation Stage Profile Attainment by Pupil Characteristics, England 2011/12.* Government document). [Online.] Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/219162/sfr30-2012.pdf [Last accessed 5th May 2015.]

Department for Education (2012b). *National Curriculum Assessments at Key Stage 2 in England, 2011/2012 (revised).* (Government document). [Online.] Available at: http://dera.ioe.ac.uk/16228/1/sfr33-2012v2.pdf [Last accessed 5th May 2015.]

Department for Education (2012c). *Phonics Screening Check and National Curriculum Assessments at Key Stage 1 in England: 2012.* (Government document). [Online.] Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/219208/main_20text_20_20sfr21-2012.pdf [Last accessed 5th May 2015.]

Department for Education (2014a). *Phonics screening check and national curriculum assessments at key stage 1 in England, 2014.* (Government document). [Online.] Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/356941/SFR34_2014_text.pdf [Last accessed 15th July 2015]

Department for Education (2014b). *National curriculum assessments at key stage 2 in England, 2014 (Revised).* (Government document). [Online.] Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/428838/SFR50_2014_Text.pdf [Last accessed 15th July 2015]

Department for Education (2014c). *Early years foundation stage profile attainment by pupil characteristics, England 2014.* (Government document). [Online.] Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/376216/SFR46_2014_text.pdf [Last accessed 5th May 2015.]

Department for Education (2014d). *Reforming assessment and accountability for primary schools*. (Government document). [Online.] Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/297595/Primary_Accountability_and_Assessment_Consultation_Response.pdf [Last accessed 5th May 2015.]

Department for Education and Employment (1997). *Excellence in Schools.* (Government document). [Online.] Available at: http://www.educationengland.org.uk/documents/wp1997/excellence-in-schools.html [Last accessed 5th May 2015.]

Department for Education and Skills (2001). *Schools Achieving Success*. (Government document). [Online.] Available at: http://www.educationengland.org.uk/documents/pdfs/2001-schools-achieving-success.pdf [Last accessed 5th May 2015.]

Department for Education and Skills (2005). *Higher Standards, Better Schools for All: More choice for parents and pupils.* (Government document). [Online.] Available at: http://dera.ioe.ac.uk/5496/1/DfES-Schools%20White%20Paper.pdf [Last accessed 5th May 2015.]

Dhuey, E. & Lipscomb, S. (2010). 'Disabled or Young? Relative Age and Special Education Diagnoses in Schools'. *Economics of Education Review 29*, 857–872.

Dockrell, J., Lindsay, G., Palikara, O., Cullen, M. (2007). *Raising the Achievements of Children and Young People with Specific Speech and Language Difficulties and other Special Educational Needs through School to Work and College*. (Government document). [Online.] Available at: http://dera.ioe.ac.uk/7860/1/Dockrell2007raising.pdf [Last accessed 5th May 2015.]

Dunne, M., Humphreys, S., Sebba, J., Dyson, A., Gallannaugh, F., & Muijs, D. (2007). *Effective Teaching and Learning for Pupils in Low Attaining Groups.* (Government document). [Online.] Available at: http://dera.ioe.ac.uk/6622/1/DCSF-RR011.pdf [Last accessed 5th May 2015.]

Earp, B. D. (2010). 'Automaticity in the classroom: Unconscious mental processes and the racial achievement gap'. *Journal of Multiculturalism in Education*, 6(1), 1-22.

Eckert, T. L., Dunn, E. K., Codding, R. S., Begeny, J. C., & Kleinmann, A. E. (2006). 'Assessment of mathematics and reading performance: An examination of the correspondence between direct assessment of student performance and teacher report'. *Psychology in the Schools*, 43(3), 247-265.

Education Committee (2015). Correspondence to Nick Gibb, MP. [Online]. Available at: http://www.parliament.uk/documents/commons-committees/Education/Committee-letter-to-Nick-Gibb-starting-school.pdf [Last accessed 5th May 2015.]

Education Endowment Foundation (n.d.a) About EEF Evaluation. [Online]. Available at: https://educationendowmentfoundation.org.uk/evaluation/about-eef-evaluation/ [Last accessed 5th May 2015.]

Education Endowment Foundation (n.d.b) Setting or Streaming. [Online]. Available at: https://educationendowmentfoundation.org.uk/toolkit/toolkit-a-z/ability grouping/ [Last accessed 5th May 2015.]

Elder, T. E. & Lubotsky, D. H. (2009). 'Kindergarten Entrance Age and Children's Achievement: Impacts of State Policies, Family Background, and Peers'. *Journal of Human Resources*, 44(3), 641-683.

Foster, R. G., & Roenneberg, T. (2008). 'Human responses to the geophysical daily, annual and lunar cycles'*.* *Current Biology*, 18(17), 784-794.

Francis, B., Archer, L., Hodgen, J., Pepper, D., Taylor, B., & Travers, M. (2016) Exploring the relative lack of impact of research on "ability grouping" in England: a discourse analytic account, Cambridge Journal of Education, early view online: http://dx.doi.org/10.1080/0305764X.2015.1093095

Gibb, N. (2015). Letter in response to the Education Select Committee on Summer Born Children. [Online]. Available at: http://www.parliament.uk/documents/commons-committees/Education/Nick-Gibb-letter-on-summer-born-admissions.pdf [Last accessed 22nd July 2015.]

Gledhill, J., Ford, T., & Goodman, R. (2002). 'Does season of birth matter? The relationship between age within the school year (season of birth) and educational difficulties among a representative general population sample of children and adolescents (aged 5-15) in Great Britain'. *Research in Education,* 68, 41-47.

Good, T. L. (1987), 'Two Decades of Research on Teacher Expectations: Findings and Future Directions'. *Journal of Teacher Education,* 38, 32-47.

Goodman, R., Gledhill J., & Ford, T. (2003). 'Child psychiatric disorder and relative age within school year: cross sectional survey of large population sample.' *British Medical Journal,* 327 472-475.

Guardian (Thursday 9 February 2012 17.30 GMT): 'Dividing younger pupils by ability can entrench disadvantage, study finds'. [Online]. Available at: http://www.theguardian.com/education/2012/feb/09/dividing-pupils-ability-entrench-disadvantage [Last accessed 5th May 2015.]

Hallam, S. & Ireson, J. (1999). 'Pedagogy in the Secondary School.' In Mortimore, P. (Ed.), *Understanding Pedagogy*. London: Chapman.

Hallam, S. Ireson, J., Judith, Lister, V., Andon Chaudhury, I. & Davies, J. (2003). 'Ability grouping in the primary school: a survey*'. Educational Studies*, 29(1), 69-83.

Hallam, S. & Parsons, S. (2013). 'Prevalence of streaming in UK primary schools: Evidence from the Millennium Cohort Study*'. British Educational Research Journal*, 39(3), 514-544.

Hallam, S. & Parsons, S. (2014). 'The impact of streaming on attainment at age seven: evidence from the Millennium Cohort Study'. *Oxford Review of Education,* 40(5), 567-589.

Hansen, K. & Jones, E. (2011). 'Ethnicity and gender gaps in early Childhood'. *British Educational Research Journal,* 37(6), 973-991.

Hansen, K. [Ed.] (2012). *Millennium Cohort Study: First, Second, Third and Fourth Surveys. A Guide to the Datasets (Seventh Edition).* London: Centre for Longitudinal Studies. [Online]. Available at: http://www.cls.ioe.ac.uk/shared/get-file.ashx?id=598&itemtype=document [Last accessed 5th May 2015.]

Harlen, W. (2004). ‘A systematic review of the evidence of the impact on students, teachers and the curriculum of the process of using assessment by teachers for summative purposes’. [Online]. London: EPPI-Centre. Available at: https://eppi.ioe.ac.uk/cms/LinkClick.aspx?fileticket=Pbyl1CdsDJU%3D&tabid=108&mid=1003 [Last accessed 5th May 2015.]

Harlen, W. (2005). ‘Trusting teachers’ judgement: research evidence of the reliability and validity of teachers’ assessment used for summative purposes’. *Research Papers in Education*, 20(3), 245-270.

Harlen, W. (2007). *The quality of learning: assessment alternatives for primary education. Primary Review Research Survey 3/4*. Cambridge: University of Cambridge Faculty of Education.

Hilton, J. L. & von Hipple. W. (1996). ‘Stereotypes.’ *Annual Review of Psychology,* 47, 237–71.

Huang, Y., & Gatenby, R. (2010), ‘Millennium Cohort Study Sweep 4 Teacher Survey Technical Report’. [Online]. London: Centre for Longitudinal Studies. Available at: http://www.cls.ioe.ac.uk/shared/get-file.ashx?id=489&itemtype=document [Last accessed 5th May 2015.]

Ireson, J. & Hallam, S. (1999). ‘Raising standards: Is ability grouping the answer?’ *Oxford Review of Education*, 25(3), 344–60.

Johnson, J., Rosenberg, R., Platt, L. & Parsons, S. (2011). ‘Millennium Cohort Study Fourth Survey: A Guide to the Teacher Survey Dataset 1st Edition.’ [Online]. London: Centre for Longitudinal Studies. Available at: http://www.cls.ioe.ac.uk/shared/get-file.ashx?id=1341&itemtype=document [Last accessed 5th May 2015.]

Johnson, J. and Rosenberg, R. (2013). ‘A Guide to the Linked Education Administrative Datasets: Millennium Cohort Study.’ [Online]. London: Centre for Longitudinal Studies. Available at: http://www.cls.ioe.ac.uk/shared/get-file.ashx?id=1342&itemtype=document [Last accessed 5th May 2015.]

Kendall, S., Straw, S. Jones, M. Springate, I. & Grayson, H (2008). ‘A Review of the Research Evidence (Narrowing the Gap in Outcomes for Vulnerable Groups).’ [Online]. Slough: NFER. Available at: https://www.nfer.ac.uk/publications/LNG01/LNG01.pdf [Last accessed 5th May 2015.]

Kerr, K & West, M (2010). ‘Social inequality: can schools narrow the gap?’ [Online]. Macclesfield: British Educational Research Association. Available at: https://www.bera.ac.uk/wp-content/uploads/2014/01/Insight2-web.pdf [Last accessed 5th May 2015.]

Ketende, S. (Ed.) (2010). *Millennium Cohort Study: Technical Report on Response, Third Edition.* London: Centre for Longitudinal Studies. [Online]. Available at: http://www.cls.ioe.ac.uk/shared/get-file.ashx?itemtype=document&id=1625 [Last accessed 5th May 2015]

Kuklinski, M. & Weinstein, R. (2000). ‘Classroom and Grade Level Differences in the Stability of Teacher Expectations and Perceived Differential Teacher Treatment’. *Learning Environments Research,* 3(1), 1-34.

Kutnick, P., Sebba, J., Blatchford, P., Galton, M., & Thorp, J. (2005). *The effects of pupil grouping: Literature review.* (Government document). [Online.] Available at: http://webarchive.nationalarchives.gov.uk/20130401151715/http://www.education.gov.uk/publications/eOrderingDownload/RR688.pdf [Last accessed 5th May 2015.]

Kutnick, P., Hodgkinson, S. Sebba, J., Humphreys, S., Galton, M., Steward, S., Blatchford, P., Baines, E. (2006). *Pupil Grouping Strategies and Practices at Key Stage 2 and 3: Case Studies of 24 Schools in England.* (Government document). [Online.] Available at: http://www.leics.gov.uk/grouping_pupils_for_success_full_report.pdf [Last accessed 5th May 2015.]

Landsberger, H. A. (1958). *Hawthorn Revisited.* New York: Cornell.

Lawlor, H., Clark, H., Ronalds, G., & Leon, D. (2006). 'Season of birth and childhood intelligence: findings from the Aberdeen Children of the 1950s cohort study'. *British Journal of Educational Psychology,* 76(3), 481-499.

Lupton, R. & Thomson, S. (2015). 'The Coalition's Record on Schools: Policy, Spending and Outcomes 2010-2015.' [Online]. London: London School of Economics. Available at: http://sticerd.lse.ac.uk/dps/case/spcc/wp13.pdf [Last accessed 5th May 2015.]

McEwan, P. J. & Shapiro, J. S. (2008). 'The Benefits of Delayed Primary School Enrollment: Discontinuity Estimates Using Exact Birth Dates'. *Journal of Human Resources,* 43(1), 1-29.

McGarty, C., Yzerbyt, V. Y. and Spears, R. (2002). *Stereotypes as Explanations: The Formation of Meaningful Beliefs about Social Groups.* Cambridge: Cambridge University Press.

Martin, R. P., Foels, P., Clanton, G. & Moon, K. (2004). 'Season of birth is related to child retention rates, achievement, and rates of diagnosis with specific LD'. *Journal of Learning Disabilities,* 37(4), 307-317.

Maylor, U., Smart, S., Kuyok, K. A., Ross, A. (2009). *Black Children's Achievement Programme Evaluation*. (Government document). [Online.] Available at: http://dera.ioe.ac.uk/11380/1/DCSF-RR177.pdf [Last accessed 5th May 2015.]

