Leading education and social research
Institute of Education
University of London

Selected at seven: The relationship between teachers' judgments and assessments of pupils, and pupils' stream placements

Tammy Campbell

Department of Quantitative Social Science

Working Paper No. 14-10
June 2014

Disclaimer

Any opinions expressed here are those of the author(s) and not those of the Institute of Education. Research published in this series may include views on policy, but the institute itself takes no institutional policy positions.

DoQSS Workings Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

Department of Quantitative Social Science, Institute of Education, University of London
20 Bedford Way, London WC1H 0AL, UK

Selected at seven: The relationship between teachers' judgments and assessments of pupils, and pupils' stream placements

Tammy Campbell[1]

Abstract

Streaming (grouping pupils according to a measure or conception of overall ability for most / all teaching) has greatly increased in prevalence among English primary school children since the turn of the century. Evidence indicates that streaming may disadvantage children in lower groups and increase the overall attainment gap, and this paper explores one possible mechanism through which disparities might manifest: stream-dependent teacher perceptions. Using data for over 800 seven-year-old children who are taking part in the Millennium Cohort Study, analysis investigates whether teachers' survey-reported judgements and Key Stage One assessments of children correspond to the stream in which a child is placed. Regression modelling controls for potential confounding factors including: cognitive test performance; pupil gender, ethnicity, and month of birth; parents' income and education levels; parent and teacher perceptions of children's behaviour; prior in-school judgments / attainment; special educational need diagnosis; teacher characteristics; and school-type. Both survey-reported judgements of pupils and teacher-assessed Key Stage One assessments are found to be significantly related to children's stream placement. Children in the top stream are judged to be at a higher level and children in the bottom stream at a lower level than equivalent peers. It seems therefore that streaming may indeed contribute to attainment gaps through the medium of teacher perceptions and assessments, both by advantaging pupils in higher groups and penalising children in lower placements. This suggests a need to recognise, review and potentially revise the growing use of streaming among young children.

JEL classification: I24, I28

Keywords: primary education, streaming, inequality, perceptions, judgements, assessments

[1] Department of Quantitative Social Science, Institute of Education, University of London (tcampbell@ioe.ac.uk)

Acknowledgements

This paper presents analysis forming part of my PhD, which I am very grateful to the Economic and Social Research Council for funding. It uses data from the Millennium Cohort Study. I am grateful to the Centre for Longitudinal Studies, 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. However, these organisations bear no responsibility for the analysis or interpretation of these data. Many thanks to Lorraine Dearden for useful comments and suggestions.

Non-technical summary

Streaming is the practice of grouping and teaching pupils according to a measure or conception of 'overall ability' for most or all lessons. It was common in the early 20th century, but became extremely rare by the 1990s. However, in the past two decades it has remerged – including among very young children. The most recent estimates (for 2008) indicated that around 18% of seven-year-olds were streamed (while many were also set in-class or for individual subjects) – and, if the trend has continued, this proportion may now be even higher.

Research suggests that streaming and stream placement may affect pupils' academic attainment though various channels, including: the opportunities and teaching offered to children in different streams; children's own self-concept, motivation and attitudes; and teachers' perceptions and assessments of pupils placed at varying levels.

This paper focusses on the last of these possibilities, exploring whether teachers' judgments of pupils are affected by the stream to which a child is allocated. It uses contemporary national Millennium Cohort survey data for English pupils in early primary school, and accounts for a variety of factors which may explain spurious connections between stream placement and teacher judgments.

2008 data for over 800 Year Two children reported as being streamed is investigated. Both survey-reported teacher judgments of 'ability and attainment' and teacher-assessed Key Stage One results are analysed, in order to ascertain whether teacher judgments across measures and domains are associated with a pupil's stream.

Regression modelling unpicks whether children who score at the same level in recent tests of reading, maths, and overall cognitive function are assessed differently by their teachers depending on their stream (top, middle or bottom). As well as controlling for performance in these tests, analysis also accounts for the potential confounding influences of: pupil gender, ethnicity, month of birth, parents' income and education levels, parent and teacher perceptions of children's behaviour, prior in-school judgments / attainment, special educational need diagnosis, teacher characteristics, and school-type. This indicates whether pupils who perform equivalently and who are similar according to all other measured factors are assessed at different levels according to their stream placement. Analysis finds that:

- Survey-reported teacher judgments of pupils' 'ability and attainment' are associated with pupils' streams. Children in the top stream are judged to be at a higher level and children in the bottom stream at a lower level than equivalent peers.

- Teacher-assessed Key Stage One results are also related to the stream in which a pupil is placed. Even though they score equally on cognitive tests and are otherwise similar, the levels assigned to children in the highest stream are elevated and those in the bottom stream depressed.

Findings therefore indicate that streaming may exacerbate inequalities and widen or create attainment gaps. In light of this, the following paper argues that the increasing tendency to stream pupils in early primary school should urgently be reviewed, reconsidered and potentially ceased.

Contents

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, 2008; Department for Education, 1992; Department for Education, 2010; 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% of all primary pupils in 1999 (Hallam *et al*, 2003) to nearly 18% of Year Two pupils in 2008 (Campbell, 2013a).[2]

This resurrection of streaming among young children in England appears largely to be ideologically driven, given that the majority of available evidence indicates that early grouping 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 evidence and conclude that grouping is 'disadvantageous for those in lower sets and increases the overall attainment gap' (p 8).

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

In addition to these apparent inefficiencies, streaming is potentially problematic for a number of further reasons. As well as questioning the theoretical and empirical bases for its implicit premise of invariable, measurable 'ability,' studies have 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 performance, 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, and children with less educated parents. There are also some indications of disproportionality by ethnicity (Hallam and Parsons, 2013). That these inequalities exist even after taking account of manifest educational attainment 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 these apparent disparities in placement according to pupil characteristics, the evidence that streaming can be particularly 'disadvantageous for those in lower sets' is especially problematic. 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 may 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 and 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 because there are well-established relationships between teacher perceptions and pupil attainment. From Rosenthal and colleagues in the 1960s (Rosenthal & Jacobsen, 1968) to the present, a solid body of evidence has built which suggests that teacher beliefs about, and expectations and judgments of, their pupils can influence the 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).

The current study

Teacher perceptions, judgments and assessments, the last of these three potential mechanisms linking stream placement and pupil attainment, are therefore the focus of the current investigation. While some studies have explored the relationships between 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 judgments have been minimal (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 resurgence of streaming among very young pupils has emerged 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 only in recent years.

The current study uses contemporary national survey data for a large number of English pupils in early primary school, and accounts for a variety of factors which may explain spurious

apparent connections between stream placement and teacher judgments. These factors include demonstrable pupil performance / aptitude, pupil, family and teacher characteristics, measures of pupil behaviour and teacher perceptions of behaviour, and prior attainment and assessment. In addition, this paper utilises two discrete groups of measures of teacher judgments – survey-reported perceptions, and official, teacher-assessed Key Stage One test scores – thereby examining whether any effect of streaming on judgment is sensitive to / an artefact of the 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 judgments of pupils are influenced by the stream to which the pupil is allocated.

Methodology

Sample and data

Data are derived from the Millennium Cohort Study (MCS), a longitudinal investigation of a 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, a subsample of English MCS children's teachers responded to a separate survey asking for information including their perceptions of the child's attainment, of their behaviours, and details of the grouping structures within their schools. 5598 children's teachers participated in this survey, meaning that data are available for 63% of the 8887 children comprising the main wave four sample (Johnson *et al*, 2011). 914 (17.5% of the) sample pupils in state schools are reported as being streamed, and data on stream placement itself is available for 882 English, seven-year-old, singleton pupils within this group, who form the core sample for whom analyses are performed in this paper (see University of London 2008; 2011a; 2011b; 2012a; 2012b for data source references). All estimates are weighted for the MCS's design features and for attrition to the main wave four sample, as per Mostapha (2013).[3]

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: perceptions of each pupil's 'ability and attainment' as reported by teachers during MCS surveying, and officially recorded Key Stage One scores, which are entirely teacher-assessed.

[3] Weights specifically for the teacher sample are not yet available.

Outcome group one: Survey-reported teacher judgments

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 http://www.cls.ioe.ac.uk/shared/get-file.ashx?id=920&itemtype=document). They 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 paper, 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 5-35, which serves as a measure of each teacher's general judgment of pupil ability (analysis using this outcome is modelled using linear regression).

Among the 851 sample pupils with data on both stream placement and survey-reported teacher judgments, responses for each domain are, in the main, highly correlated with this overall summed total (see Table 1). Judgments of ability in physical education and in arts are less strongly related to the total and to judgments 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 judgments on physical education and arts are also carried out (using the five remaining domains; scale 1-25).