Menet, F., Eakin, J., Stuart, M., & Rafferty, H. (2000). 'Month of Birth and Effect on Literacy, Behaviour and Referral to Psychological Service'. *Educational Psychology in Practice,* 16(2) 225-234.

Miller, K. & Satchwell, C. (2006). 'The effect of beliefs about literacy on teacher and student expectations: a further education perspective'. *Journal of Vocational Education and Training,* 58(2), 135–150.

Mostapha, T. (2013). 'Technical Report on Response in the Teacher Survey in MCS 4 (Age 7).' [Online]. London: Centre for Longitudinal Studies. Available at http://www.cls.ioe.ac.uk/shared/get-file.ashx?id=1749&itemtype=document [Last accessed 5th May 2015.]

Mühlenweg, A. (2010). 'Young and innocent: international evidence on age effects within grades on victimization in elementary school'. *Economics Letters,* 109, 157-60.

Ofqual (2012). 'GCSE English 2012'. [Online]. Available at: http://www.rewardinglearning.org.uk/docs/accreditation/projects/OfqualReport.pdf [Last accessed 5th May 2015.]

OECD (2012). 'Equity and Quality in Education: Supporting Disadvantaged Students and Schools.' [Online]. OECD Publishing. Available at: http://www.keepeek.com/Digital-Asset-Management/oecd/education/equity-and-quality-in-education_9789264130852-en_ [Last accessed 5th May 2015.]

Oshima, T.C., & Domaleski, C. S. (2006). 'Academic Performance Gap Between Summer-Birthday and Fall-Birthday Children in Grades K-8'. *The Journal of Educational Research,* 99(4), 212-217.

Plewis, I. (Ed.) (2007). *The Millennium Cohort Study: Technical Report on Sampling: Fourth Edition.* London: Centre for Longitudinal Studies. [Online]. Available at: http://www.cls.ioe.ac.uk/shared/get-file.ashx?id=409&itemtype=document [Last accessed 5th May 2015.]

Plewis, I., Ketende, S., Calderwood, L. (2010). ‘Assessing the accuracy of response propensities in longitudinal studies’. [Online]. Manchester: The Cathie Marsh Centre for Census and Survey Research. Available at: http://surveynet.ac.uk/sdmi/ccsr_2010-08.pdf [Last accessed 5th May 2015.]

Polizzi, N., Martin, R., & Dombrowski, S. (2007). ‘Season of birth of students receiving special education services under a diagnosis of emotional and behavioural disorder’. *School Psychology Quarterly,* 22(1) 44-57.

Pre-School Learning Alliance (2015). Written evidence submitted by the Pre-school Learning Alliance to the Commons Select Committee. [Online]. Available at: http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/education-committee/evidence-check-starting-school/written/18264.html [Last accessed 5th May 2015.]

Reay, D. (2006). ‘The Zombie stalking English schools: Social class and Educational Inequality’. *British Journal of Educational Studies* 54(3), 288-307.

Reeves, D. J., Boyle, W. F. & Christie, T. (2001). ‘The Relationship between Teacher Assessments and Pupil Attainments in Standard Test Tasks at Key Stage 2, 1996-98’. *British Educational Research Journal,* 27(2), 141-160.

Riggall, A., & Sharp, C. (2008). ‘The Structure of Primary Education: England and Other Countries. (Primary Review Research Report 9/1).’ [Online]. Available at: https://www.nfer.ac.uk/publications/PRO01/PRO01.pdf [Last accessed 5th May 2015.]

Robinson, J. P. and Lubienski, S. T. (2011). ‘The Development of Gender Achievement Gaps in Mathematics and Reading During Elementary and Middle School: Examining Direct Cognitive Assessments and Teacher Ratings’. *American Educational Research Journal,* 48(2), 268–302.

Rosenthal, R., & Jacobsen, L. (1968). ‘Pygmalion in the classroom’. *The Urban Review,* 3(1), 16-20.

Rubie-Davies, C. M. (2010). ‘Teacher expectations and perceptions of student characteristics: Is there a relationship?’ *British Journal of Educational Psychology,* 80(1), 121-135.
Sampaio, B., da Matta, R., Ribas, R. & Sampaio, G. (2011). ‘The effect of age on college entrance test score and enrolment: a regression discontinuity approach.’ [Online]. Available at: http://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID1766222_code1337569.pdf?abstractid=1471686&mirid=1 [Last accessed 5th May 2015.]

Sharp, C., George, N., Sargent, C., O’Donnell, S., & Heron, M. (2009). ‘The influence of relative age on learner attainment and development.’ [Online]. Available at https://www.nfer.ac.uk/what-we-do/information-and-reviews/inca/RelativeAgeReviewRevised2012.pdf [Last accessed 5th May 2015.]

Shih, M., Pittinsky, T. L., and Trahan, A. (2005). ‘Domain specific effects of stereotypes on performance: Working paper no. RWP05-026.’ [Online]. US: Harvard University. Available at: http://www.cs.cmu.edu/~cfrieze/courses/Shih.pdf [Last accessed 5th May 2015.]

Sprietsma, M. (2007). ‘The effect of relative age in the first grade of primary school on long-term scholastic results: International comparative evidence using PISA 2003’. *Education Economics*, 18(1), 1-32.

Steele, C. M., & Aronson, J. (1995). ‘Stereotype threat and the intellectual test performance of African Americans’. *Journal of Personality and Social Psychology* 69, 797-881.

Slavin, R. E. (1990). ‘Achievement effects of ability grouping in secondary schools: a best evidence synthesis’. *Review of Educational Research,* 60, 471-499.

Strand, S. (2007). *Minority Ethnic Pupils in the Longitudinal Study of Young People in England (LSYPE).* (Government document). [Online.] Available at: http://www.irr.org.uk/pdf/DCSF_Strand_full.pdf [Last accessed 5th May 2015.]

Strom, B. (2004). *Student achievement and birthday effects*. Paper prepared for presentation at the CESifo-Harvard University/PEPG Conference on Schooling and Human Capital in the Global Economy: Revisiting the Equity-Efficiency Quandary. CESifo Conference Center, Munich, September 3-4, 2004. [Online]. Available at: http://www.hks.harvard.edu/pepg/PDF/events/Munich/PEPG-04-24Strom.pdf [Last accessed 5th May 2015.]

Strand, S., de Coulon, A., Meschi, E., Vorhouse, J., Frumkin, L., Ivins, C., Small, L. Sood, A., Gervais, M. C., Rehman, H. (2010). *Drivers and Challenges in Raising the Achievement of Pupils from Bangladeshi, Somali and Turkish Backgrounds*. (Government document). [Online.] Available at: https://www.education.gov.uk/publications/eOrderingDownload/DCSF-RR226.pdf [Last accessed 5th May 2015.]

Sullivan, A., Heath, A. & Rothon, C. (2011). ‘Equalisation or inflation? Social class and gender differentials in England and Wales’. Oxford Review of Education, 37(2), 215-240.

Sutton Trust / Educational Endowment Foundation (2014). ‘Teaching and Learning Toolkit.’ [Online]. Available at: http://educationendowmentfoundation.org.uk/uploads/toolkit/EEF_Teaching_and_learning_toolkit_Feb_2014.pdf [Last accessed 5th May 2015.]

Sykes, E. D. A., Bell, J. F., & Rodeiro, C. V. (2009). 'Birthdate Effects: A Review of the Literature from 1990-on'. [Online]. Cambridge: University of Cambridge. Available at: http://www.cambridgeassessment.org.uk/Images/109784-birthdate-effects-a-review-of-the-literature-from-1990-on.pdf [Last accessed 5th May 2015.]

The Association for Professional Development in Early Years (2015). Written evidence submitted by the TACTYC to the Commons Select Committee. [Online]. Available at: http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/education-committee/evidence-check-starting school/written/18334.pdf [Last accessed 5th May 2015.]

Thomas, S., Smees, R., Madaus, G. F., and Raczek, A. E. (1998). 'Comparing Teacher Assessment and Standard Task Results in England: the relationship between pupil characteristics and attainment'. *Assessment in Education: Principles, Policy & Practice*, 5(2), 213-246.

Tikly, L., Haynes, J., Caballero, C., Hill, J., Gillborn, D. (2008). *Evaluation of Aiming High: African Caribbean Achievement Project*. (Government document). [Online.] Available at: http://webarchive.nationalarchives.gov.uk/20130401151715/http://www.education.gov.uk/publications/eOrderingDownload/RR801.pdf [Last accessed 5th May 2015.]

University of London. Institute of Education. Centre for Longitudinal Studies, *Millennium Cohort Study: First Survey, 2001-2003* [computer file]. *11th Edition.* Colchester, Essex: UK Data Archive [distributor], December 2012. SN: 4683, http://dx.doi.org/10.5255/UKDA-SN-4683-3.

University of London. Institute of Education. Centre for Longitudinal Studies, *Millennium Cohort Study: Fourth Survey, Teacher Survey, 2008* [computer file]. Colchester, Essex: UK Data Archive [distributor], August 2011a. SN: 6848 , http://dx.doi.org/10.5255/UKDA-SN-6848-1

University of London. Institute of Education. Centre for Longitudinal Studies, *Millennium Cohort Study: Third Survey, Teacher Survey and Foundation Stage Profile, 2006* [computer file]. Colchester, Essex: UK Data Archive [distributor], August 2011a. SN: 6847, http://dx.doi.org/10.5255/UKDA-SN-6847-1

University of London. Institute of Education. Centre for Longitudinal Studies, *Millennium Cohort Study, 2001-2008: Linked Education Administrative Dataset, England: Secure Access* [computer file]. Colchester, Essex: UK Data Archive [distributor], November 2011b. SN: 6862, http://dx.doi.org/10.5255/UKDA-SN-6862-2

University of London. Institute of Education. Centre for Longitudinal Studies, *Millennium Cohort Study: Fourth Survey, 2008* [computer file]. *4th Edition.* Colchester, Essex: UK Data Archive [distributor], December 2012a. SN: 6411, http://dx.doi.org/10.5255/UKDA-SN-6411-3

University of London. Institute of Education. Centre for Longitudinal Studies, *Millennium Cohort Study: Third Survey, 2006* [computer file]. *6th Edition.* Colchester, Essex: UK Data Archive [distributor], December 2012b. SN: 5795, http://dx.doi.org/10.5255/UKDA-SN-5795-3

Wallingford, E. L., & Prout, H. T. (2000). 'The relationship between season of birth and special education referral'. *Psychology in the Schools*, 37(4), 379-387.

William, D. & Bartholomew, H. (2004). 'It's not which school but which set you're in that matters: the influence on ability grouping practices on student progress in mathematics'. *British Educational Research Journal,* 30(2), 279–294.

Whitty, G & Anders, J. (2014). '(How) did New Labour narrow the achievement and participation gap?' [Online]. London: The Centre for Learning and Life Chances in Knowledge Economies and Societies. Available at: http://www.llakes.org/wp-content/uploads/2014/01/46.-Whitty-and-Anders-final.pdf [Last accessed 5th May 2015.]

Wilson, G. (2000). 'The effects of season of birth, sex and cognitive abilities on the assessment of special educational needs'*. Educational Psychology,* 20(2), 153-166.

Wyse, D., McCreery, E., & Torrance, H. (2008). 'The trajectory and impact of national reform: curriculum and assessment in English primary education. Primary Review Research Survey 3/2.' [Online]. Cambridge: University of Cambridge Faculty of Education. Available at: http://image.guardian.co.uk/sys-files/Education/documents/2008/02/29/RSreport1.pdf [Last accessed 5th May 2015.]

Yopyk, D. J. A. (2005). 'Am I an athlete or a student? Identity salience and stereotype threat in student-athletes'. *Basic and Applied Social Psychology* 27(4), 329-336.

Annex 3A

Characteristics of those English MCS wave four, teacher sample, singleton, state school pupils who are streamed / not streamed

Table 3A1 presents (a) discrete descriptive statistics for percentage of MCS wave four teacher survey pupils with each respective characteristic who are streamed, and (b) coefficients and p-values from a probit regression where the outcome is streamed / not and each characteristic is simultaneously included as a predictor.

The descriptive statistics provide some indication that sample pupils of certain ethnic groups are more likely to be streamed than others, as well as low-income children, those whose parents have lower or overseas qualifications, and those whose families speak languages in addition to English at home. There are also some discrepancies according to birth month. However, when all characteristics are accounted for at once in the probit regression, only having a main parent with overseas qualifications and being born in June remain significantly related to being streamed (while being of Indian or Pakistani / Bangladeshi ethnicity is of borderline significance). Pupils with all other characteristics appear equally as likely as their reference comparators to be streamed.

Table 3A1: Percentage of sample^ pupils who are streamed and coefficients from probit regression of whether streamed / not where each characteristic is simultaneously included as predictor^^

	Percentage streamed (a)	Probit regression coefficient (b)
All sample pupils (n = 4999 / 4951)	17.6	
Boys (n = 2508)	17.8	(reference)
Girls (n = 2491)	17.2	.022 (.045)
White (4000)	17.1	(reference)
Mixed ethnicity (169)	20.3	.143 (.138)
Indian (148)	24.7	.295 (.172)*
Pakistani / Bangladeshi (363)	23.8	.238 (.142)*
Black / Black British (193)	14.4	-.072 (.158)
Other ethnic group (81)	15.8	-.063 (.249)
Higher-income (3577)	17.2	(reference)
Low-income (1418)	18.5	-.051 (.062)
Parent NVQ level 1 (373)	19.2	.144 (.141)
Parent NVQ level 2 (1413)	18.7	.156 (.120)
Parent NVQ level 3 (722)	18.3	.143 (.130)
Parent NVQ level 4 (1489)	14.5	-.026 (.111)
Parent NVQ level 5 (318)	15.2	(reference)
Overseas qualifications only (167)	25.6	.371 (.159)**
No qualifications (515)	19.6	.187 (.133)
Speaks other languages at home (689)	20.7	.041 (.118)
Speaks English only (4310)	17.2	(reference)
August-born (357)	17.0	.031 (.141)
July-born (374)	17.4	.045 (.132)
June-born (434)	23.5	.258 (.106)**
May-born (396)	18.4	.085 (.113)
April-born (402)	14.4	-.068 (.118)
March-born (422)	18.1	.071 (.107)
February-born (374)	13.2	-.130 (.123)
January-born (429)	18.6	.091 (.106)
December-born (453)	20.4	.163 (.108)
November-born (463)	15.9	-.022 (.102)
October-born (430)	16.1	-.010 (.107)
September-born (465)	16.6	(reference)

Standard errors in brackets. *** = p < .001; ** = p < .05; * = p < .10

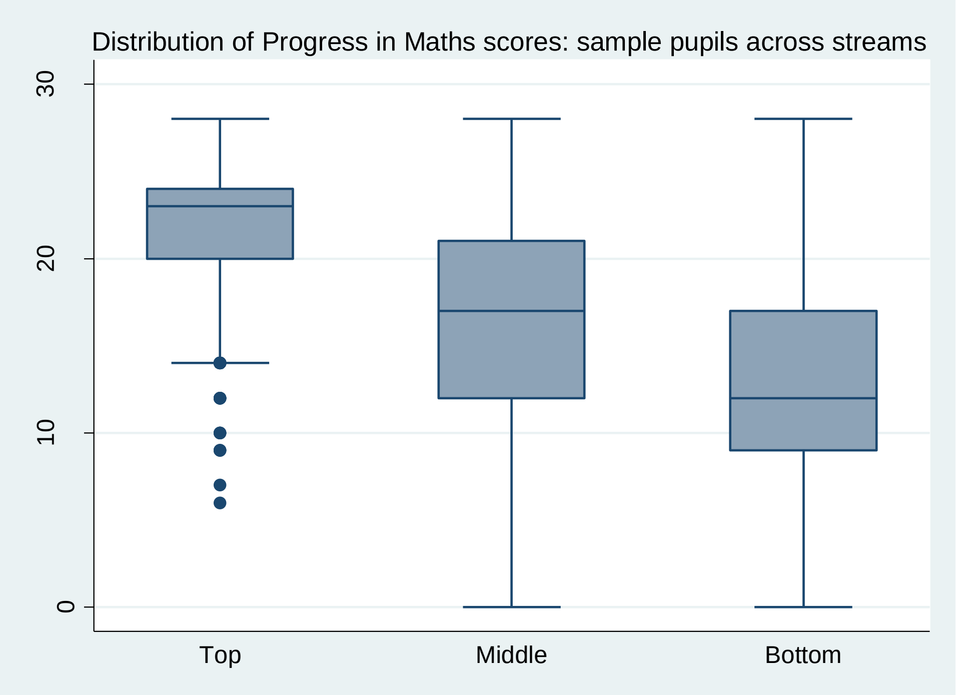
^MCS wave four teacher sample pupils interviewed in England, singleton children in state schools only. ^^All estimates weighted for survey design and attrition to main wave four survey.

Ns are unweighted and are for descriptive statistics (sample sizes are slightly smaller for the regression due to list-wise deletion – 4,951 [vs 4999] cases in total are included in the model)

Annex 3B

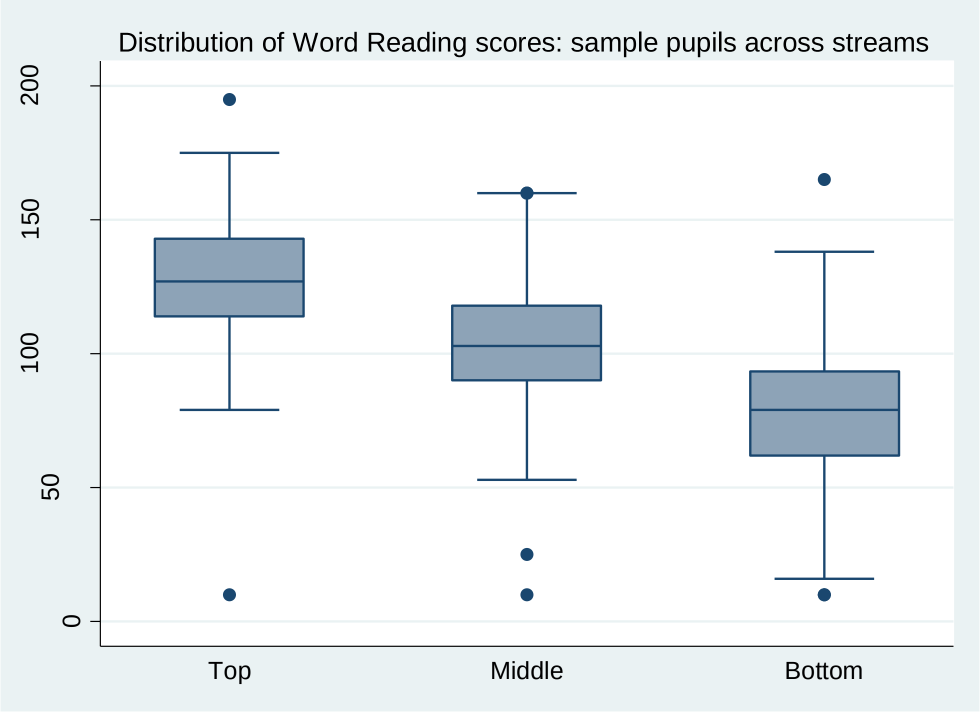
Distribution of scores on the three cognitive tests for pupils situated in each stream, in sample with KS1 scores

Figure 3B1:



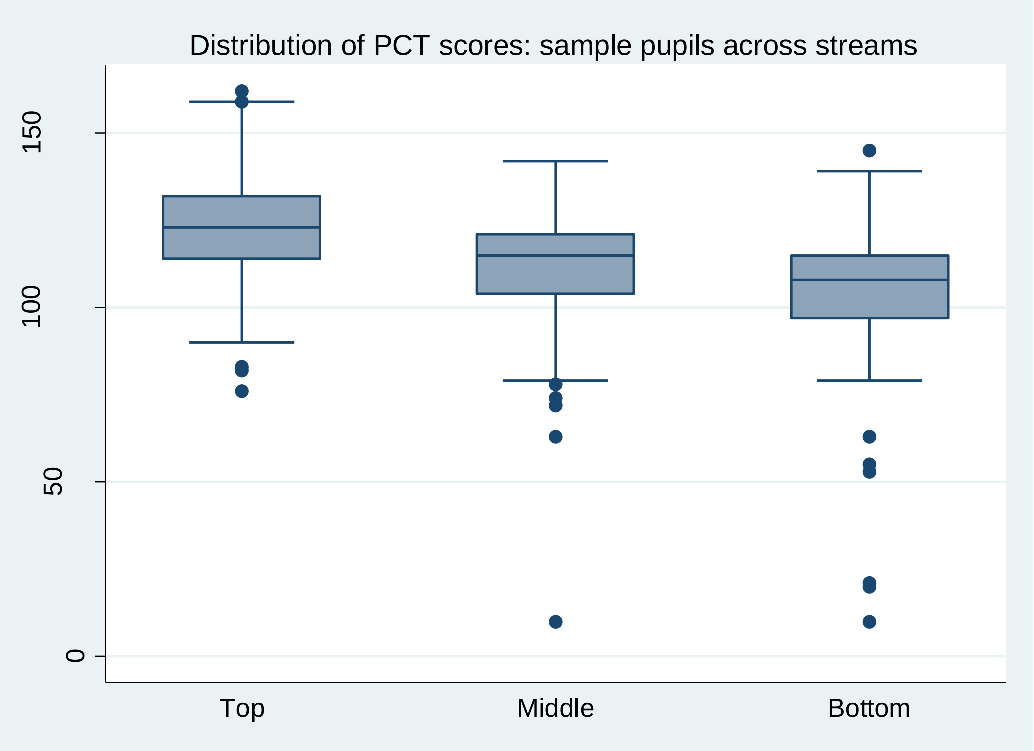
n = 644; Mean for all pupils = 18.2. Line represents median, box represents 25th and 75th percentiles (Q1 and Q3, respectively), whiskers represent Q3+1.5(Q3-Q1) / Q1-1.5*(Q3-Q1).

Figure 3B2:



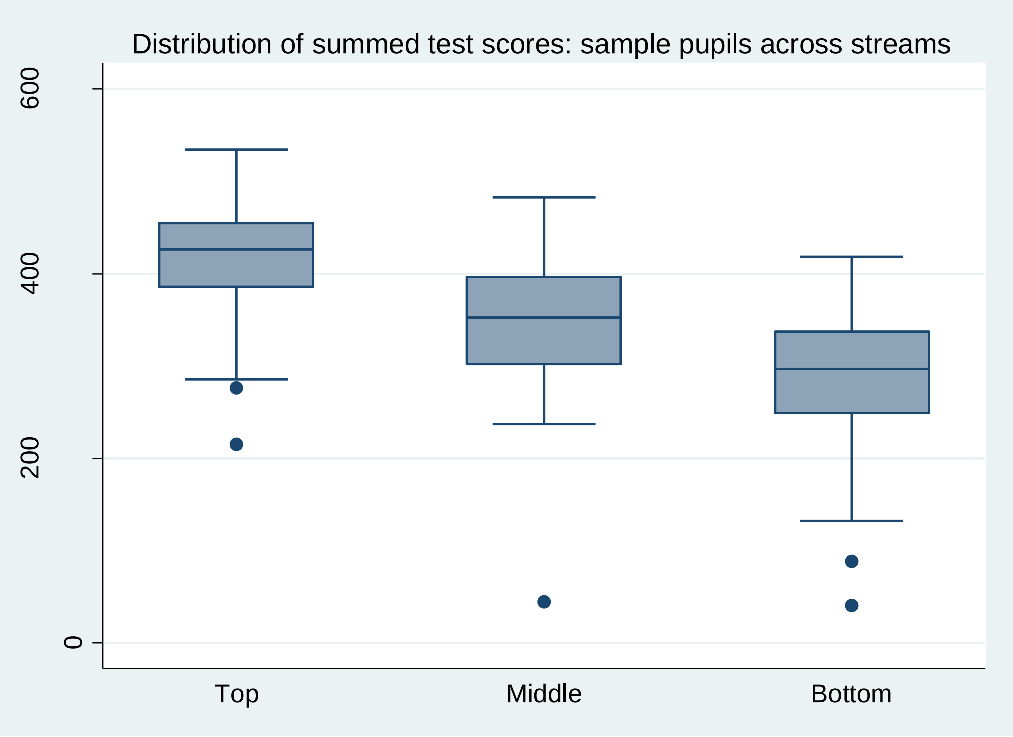
n = 644; Mean for all pupils = 108.9. Line represents median, box represents 25th and 75th percentiles (Q1 and Q3, respectively), whiskers represent Q3+1.5(Q3-Q1) / Q1-1.5*(Q3-Q1).

Figure 3B3:



n = 642; Mean for all pupils = 115.1. Line represents median, box represents 25th and 75th percentiles (Q1 and Q3, respectively), whiskers represent Q3+1.5(Q3-Q1) / Q1-1.5*(Q3-Q1).

Figure 3B4:



n = 639; Mean for all pupils = 367.9. Line represents median, box represents 25th and 75th percentiles (Q1 and Q3, respectively), whiskers represent Q3+1.5(Q3-Q1) / Q1-1.5*(Q3-Q1)

Annex 3C

Difference in teacher-assessed Key Stage One average point score according to pupils’ stream placement: all covariates

Table 3C1: Difference in teacher-assessed Key Stage One average point score according to pupils’ stream placement^

	Spec 1	**Spec 2**	**Spec 3**	**Spec 4**	**Spec 5**	**Spec 6**
Top stream	1.335***	1.371***	1.375***	1.229***	1.230***	1.209***
	(0.208)	(0.210)	(0.193)	(0.198)	(0.199)	(0.198)
(Middle stream)	0	0	0	0	0	0
	(.)	(.)	(.)	(.)	(.)	(.)
Bottom stream	-1.677***	-1.586***	-1.395***	-1.376***	-1.275***	-1.266***
	(0.234)	(0.231)	(0.238)	(0.236)	(0.250)	(0.255)
Maths Test score	0.101***	0.0963***	0.0816***	0.0779***	0.0755***	0.0781***
	(0.018)	(0.019)	(0.017)	(0.016)	(0.016)	(0.016)
Word Reading Test score	0.0520***	0.0498***	0.0488***	0.0470***	0.0462***	0.0458***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Pattern Construction Test score	0.0256***	0.0240***	0.0203***	0.0201***	0.0206***	0.0198***
	(0.006)	(0.006)	(0.005)	(0.005)	(0.005)	(0.005)
Age at tests	-0.118*	-0.114^{+}	-0.129*	-0.126*	-0.126*	-0.119*
	(0.054)	(0.058)	(0.057)	(0.056)	(0.055)	(0.055)
August-born	-1.482*	-1.401*	-1.532**	-1.266*	-1.228*	-1.222*
	(0.592)	(0.598)	(0.583)	(0.586)	(0.576)	(0.577)
July-born	-1.781**	-1.744**	-2.006**	-1.762**	-1.776**	-1.758**
	(0.606)	(0.640)	(0.611)	(0.607)	(0.591)	(0.589)
June-born	-1.699***	-1.613**	-1.779***	-1.588**	-1.624***	-1.579***
	(0.486)	(0.504)	(0.491)	(0.477)	(0.464)	(0.462)

May-born	-1.175*	-1.096*	-1.390**	-1.268*	-1.274**	-1.204*
	(0.474)	(0.512)	(0.496)	(0.489)	(0.479)	(0.477)
April-born	-1.137**	-1.065*	-1.106**	-0.910*	-0.898*	-0.932*
	(0.427)	(0.423)	(0.403)	(0.421)	(0.414)	(0.410)
March-born	-0.507	-0.426	-0.458	-0.348	-0.377	-0.368
	(0.375)	(0.374)	(0.375)	(0.375)	(0.375)	(0.384)
February-born	-0.406	-0.300	-0.339	-0.237	-0.286	-0.307
	(0.548)	(0.515)	(0.487)	(0.488)	(0.479)	(0.482)
January-born	-0.694	-0.622	-0.607	-0.473	-0.467	-0.504
	(0.436)	(0.430)	(0.385)	(0.374)	(0.375)	(0.374)
December-born	-0.305	-0.254	-0.403	-0.327	-0.343	-0.288
	(0.383)	(0.369)	(0.332)	(0.318)	(0.318)	(0.312)
November-born	-0.269	-0.242	-0.359	-0.329	-0.349	-0.342
	(0.379)	(0.363)	(0.347)	(0.343)	(0.350)	(0.342)
October-born	0.112	0.224	0.0597	0.110	0.115	0.117
	(0.352)	(0.356)	(0.339)	(0.341)	(0.345)	(0.337)
(September-born)	0	0	0	0	0	0
	(.)	(.)	(.)	(.)	(.)	(.)
Boy		-0.186	-0.0948	-0.0666	-0.0549	-0.0539
		(0.146)	(0.138)	(0.137)	(0.140)	(0.140)
(Girl)		0	0	0	0	0
		(.)	(.)	(.)	(.)	(.)
(White ethnicity)		0	0	0	0	0
		(.)	(.)	(.)	(.)	(.)
Mixed / ‘other’ / missing data		0.548^{+}	0.524*	0.558*	0.494^{+}	0.527^{+}
		(0.283)	(0.256)	(0.264)	(0.262)	(0.275)
Indian		0.710*	0.465	0.520	0.439	0.424
		(0.343)	(0.355)	(0.335)	(0.328)	(0.336)
Pakistani / Bangladeshi		-0.154	-0.132	-0.0917	-0.179	-0.127
		(0.271)	(0.298)	(0.315)	(0.304)	(0.314)
Black / Black British		-0.421	-0.610	-0.541	-0.586	-0.573
		(0.463)	(0.466)	(0.504)	(0.489)	(0.453)

(Higher-income)	0 (.)	0 (.)	0 (.)	0 (.)	0 (.)
Low-income	-0.435* (0.195)	-0.375* (0.186)	-0.342^{+} (0.175)	-0.315^{+} (0.176)	-0.367* (0.181)
Parent level 1 qual	0.0212 (0.381)	0.0762 (0.369)	0.133 (0.361)	0.131 (0.360)	0.167 (0.359)
Parent level 2 qual	0.0861 (0.323)	0.108 (0.327)	0.118 (0.323)	0.105 (0.322)	0.144 (0.301)
Parent level 3 qual	0.115 (0.357)	0.126 (0.359)	0.150 (0.354)	0.120 (0.352)	0.143 (0.336)
Parent level 4 qual	0.248 (0.302)	0.246 (0.303)	0.254 (0.298)	0.235 (0.296)	0.256 (0.280)
(Parent level 5 qual – ref)	0 (.)	0 (.)	0 (.)	0 (.)	0 (.)
Parent overseas qual	0.318 (0.475)	0.368 (0.432)	0.421 (0.422)	0.455 (0.431)	0.522 (0.414)
Parent no qual	-0.477 (0.435)	-0.352 (0.423)	-0.255 (0.422)	-0.237 (0.419)	-0.265 (0.400)
Community mainstream school	0 (.)	0 (.)	0 (.)	0 (.)	0 (.)
Voluntary aided school	-0.0762 (0.230)	-0.0353 (0.228)	-0.0104 (0.229)	-0.0247 (0.231)	-0.0345 (0.225)
Voluntary controlled / foundation	0.269 (0.292)	0.298 (0.277)	0.206 (0.268)	0.212 (0.277)	0.196 (0.274)
(Did not join school in current academic year)	0 (.)	0 (.)	0 (.)	0 (.)	0 (.)
Joined school in current academic year	-0.244 (0.267)	-0.197 (0.280)	-0.244 (0.276)	-0.268 (0.278)	-0.237 (0.280)
(Did not join school in last academic year)	0 (.)	0 (.)	0 (.)	0 (.)	0 (.)

Joined school in last academic year	0.0611 (0.346)	0.100 (0.309)	0.109 (0.316)	0.115 (0.324)	0.0696 (0.310)
(Age five SDQ emotional – ‘normal’)		0 (.)	0 (.)	0 (.)	0 (.)
Age five SDQ emotional – ‘borderline’		0.0817 (0.327)	0.111 (0.338)	0.0696 (0.342)	0.0457 (0.342)
Age five SDQ emotional – ‘abnormal’		-0.397 (0.327)	-0.440 (0.316)	-0.439 (0.316)	-0.405 (0.314)
Age five SDQ emotional – missing data		0.203 (0.668)	0.281 (0.649)	0.286 (0.638)	0.200 (0.648)
(Age five SDQ conduct – ‘normal’)		0 (.)	0 (.)	0 (.)	0 (.)
Age five SDQ conduct – ‘borderline’		-0.00717 (0.228)	0.00468 (0.225)	-0.0231 (0.227)	-0.00489 (0.235)
Age five SDQ conduct – ‘abnormal’		-0.506* (0.221)	-0.517* (0.211)	-0.565** (0.207)	-0.497* (0.211)
Age five SDQ conduct – missing data		-5.132** (1.644)	-5.532*** (1.646)	-5.853*** (1.652)	-6.149*** (1.747)
(Age five SDQ hyperactive – ‘normal’)		0 (.)	0 (.)	0 (.)	0 (.)
Age five SDQ hyperactive – ‘borderline’		0.137 (0.287)	0.161 (0.280)	0.164 (0.274)	0.146 (0.266)
Age five SDQ hyperactive – ‘abnormal’		-0.696** (0.265)	-0.641* (0.259)	-0.597* (0.259)	-0.593* (0.257)
Age five SDQ hyperactive – missing data		0.568 (0.382)	0.682^{+} (0.389)	0.630 (0.390)	0.650 (0.401)
(Age five SDQ peer – ‘normal’)		0 (.)	0 (.)	0 (.)	0 (.)
Age five SDQ peer – ‘borderline’		0.133 (0.241)	0.195 (0.234)	0.204 (0.233)	0.199 (0.229)

Age five SDQ peer – ‘abnormal’	0.212 (0.325)	0.260 (0.307)	0.278 (0.321)	0.201 (0.317)
Age five SDQ peer – missing data	1.252 (0.883)	1.426 (0.885)	1.474^{+} (0.880)	1.689^{+} (0.908)
(Age five SDQ pro-social – ‘normal’)	0 (.)	0 (.)	0 (.)	0 (.)
Age five SDQ pro-social – ‘borderline’	0.389 (0.271)	0.486^{+} (0.278)	0.497^{+} (0.283)	0.445 (0.283)
Age five SDQ pro-social – ‘abnormal’	0.899* (0.374)	0.899* (0.355)	0.893* (0.356)	0.913* (0.358)
Age five SDQ pro-social – missing data	2.267** (0.713)	2.227** (0.702)	2.629*** (0.748)	2.759*** (0.759)
Age seven SDQ emotional	-0.00853 (0.036)	0.0108 (0.035)	0.00875 (0.035)	0.00383 (0.035)
Age seven SDQ conduct	0.0826 (0.073)	0.0836 (0.071)	0.0698 (0.070)	0.0832 (0.072)
Age seven SDQ hyperactive	-0.0776^{+} (0.044)	-0.0824^{+} (0.044)	-0.0783^{+} (0.044)	-0.0690 (0.045)
Age seven SDQ peer	-0.0256 (0.059)	-0.0293 (0.056)	-0.0260 (0.056)	-0.0164 (0.055)
Age seven SDQ pro-social	-0.0175 (0.045)	-0.0237 (0.047)	-0.0237 (0.047)	-0.0176 (0.047)
(No behaviour difficulties)	0 (.)	0 (.)	0 (.)	0 (.)
Minor behaviour difficulties	0.0330 (0.236)	0.0557 (0.235)	0.0630 (0.237)	0.0522 (0.238)
Definite behaviour difficulties	0.110 (0.382)	0.150 (0.381)	0.250 (0.387)	0.223 (0.407)

Severe behaviour difficulties	-1.734^{+} (0.936)	-1.692^{+} (0.939)	-1.690^{+} (0.931)	-1.671^{+} (0.948)
(FSP score – bottom quintile)		0 (.)	0 (.)	0 (.)
FSP score – second quintile		0.161 (0.221)	0.135 (0.220)	0.130 (0.220)
FSP score – third quintile		0.378 (0.266)	0.306 (0.267)	0.281 (0.276)
FSP score – fourth quintile		0.388 (0.284)	0.312 (0.280)	0.308 (0.284)
FSP score – top quintile		0.857** (0.306)	0.793** (0.303)	0.814** (0.312)
FSP score – missing data		0.769* (0.348)	0.690* (0.342)	0.613^{+} (0.336)
Recognised SEN			-0.376^{+} (0.215)	-0.398^{+} (0.220)
(No SEN / do not know)			0 (.)	0 (.)
(Female teacher)				0 (.)
Male teacher				0.166 (0.347)
Teacher gender missing data				-0.0537 (0.396)
Teacher years taught: missing data				-0.573 (0.474)
Teacher years taught: 24-48 years				-0.0666 (0.372)
Teacher years taught: 14-23 years				-0.627 (0.392)

Teacher years taught: 8-13 years						-0.407 (0.338)
Teacher years taught: 4-7 years						-0.363 (0.306)
(Teacher years taught: 1-3 years – ref)						0 (.)
Teacher years at school: missing data						0.665^{+} (0.383)
Teacher years at school: 8-48 years						0.685* (0.323)
Teacher years at school: 4-7 years						0.294 (0.286)
(Teacher years at school: 1-3 years – ref)						0 (.)
Constant	15.99** (5.045)	16.11** (5.327)	18.72*** (5.198)	18.23*** (5.169)	18.44*** (5.049)	17.78*** (5.037)
N	639	639	635	635	635	635
***R*²**	0.799	0.809	0.825	0.829	0.830	0.833

Standard errors in parentheses. Reference category in brackets. Coefficients from linear regression model.

Ns are unweighted; coefficients are weighted for initial survey design and for attrition to the level of the main wave four survey.

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

^Outcome is KS1 Average Points Score; range: 3-22.5.

Annex 3D

Difference in survey-reported summed teacher judgment of ‘ability and attainment’ according to pupils’ stream placement: all covariates

Table 3D1: Difference in survey-reported summed teacher judgment of ‘ability and attainment’ according to pupils’ stream placement^

	Spec 1	**Spec 2**	**Spec 3**	**Spec 4**	**Spec 5**	**Spec 6**
Top stream	3.157***	2.874***	2.661***	2.586***	2.611***	2.569***
	(0.286)	(0.274)	(0.260)	(0.253)	(0.250)	(0.258)
(Middle stream)	0	0	0	0	0	0
	(.)	(.)	(.)	(.)	(.)	(.)
Bottom stream	-2.702***	-2.384***	-1.964***	-1.897***	-1.686***	-1.704***
	(0.327)	(0.328)	(0.318)	(0.299)	(0.289)	(0.280)
Maths Test score	0.0951***	0.0971***	0.0681**	0.0646**	0.0602**	0.0611**
	(0.023)	(0.024)	(0.021)	(0.021)	(0.021)	(0.021)
Word Reading Test score	0.0489***	0.0502***	0.0484***	0.0456***	0.0437***	0.0440***
	(0.005)	(0.005)	(0.004)	(0.004)	(0.004)	(0.004)
Pattern Construction Test score	0.0313***	0.0258***	0.0168*	0.0166*	0.0172*	0.0159*
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Age at tests	0.0409	-0.239**	-0.245**	-0.245**	-0.241**	-0.240**
	(0.063)	(0.086)	(0.080)	(0.079)	(0.078)	(0.077)
Age at teacher survey: missing data	0.193	-0.136	-0.324	-0.306	-0.202	-0.192
	(0.553)	(0.555)	(0.540)	(0.525)	(0.523)	(0.516)
82-87 months	0.405	0.295	0.177	0.157	0.243	0.280
	(0.554)	(0.501)	(0.508)	(0.498)	(0.494)	(0.499)
88-89 months	-0.0293	-0.433	-0.0565	-0.0379	-0.00995	0.0168
	(0.453)	(0.431)	(0.410)	(0.414)	(0.413)	(0.415)

90-91 months	0.470 (0.462)	0.251 (0.443)	0.191 (0.418)	0.148 (0.408)	0.166 (0.410)	0.191 (0.410)
92-93 months	0.533 (0.405)	0.175 (0.330)	0.0909 (0.343)	0.163 (0.342)	0.200 (0.343)	0.230 (0.350)
(94-104 months)	0 (.)	0 (.)	0 (.)	0 (.)	0 (.)	0 (.)
Boy		-0.423^{+} (0.225)	-0.146 (0.220)	-0.115 (0.221)	-0.0931 (0.222)	-0.0752 (0.219)
(Girl)		0 (.)	0 (.)	0 (.)	0 (.)	0 (.)
August-born		-2.899*** (0.800)	-2.925*** (0.759)	-2.563*** (0.736)	-2.525** (0.755)	-2.566*** (0.759)
July-born		-3.412*** (0.765)	-3.492*** (0.733)	-3.146*** (0.722)	-3.103*** (0.741)	-3.170*** (0.737)
June-born		-2.593*** (0.691)	-2.694*** (0.665)	-2.441*** (0.645)	-2.415*** (0.656)	-2.456*** (0.666)
May-born		-2.258*** (0.633)	-2.294*** (0.558)	-2.062*** (0.534)	-2.064*** (0.539)	-2.105*** (0.536)
April-born		-2.812*** (0.688)	-2.783*** (0.650)	-2.506*** (0.658)	-2.519*** (0.649)	-2.582*** (0.652)
March-born		-0.988^{+} (0.571)	-1.146* (0.545)	-0.984^{+} (0.528)	-1.004^{+} (0.532)	-1.032^{+} (0.531)
February-born		-1.041 (0.730)	-1.211^{+} (0.677)	-1.055 (0.674)	-1.098 (0.669)	-1.167^{+} (0.685)
January-born		-1.411* (0.620)	-1.498* (0.583)	-1.338* (0.570)	-1.309* (0.584)	-1.360* (0.592)
December-born		-0.993^{+} (0.567)	-1.206* (0.547)	-1.050* (0.530)	-1.040^{+} (0.544)	-1.020^{+} (0.528)
November-born		-0.878^{+} (0.521)	-0.938^{+} (0.499)	-0.880^{+} (0.493)	-0.884^{+} (0.509)	-0.873^{+} (0.496)
October-born		0.127 (0.504)	-0.269 (0.519)	-0.177 (0.505)	-0.148 (0.520)	-0.166 (0.513)

(September-born)	0 (.)	0 (.)	0 (.)	0 (.)	0 (.)
(White ethnicity)	0 (.)	0 (.)	0 (.)	0 (.)	0 (.)
Mixed / ‘other’ / missing data	0.00334 (0.502)	-0.161 (0.467)	-0.0579 (0.474)	-0.132 (0.475)	-0.118 (0.471)
Indian	0.257 (0.525)	0.442 (0.545)	0.555 (0.530)	0.447 (0.510)	0.513 (0.553)
Pakistani / Bangladeshi	-0.843* (0.362)	-0.730* (0.369)	-0.588 (0.390)	-0.709^{+} (0.383)	-0.654^{+} (0.389)
Black / Black British	-1.299^{+} (0.668)	-1.625** (0.590)	-1.577* (0.629)	-1.588* (0.639)	-1.463* (0.622)
(Higher-income)	0 (.)	0 (.)	0 (.)	0 (.)	0 (.)
Low-income	-0.463^{+} (0.277)	-0.427 (0.273)	-0.395 (0.262)	-0.364 (0.263)	-0.392 (0.265)
Parent level 1 qual	-0.564 (0.514)	-0.613 (0.497)	-0.563 (0.496)	-0.577 (0.504)	-0.614 (0.495)
Parent level 2 qual	-0.857^{+} (0.452)	-0.961* (0.449)	-0.982* (0.450)	-0.994* (0.453)	-0.977* (0.454)
Parent level 3 qual	-0.0611 (0.495)	-0.0986 (0.460)	-0.0707 (0.456)	-0.127 (0.459)	-0.0977 (0.462)
Parent level 4 qual	-0.292 (0.452)	-0.360 (0.442)	-0.390 (0.445)	-0.421 (0.449)	-0.431 (0.450)
(Parent level 5 qual – ref)	0 (.)	0 (.)	0 (.)	0 (.)	0 (.)
Parent overseas qual	0.453 (0.918)	0.386 (0.864)	0.435 (0.859)	0.456 (0.902)	0.437 (0.904)
Parent no qual	-1.410* (0.570)	-1.353* (0.556)	-1.240* (0.565)	-1.217* (0.567)	-1.205* (0.560)

(Age five SDQ emotional – ‘normal’)	0 (.)	0 (.)	0 (.)	0 (.)
Age five SDQ emotional – ‘borderline’	0.0528 (0.468)	0.114 (0.464)	0.0453 (0.468)	0.135 (0.484)
Age five SDQ emotional – ‘abnormal’	-0.194 (0.403)	-0.152 (0.394)	-0.171 (0.396)	-0.143 (0.408)
Age five SDQ emotional – missing data	1.779 (1.244)	1.773 (1.105)	1.720 (1.092)	1.675 (1.103)
(Age five SDQ conduct – ‘normal’)	0 (.)	0 (.)	0 (.)	0 (.)
Age five SDQ conduct – ‘borderline’	-0.245 (0.345)	-0.179 (0.328)	-0.187 (0.325)	-0.166 (0.333)
Age five SDQ conduct – ‘abnormal’	-0.259 (0.303)	-0.310 (0.296)	-0.383 (0.302)	-0.356 (0.306)
Age five SDQ conduct – missing data	-4.298* (2.042)	-4.846* (2.096)	-4.945* (2.140)	-5.153* (2.224)
(Age five SDQ hyperactive – ‘normal’)	0 (.)	0 (.)	0 (.)	0 (.)
Age five SDQ hyperactive – ‘borderline’	0.0568 (0.334)	0.0836 (0.337)	0.0471 (0.328)	0.0493 (0.328)
Age five SDQ hyperactive – ‘abnormal’	-0.241 (0.344)	-0.0994 (0.344)	-0.0518 (0.350)	-0.0373 (0.357)
Age five SDQ hyperactive – missing data	1.245 (0.859)	1.476^{+} (0.817)	1.326 (0.852)	1.418 (0.884)
(Age five SDQ peer – ‘normal’)	0 (.)	0 (.)	0 (.)	0 (.)
Age five SDQ peer – ‘borderline’	-0.498 (0.363)	-0.433 (0.364)	-0.440 (0.360)	-0.516 (0.377)
Age five SDQ peer – ‘abnormal’	-0.310 (0.331)	-0.233 (0.337)	-0.155 (0.338)	-0.210 (0.349)

Age five SDQ peer – missing data	3.953^{+}	4.072^{+}	3.912^{+}	4.253^{+}
	(2.043)	(2.170)	(2.149)	(2.291)
(Age five SDQ pro-social – ‘normal’)	0	0	0	0
	(.)	(.)	(.)	(.)
Age five SDQ pro-social – ‘borderline’	0.395	0.544	0.496	0.536
	(0.424)	(0.405)	(0.401)	(0.399)
Age five SDQ pro-social – ‘abnormal’	0.157	0.116	0.187	0.142
	(0.528)	(0.490)	(0.498)	(0.511)
Age five SDQ pro-social – missing data	-3.060**	-3.175**	-2.601*	-2.800*
	(1.021)	(1.035)	(1.150)	(1.203)
Age seven SDQ emotional	-0.0774	-0.0672	-0.0720	-0.0749
	(0.049)	(0.049)	(0.049)	(0.051)
Age seven SDQ conduct	0.271**	0.270**	0.258**	0.264**
	(0.095)	(0.096)	(0.095)	(0.096)
Age seven SDQ hyperactive	-0.154**	-0.164**	-0.157**	-0.159**
	(0.057)	(0.056)	(0.055)	(0.055)
Age seven SDQ peer	-0.149^{+}	-0.156^{+}	-0.148^{+}	-0.145^{+}
	(0.086)	(0.086)	(0.085)	(0.087)
Age seven SDQ pro-social	0.148**	0.150**	0.154**	0.151**
	(0.054)	(0.055)	(0.054)	(0.056)
(No behaviour difficulties)	0	0	0	0
	(.)	(.)	(.)	(.)
Minor behaviour difficulties	0.0112	0.0755	0.130	0.151
	(0.284)	(0.287)	(0.288)	(0.292)
Definite behaviour difficulties	-0.742	-0.711	-0.523	-0.522
	(0.516)	(0.519)	(0.533)	(0.536)
Severe behaviour difficulties	-2.608**	-2.510**	-2.390**	-2.426**
	(0.929)	(0.919)	(0.912)	(0.912)

(FSP score – bottom quintile)	0 (.)	0 (.)	0 (.)
FSP score – second quintile	0.452 (0.358)	0.407 (0.346)	0.469 (0.346)
FSP score – third quintile	0.698^{+} (0.387)	0.630^{+} (0.377)	0.700^{+} (0.382)
FSP score – fourth quintile	0.328 (0.418)	0.216 (0.410)	0.285 (0.416)
FSP score – top quintile	1.240^{*} (0.495)	1.155^{*} (0.490)	1.208^{*} (0.491)
FSP score – missing data	0.959^{+} (0.495)	0.803^{+} (0.484)	0.866^{+} (0.474)
Recognised SEN		-0.709^{*} (0.356)	-0.678^{+} (0.350)
(No SEN / do not know)		0 (.)	0 (.)
(Female teacher)			0 (.)
Male teacher			0.286 (0.466)
Teacher gender missing data			0.348 (0.520)
Teacher years taught: missing data			-0.387 (0.577)
Teacher years taught: 24-48 years			-0.196 (0.628)
Teacher years taught: 14-23 years			-0.120 (0.556)
Teacher years taught: 8-13 years			-0.340 (0.450)
Teacher years taught: 4-7 years			-0.603 (0.462)

(Teacher years taught: 1-3 years – ref)						0 (.)
Teacher years at school: missing data						0.189 (0.463)
Teacher years at school: 8-48 years						0.277 (0.449)
Teacher years at school: 4-7 years						0.613 (0.382)
(Teacher years at school: 1-3 years – ref)						0 (.)
Constant	6.932 (5.809)	34.41*** (7.845)	36.48*** (7.509)	36.02*** (7.417)	35.91*** (7.317)	35.84*** (7.194)
N	829	829	823	823	823	823
R^2	0.703	0.737	0.769	0.773	0.775	0.776

Standard errors in parentheses. Reference category in brackets. Coefficients from linear regression model.
Ns are unweighted; coefficients are weighted for initial survey design and for attrition to the level of the main wave four survey.
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
^Outcome is summed teacher survey-reported judgment; range: 7-35

Annex 3E

Difference in survey-reported summed teacher judgment of academic domain ‘ability and attainment’ according to pupils’ stream placement

Table 3E1: Difference in survey-reported summed teacher judgment of academic domain ‘ability and attainment’ according to pupils’ stream placement^ ^^

	Spec 1	Spec 2	Spec 3	Spec 4	Spec 5	Spec 6
Top stream	2.779*** (0.216)	2.519*** (0.208)	2.406*** (0.211)	2.327*** (0.207)	2.347*** (0.205)	2.308*** (0.209)
(Middle stream)	0 (.)	0 (.)	0 (.)	0 (.)	0 (.)	0 (.)
Bottom stream	-2.304*** (0.253)	-2.103*** (0.253)	-1.859*** (0.249)	-1.806*** (0.235)	-1.633*** (0.237)	-1.640*** (0.229)
Maths test score	0.0728*** (0.016)	0.0735*** (0.017)	0.0585*** (0.017)	0.0551*** (0.016)	0.0516** (0.016)	0.0507** (0.016)
Word Reading Test score	0.0439*** (0.004)	0.0460*** (0.004)	0.0447*** (0.003)	0.0419*** (0.004)	0.0404*** (0.004)	0.0408*** (0.004)
Pattern Construction Ability test score	0.0215*** (0.005)	0.0168** (0.006)	0.0119* (0.006)	0.0115* (0.006)	0.0120* (0.006)	0.0108^{+} (0.006)
Constant	4.915 (4.354)	26.75*** (5.837)	28.09*** (5.807)	27.51*** (5.734)	27.42*** (5.647)	27.39*** (5.568)
N	836	836	830	830	830	830
***R*^{2}**	0.746	0.773	0.789	0.793	0.795	0.798

Standard errors in parentheses. Reference category in brackets. Coefficients from linear regression model.
Ns are unweighted; coefficients are weighted for initial survey design and for attrition to the level of the main wave four survey.
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
^Outcome is summed teacher survey-reported judgment; range: 5-25
^^Specification one controls for age at tests and age at teacher survey, specification two adds pupil gender, pupil month of birth, pupil ethnicity, family income level, main parent’s highest qualification; specification three adds age five parent-assessed SDQ, age seven teacher assessed SDQ, age seven teacher judgment of pupil’s behaviour; specification four adds Foundation Stage Profile score; specification five adds pupil special educational needs diagnosis; specification six adds teacher gender, teacher years teaching, teacher years teaching at this school.

Annex 3F

Difference in survey-reported summed teacher judgment of ‘ability and attainment’ according to pupils’ stream placement: unweighted, clustered estimates

Table 3F1: Difference in survey-reported summed teacher judgment of ‘ability and attainment’ according to pupils’ stream placement: unweighted coefficients, standard errors clustered at school-level: specification six^ ^^

	Original	**No weights, clustering**
Top stream	2.569***	2.706***
	(0.258)	(0.269)
(Middle stream)	-	-
Bottom stream	-1.704***	-1.791***
	(0.280)	(0.318)
Maths test score	0.0611**	0.057**
	(0.021)	(.022)
Word Reading Test score	0.0440***	0.044***
	(0.004)	(0.005)
Pattern Construction Ability test score	0.0159* (0.007)	0.015* (0.007)
Constant	35.84***	33.31***
	(7.194)	(6.838)
N	**823**	**823**
R^2	0.776	0.768

Standard errors in parentheses. Reference category in brackets. Coefficients from linear regression model.

Ns are unweighted; coefficients are unweighted, standard errors are clustered at the school-level.

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

^Outcome is summed teacher survey-reported judgment; range: 7-35

^^Specification one controls for age at tests and age at teacher survey, specification two adds pupil gender, pupil month of birth, pupil ethnicity, family income level, main parent’s highest qualification; specification three adds age five parent-assessed SDQ, age seven teacher assessed SDQ, age seven teacher judgment of pupil’s behaviour; specification four adds Foundation Stage Profile score; specification five adds pupil special educational needs diagnosis; specification six adds teacher gender, teacher years teaching, teacher years teaching at this school.

Annex 3G

Difference in teacher-assessed Key Stage One / survey-assessed reading / maths level according to pupils' stream placement: linear models

Table 3G1: Difference in teacher-assessed Key Stage One reading / maths level according to pupils' stream placement: linear models (specification six): linear models^ ^^

	Reading level	Maths level
Top stream	0.345*** (.083)	0.187* (.089)
(Middle stream)	-	-
Bottom stream	-0.336** (.110)	-0.547*** (.115)
Maths test score	-0.002 (.006)	0.038*** (.007)
Word Reading Test score	0.022*** (.002)	0.009*** (.002)
Pattern Construction test score	.004 (.002)	.009*** (.002)
Constant	4.048+ (2.127)	5.095* (1.987)
N	437	460

Standard errors in parentheses. Reference category in brackets. Coefficients from linear regression models.

Ns are unweighted; coefficients are weighted for initial survey design and for attrition to the level of the main wave four survey.

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

^Outcome is KS1 reading / maths level: 'working towards level 1' / achieved level 1' / 'achieved level 2c' / 'achieved level 2b' / 'achieved level 2a.'

^^Controlled for age at tests, month of birth, pupil gender, pupil ethnicity, family income-level, main parent's highest qualification, school type, pupil's length of time attending school; age five parent-assessed SDQ, age seven teacher assessed SDQ, age seven teacher judgment of pupil's behaviour; Foundation Stage Profile score; pupil special educational needs diagnosis; teacher gender, teacher years teaching, teacher years teaching at this school.

Table 3G2: **Differences in survey-reported teacher judgements of level of reading / maths 'ability and attainment' according to pupils' stream placement (specification six): linear models^ ^^**

	Reading level	Maths level
Top stream	0.555*** (.063)	0.536*** (.062)
(Middle stream)	-	-
Bottom stream	-0.367*** (.072)	-.0476*** (0.078)
Maths test score	0.003 (.005)	0.023*** (.005)
Word Reading Test score	.015*** (.001)	0.005*** (.001)
Pattern Construction test score	0.001 (.001)	0.005** (.0.002)
Constant	6.00*** (1.304)	6.406*** (1.620)
N	**843**	**839**

Standard errors in parentheses. Reference category in brackets. Coefficients from linear regression models.

Ns are unweighted; coefficients are weighted for initial survey design and for attrition to the level of the main wave four survey.

$^{+}$ p < 0.10, * p < 0.05, ** p < 0.01, *** p < 0.001

^ Outcomes are survey-reported teacher judgements of reading / maths ability; range: 1-5

^^Controlled for age at tests, month of birth, pupil gender, pupil ethnicity, family income-level, main parent's highest qualification, school type, pupil's length of time attending school; age five parent-assessed SDQ, age seven teacher assessed SDQ, age seven teacher judgment of pupil's behaviour; Foundation Stage Profile score; pupil special educational needs diagnosis; teacher gender, teacher years teaching, teacher years teaching at this school.

Annex 4A

Key descriptive statistics for respective Millennium Cohort Study English single-cohort baby household samples

Table 4A1: Key descriptive statistics for respective Millennium Cohort Study English single-cohort baby household samples, for comparison*

	Wave one: whole English sample, design weights only	Wave one: whole English sample, design weights plus non-response weights	Wave four: English sample with teacher survey response, design weights only	Wave four: English sample without teacher survey response, design weights only	Wave four sample used in paper for reading analysis, design weights only	Wave four: English sample with teacher survey response, design weights plus attrition weights	Wave four: English sample without teacher survey response, design weights plus attrition weights	Wave four sample used in paper for reading analysis, design weights plus attrition weights
Percent low-income (OECD indicator) – at wave one	28.2	29.5	23.6	25.4	23.4	30.3	33.4	30.0
Percent low-income (OECD indicator) – at wave four	-	-	21.3	24.0	21.8	27.0	30.6	27.4
Percent girls	48.8	48.9	50.1	48.5	50.3	49.6	47.5	50.0
Percent White	85.3	84.6	90.9	85.3	89.6	88.0	80.8	86.4
Percent Indian	2.1	2.1	1.7	2.7	1.6	2.0	3.0	1.8
Percent Pakistani	3.4	3.6	2.6	4.0	2.5	3.5	5.4	3.4
Percent Bangladeshi	1.1	1.2	0.7	1.5	0.6	1.0	2.2	0.9
Percent Black Caribbean	1.1	1.2	0.8	1.6	0.8	1.2	2.1	1.1
Percent Black African	1.7	1.8	1.1	2.1	1.1	1.7	2.8	1.7

Percent speaking English only	88.9	88.4	91.8	87.6	93.1	89.8	84.0	91.1
Mean Word Reading Test score	-	-	109.2	108.2	110.2	107.3	105.4	108.5
Mean Progress in Maths test score	**-**	**-**	18.8	18.4	18.8	18.5	17.9	18.4
n =	**11374**	**11374**	**5184**	**3107**	**4997**	**5184**	**3107**	**4997**

*Figures are presented firstly with design weights only, which account simply for known unequal selection probabilities into the initial sample (children in areas with higher number of minority ethnic and low-income families were oversampled so had a higher probability of inclusion). Secondly, adjustments for non-response (at wave one) / attrition (at wave four) are presented – these weight the sample according to differential tendencies to participation according to selected measured characteristics. Ns are unweighted.

Annex 4B

Robustness check: Excluding cases with data missing on time lag

Table 4B1: Difference in likelihood of pupils with each respective characteristic being judged 'above average' at reading by their teacher, compared to pupils with the reference characteristic, controlling for reading cognitive test score - Robustness check 2: Excluding cases with data missing on time lag**, with full coefficients^**

	Original results (*B*)	**Check 2: Age and timing controls (*B*)^^^**
Low-income (ref = higher)	-.11 (.014)***	-.11 (.015)***
Age in months^^		.00 (.002)*
2 month lag (ref = 0-1 month)		.01 (.030)
3 month lag (ref = 0-1 month)		.04 (.033)
4 month lag (ref = 0-1 month)		.02 (.033)
5 month lag (ref = 0-1 month)		-.04 (.036)
6 month lag (ref = 0-1 month)		-.04 (.036)
7 month lag (ref = 0-1 month)		-.05 (.039)
8-20 month lag (ref = 0-1 month)		-.06 (.045)
Boy (ref = girl)	-.04 (.013)**	-.04 (.013)**
Age in months^^		.00 (.002)
2 month lag (ref = 0-1 month)		.00 (.030)
3 month lag (ref = 0-1 month)		.04 (.034)
4 month lag (ref = 0-1 month)		.01 (.033)
5 month lag (ref = 0-1 month)		-.04 (.036)
6 month lag (ref = 0-1 month)		-.05 (.037)
7 month lag (ref = 0-1 month)		-.05 (.038)
8-20 month lag (ref = 0-1 month)		-.06 (.045)
SEN (ref = no SEN)	-.12 (.015)***	-.11 (.017)***
Age in months^^		.00 (.002)
2 month lag (ref = 0-1 month)		.01 (.030)
3 month lag (ref = 0-1 month)		.04 (.033)
4 month lag (ref = 0-1 month)		.01 (.033)

5 month lag (ref = 0-1 month)		-.04 (.035)
6 month lag (ref = 0-1 month)		-.05 (.037)
7 month lag (ref = 0-1 month)		-.05 (.038)
8-20 month lag (ref = 0-1 month)		-.06 (.044)
Indian (ref = White)	-.09 (.046)*	-.11 (.046)**
Pakistani (ref = White)	-.17 (.027)***	-.18 (.027)***
Bangladeshi (ref = White)	-.15 (.058)**	-.17 (.057)**
Black Caribbean (ref = White)	-.11 (.039)**	-.12 (.043)**
Black African (ref = White)	-.14 (.056)**	-.13 (.059)**
Age in months^^		.00 (.002)*
2 month lag (ref = 0-1 month)		.00 (.031)
3 month lag (ref = 0-1 month)		.04 (.033)
4 month lag (ref = 0-1 month)		.02 (.033)
5 month lag (ref = 0-1 month)		-.04 (.036)
6 month lag (ref = 0-1 month)		-.05 (.036)
7 month lag (ref = 0-1 month)		-.05 (.039)
8-20 month lag (ref = 0-1 month)		-.05 (.045)
Other languages (ref = English only)	-.12 (.021)***	-.13 (.020)***
Age in months^^		.00 (.002)
2 month lag (ref = 0-1 month)		.00 (.030)
3 month lag (ref = 0-1 month)		.04 (.033)
4 month lag (ref = 0-1 month)		.02 (.033)
5 month lag (ref = 0-1 month)		-.04 (.036)
6 month lag (ref = 0-1 month)		-.05 (.036)
7 month lag (ref = 0-1 month)		-.05 (.039)
8-20 month lag (ref = 0-1 month)		-.05 (.044)
All n.s =	4997	4641

*** = p < .001; ** = p < .05; * = p < .10. Standard errors in brackets.

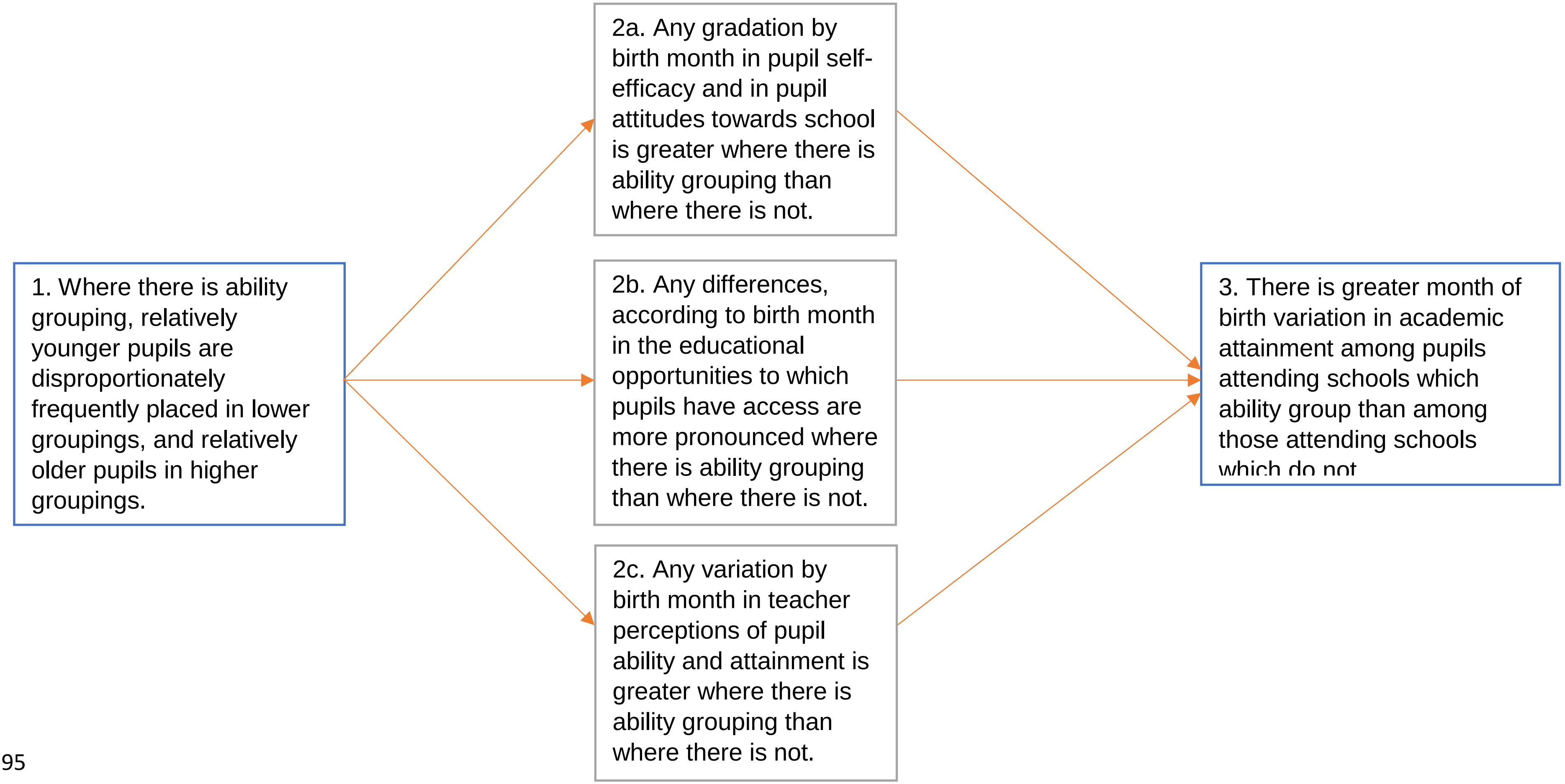
^All estimates weighted for survey design and for attrition to the main wave four survey. Ns are unweighted.

^^Range = 76-97

Annex 5A

Potential channels from in-school ability grouping to variation in attainment by birth month

Figure 5A1: Premises for potential channels through which in-school ability grouping may lead to month of birth attainment variation (from Campbell, 2013a)



1. Where there is ability grouping, relatively younger pupils are disproportionately frequently placed in lower groupings, and relatively older pupils in higher groupings.

2a. Any gradation by birth month in pupil self-efficacy and in pupil attitudes towards school is greater where there is ability grouping than where there is not.

2b. Any differences, according to birth month in the educational opportunities to which pupils have access are more pronounced where there is ability grouping than where there is not.

2c. Any variation by birth month in teacher perceptions of pupil ability and attainment is greater where there is ability grouping than where there is not.

3. There is greater month of birth variation in academic attainment among pupils attending schools which ability group than among those attending schools which do not

Annex 5B

Percentage pupils judged to be at each level of ability and attainment

Table 5B1: Percentage in whole teacher survey sample judged to be at each level of *ability and attainment* in each subject domain

	Speaking and listening (n = 5429)	Reading (n = 5426)	Writing (n = 5426)	Science (n = 5423)	Maths (n = 5412)	PE (n = 5429)	ICT (n = 5418)	Arts (n = 5425)
Well above average	9.3	12.8	6.7	5.9	9.3	4.2	2.8	4.0
Above average	30.4	33.6	25.7	28.6	31.5	23.7	23.4	22.3
Average	43.7	32.1	38.0	51.5	38.9	63.0	62.2	60.7
Below average	13.4	15.9	22.9	11.3	16.1	7.5	9.7	11.1
Well below average	3.3	5.6	6.8	2.7	4.2	1.6	1.8	1.8

Annex 5C

Details of variables used at each stage of regression analysis

Table 5C1: Details of variables used at each stage of regression analysis

Intends to measure…	**Original variable name in MCS dataset**	**Whether recoded**	**Response possibilities (in original variable or in recoded variable if applicable). Reference categories in bold. Proportion in each category, or 25th / 50th / 75th percentiles, in brackets***
Dependent variables: Whether at age seven teacher assessment of 'ability and attainment' in given category is *above average* / *well above average*			
Speaking and listening	DQ2160	Yes	Above average (39.6%) / **average or below average** (60.4%)
Reading	DQ2162	Yes	Above average (46.4%) / **average or below average** (53.6%)
Writing	DQ2164	Yes	Above average (32.3%) / **average or below average** (67.7%)
Science	DQ2166	Yes	Above average (34.5%) / **average or below average** (65.5%)
Key predictors			
Child's season of birth	dhcdbma0	Yes	Summer (23.2%) / Spring (24.3%) / Winter (25.2%) / **Autumn** (27.3%)
Whether in-class ability-grouped or not	DQ2466	Yes	Yes (78.8%) / **No** (21.2%)
Stage two controls: pupil and family characteristics			
Pupil gender	dhcsexa0	No	Male (50.2%) / **female** (49.8%)
Pupil ethnicity	ddc06ea0	Yes	**White** (80%) / mixed (3.4%) / Indian (3.3%) / Pakistani and Bangladeshi (6.9%) / Black or Black British (3.9%) / Other or missing (2.6%)
BAS Naming Vocabulary T-score at age 5	cdnvtscr	No	20 – 80 (48, 56, 62)
BAS Pattern Construction T-score at age 5	cdpctscr	No	20 – 80 (46, 51, 57)
BAS Picture Similarities T-score at age 5	cdpstscr	No	20 – 80 (49, 55, 61)
Family income level when child is age 7	doedp000	Yes	**Above 60% median level** (72.7%) / below 60% or missing data (27.3%)

Family housing tenure when child is age 7	ddroow00	Yes	**Own with mortgage or loan** (60.7%) / rent (30.7%) / other (8.6%)
Whether English is spoken as an additional language in child's household at age 7	ddhlan00	Yes	**English only or missing** (86.3%) / Mostly English (5%) / Half English and half other language (4.6%) / Mostly or only other language (4.1%)
Main parent's highest academic qualification when pupil was born	amacqu00	Yes	Higher degree (3.5%) / First degree (14.4%) / Dip HE (9%) / A or AS level (8.8%) / **O level or GCSE A-C** (32.1%) / GCSE D-G (10.5%) / Other academic inc overseas (2.6%) / None, or missing data (19.2%)
Main parent's highest vocational qualification when pupil was born	amvcqu00	Yes	Professional at degree level (12.3) / Nursing or other medical (4.6%) / NVQ 3 (9.9%) / NVQ 2 (9.3%) / NVQ 1 (7.6%) / Other (6.9%) / **None, or missing data** (49.5%)
Whether single parent when child was born	adhtys00	Yes	One parent resident (12%) / **Two parents resident** (88%)
Whether internet available in home at age 7	dminlna0	Yes	No or missing data (16.8%) / **Yes** (83.2%)
Whether / length of time for which breastfed	ambfeaa0	Yes	Less than a week (11%) / Some weeks (16.6%) / Some months (28.6%) / Still breastfeeding at wave one interview (13.8%) / **Did not try breastfeeding, or baby would not breastfeed** (30%)
Stage three controls: School and respondent teacher characteristics			
Whether this is the same school as attended at Wave 3	dmsamsa	Yes	No, don't know, not applicable (15.5%) / **Yes** (84.5%)
Whether child is in Year Two	dmstsca0	Yes	No, in different year (5.9%) / **Yes, in year 2** (94.1%)
Whether parent reports paying fees for the school	dmsctya0	Yes	Yes (4.8%) / **No** (95.2%)
Whether family displayed religiosity for school admission	dmfthsa0	Yes	Yes (28.3%) / **No, not a faith school, or missing data** (71.7%)
Whether pupil's class contains mixed year groups	DQ2513	Yes	Yes (14%) / **No** (46.5%) / Question non-response (39.5%)

Number of children in class	DQ2511	Yes	1-25 (21.6%) / 26-29 (19.9%) / **30** (14.9%) / 31+ (3.2%) / Question non-response (40.4%)
Number of classes in pupil's year	DQ2524	Yes	**One** (21.2%) / Two (24.5%) / Three or more (13.8%) / Question non-response (40.5%)
Teacher gender	DQ2479	Yes	Male (4.1%) / **Female** (46.1%) / Question non-response (39.7%)
Number of years teacher has taught	DQ2481	Yes	**1-3** (11.2%) / 4-7 (12.4%) / 8-13 (10.6%) / 14-23 (11.6%) / 24-48 (11.1%) / Question non-response (43.1%)
Number of years teacher has taught at this school	DQ2487	Yes	**1-3** (19.9%) / 4-7 (16.7%) / 8-48 (20.4%) / Question non-response (43%)
Stage four controls: previous school / teacher assessments of pupil			
Foundation Stage Profile: total score – at age 5	FSPTOTAL	No	0 – 117 (77, 91, 102)
Whether teacher reports that child has any SEN at age 7	DQ2328	Yes	Yes (22.6%) / **No or missing data** (77.4%)

Annex 5D

Alternative analysis: relationships between month of birth / ability grouping and probability of being judged ‘below average’ by teacher

Table 5D1: Key coefficients at stage four of analysis for relationships between month of birth / ability grouping and probability of being judged ‘below average’ by teacher

Speaking and listening	
Autumn (ref: summer)	.000 (.029)
Winter (ref: summer)	.012 (.029)
Spring (ref: summer)	-.010 (.029)
Ability grouped (ref: not grouped)	.009 (.023)
Autumn x ability grouped	-.045 (.032)
N.	4531
Reading	
Autumn (ref: summer)	-.046 (.030)
Winter (ref: summer)	-.016 (.030)
Spring (ref: summer)	-.052 (.030)
Ability grouped (ref: not grouped)	.000 (.024)
Autumn x ability grouped	.012 (.033)
N.	4530
Writing	
Autumn (ref: summer)	-.041 (.034)
Winter (ref: summer)	.035 (.034)
Spring (ref: summer)	-.043 (.033)
Ability grouped (ref: not grouped)	.008 (.027)
Autumn x ability grouped	-.023 (.037)
N.	4530
Science	
Autumn (ref: summer)	-.038 (.027)
Winter (ref: summer)	-.026 (.027)
Spring (ref: summer)	-.048 (.027)

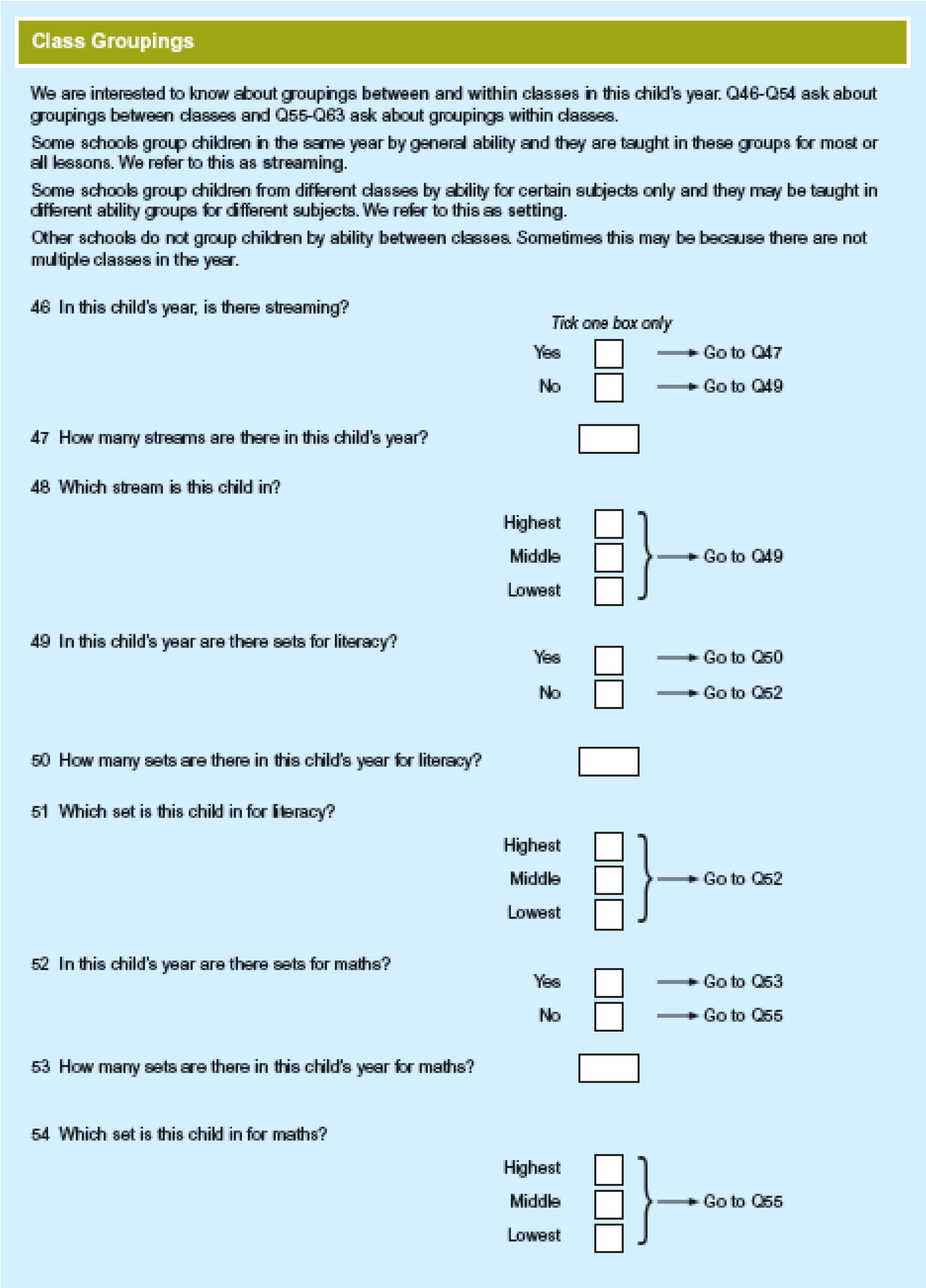
Ability grouped (ref: not grouped)	-.003 (.021)
Autumn x ability grouped	-.001 (.030)
N.	4526

*** = p < .001; ** = p < .01; * = p < .05. Standard errors in brackets.

Each coefficient indicates percentage change in predicted probability of being judged ‘below average.’ Controlled for pupil and family characteristics; school and teacher factors; pupil FSP score / presence of SEN diagnosis.

Annex I

Excerpt from Millennium Cohort Study wave four teacher questionnaire: Ability grouping questions



Class Groupings

We are interested to know about groupings between and within classes in this child's year. Q46-Q54 ask about groupings between classes and Q55-Q63 ask about groupings within classes.

Some schools group children in the same year by general ability and they are taught in these groups for most or all lessons. We refer to this as streaming.

Some schools group children from different classes by ability for certain subjects only and they may be taught in different ability groups for different subjects. We refer to this as setting.

Other schools do not group children by ability between classes. Sometimes this may be because there are not multiple classes in the year.

46 In this child's year, is there streaming?

Tick one box only

Yes □ → Go to Q47
No □ → Go to Q49

47 How many streams are there in this child's year? □

48 Which stream is this child in?

Highest □
Middle □
Lowest □
→ Go to Q49

49 In this child's year are there sets for literacy?

Yes □ → Go to Q50
No □ → Go to Q52

50 How many sets are there in this child's year for literacy? □

51 Which set is this child in for literacy?

Highest □
Middle □
Lowest □
→ Go to Q52

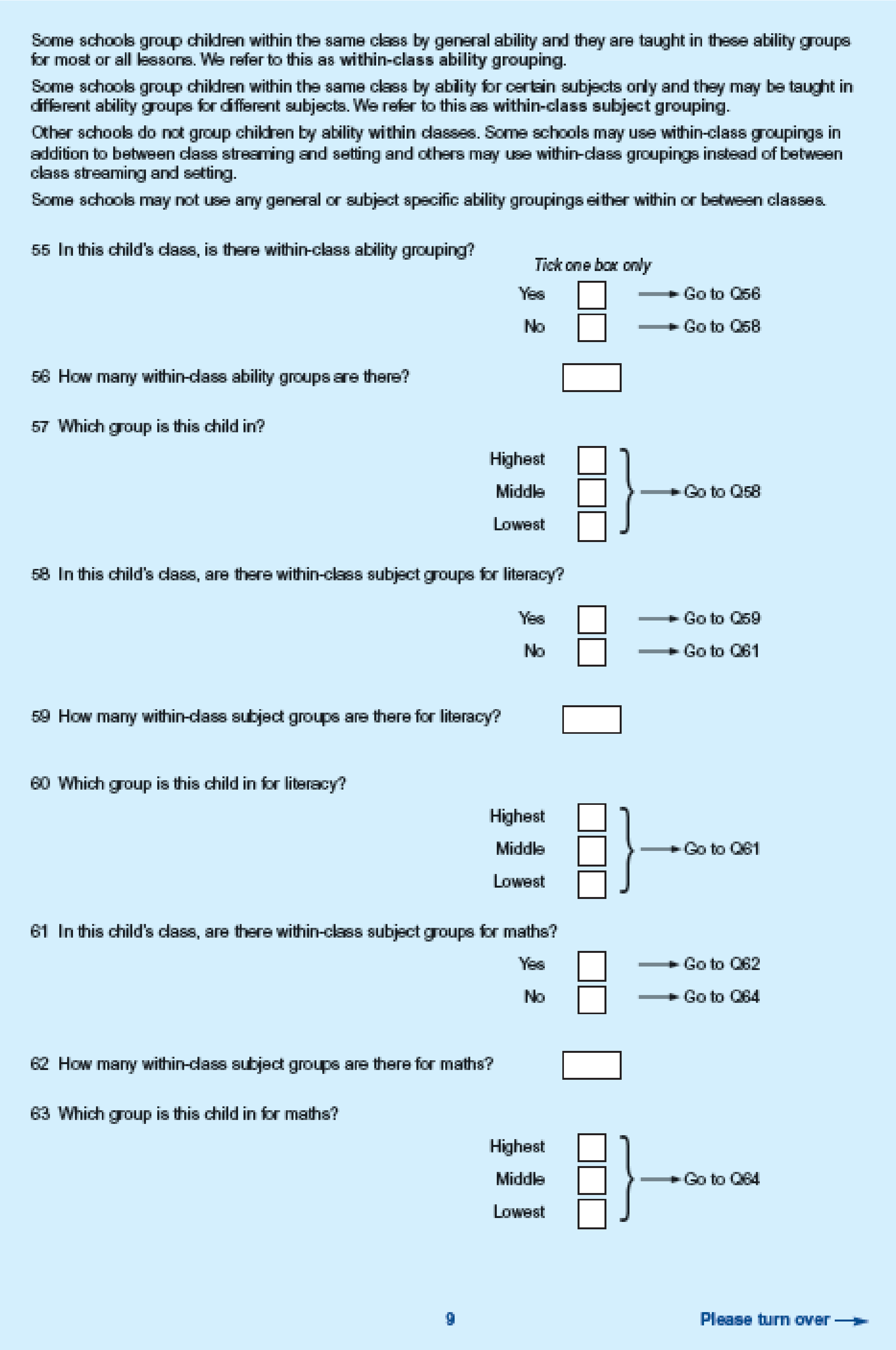
52 In this child's year are there sets for maths?

Yes □ → Go to Q53
No □ → Go to Q55

53 How many sets are there in this child's year for maths? □

54 Which set is this child in for maths?

Highest □
Middle □
Lowest □
→ Go to Q55



Some schools group children within the same class by general ability and they are taught in these ability groups for most or all lessons. We refer to this as **within-class ability grouping.**

Some schools group children within the same class by ability for certain subjects only and they may be taught in different ability groups for different subjects. We refer to this as **within-class subject grouping.**

Other schools do not group children by ability within classes. Some schools may use within-class groupings in addition to between class streaming and setting and others may use within-class groupings instead of between class streaming and setting.

Some schools may not use any general or subject specific ability groupings either within or between classes.

55 In this child's class, is there within-class ability grouping?

Tick one box only

Yes □ → Go to Q56

No □ → Go to Q58

56 How many within-class ability groups are there? □

57 Which group is this child in?

Highest □
Middle □
Lowest □
} → Go to Q58

58 In this child's class, are there within-class subject groups for literacy?

Yes □ → Go to Q59

No □ → Go to Q61

59 How many within-class subject groups are there for literacy? □

60 Which group is this child in for literacy?

Highest □
Middle □
Lowest □
} → Go to Q61

61 In this child's class, are there within-class subject groups for maths?

Yes □ → Go to Q62

No □ → Go to Q64

62 How many within-class subject groups are there for maths? □

63 Which group is this child in for maths?

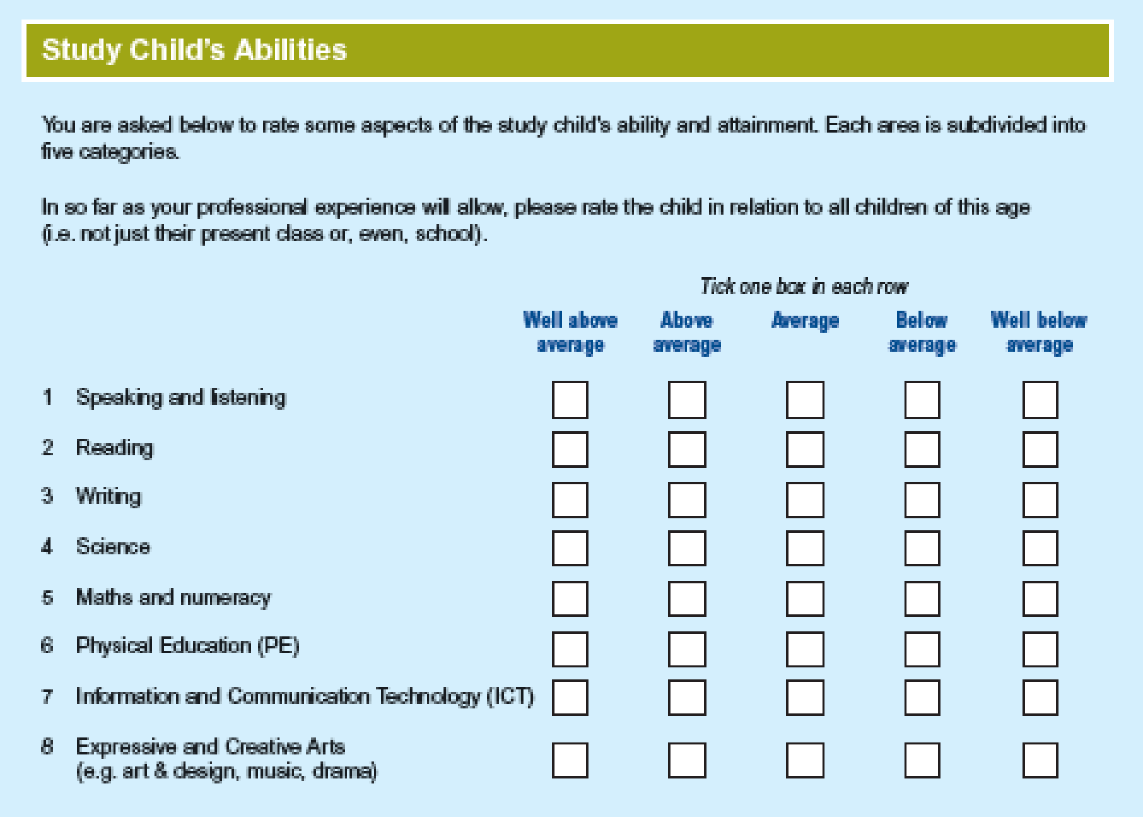
Highest □
Middle □
Lowest □
} → Go to Q64

9

Please turn over →

Annex II

Excerpt from Millennium Cohort Study wave four teacher questionnaire: ‘Ability and attainment’ questions



Study Child's Abilities

You are asked below to rate some aspects of the study child's ability and attainment. Each area is subdivided into five categories.

In so far as your professional experience will allow, please rate the child in relation to all children of this age (i.e. not just their present class or, even, school).

Tick one box in each row

		Well above average	Above average	Average	Below average	Well below average
1	Speaking and listening	□	□	□	□	□
2	Reading	□	□	□	□	□
3	Writing	□	□	□	□	□
4	Science	□	□	□	□	□
5	Maths and numeracy	□	□	□	□	□
6	Physical Education (PE)	□	□	□	□	□
7	Information and Communication Technology (ICT)	□	□	□	□	□
8	Expressive and Creative Arts (e.g. art & design, music, drama)	□	□	□	□	□