Table 1: Correlations between summed teacher judgment and judgments 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 judgments of reading and of maths ability, respectively (here, the scale is 1-5), using ordered probit modelling. Three main survey-reported teacher judgments of 'ability and attainment' are therefore used as outcomes:

1. Aggregated overall judgment (range: 5-35) – 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.

Outcome group two: Teacher-assessed Key Stage One scores

The second 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. KS1 attainment is entirely teacher-assessed, so pupils' recorded attainment at this stage is wholly dependent on the perceptions, judgments and decisions made by the respondent class teachers. This alternative outcome measure indicates whether stream placement is associated with teacher judgment when that judgment is required for official assessment rather than requested as part of a voluntary, non-school-based survey. As well as providing a test of consistency, modelling using KS1 scores indicates 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 second 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 judgments of performance across reading, writing, maths and science (equally weighted). In the sample used in this paper, scores ranges from 3 to 22.5. 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,' achieved level 1,' 'achieved level 2c,' 'achieved level 2b,' 'achieved level 2a.[4]

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.

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% are reported as being in the top stream, 31% in the middle stream, and 28% in the bottom stream. 17.2% of the slightly smaller sample[5] of MCS pupils with data on KS1 scores and with teacher response regarding whether streamed are reported to be subject to the practice (an almost identical proportion to that reported for the main survey sample). A working 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% are reported to be in the top stream, 31% in the middle stream, and 24% in the bottom stream.

[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] 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).

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 A).

Key controls: cognitive test scores

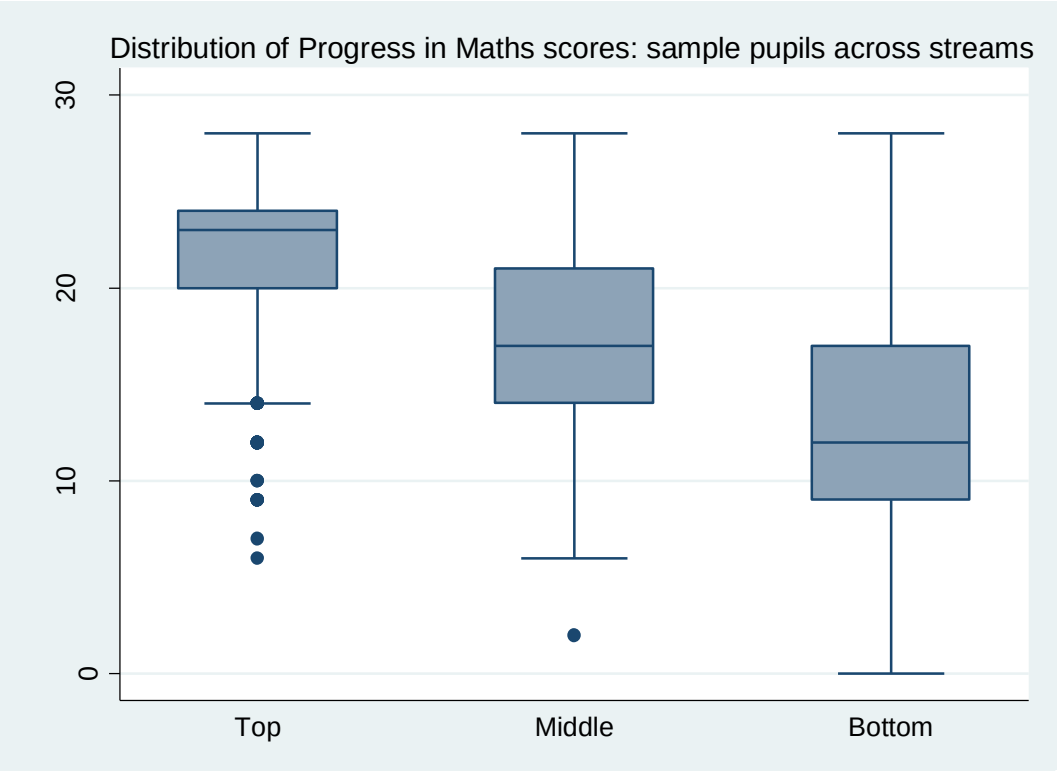
Shortly before children's teachers were contacted for their survey, the seven-year-old MCS pupils were visited in their homes by interviewers who administered three separate cognitive ability tests. The mean lag between pupil cognitive tests and teacher survey was 3.8 months, the median 3 months, and the mode 2 months. Performance scores on these tests provide key counterpoint controls in modelling to teacher judgments, 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 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 (proxied by month of birth) are controlled for in all analyses, to ensure that these factors do not confound results.

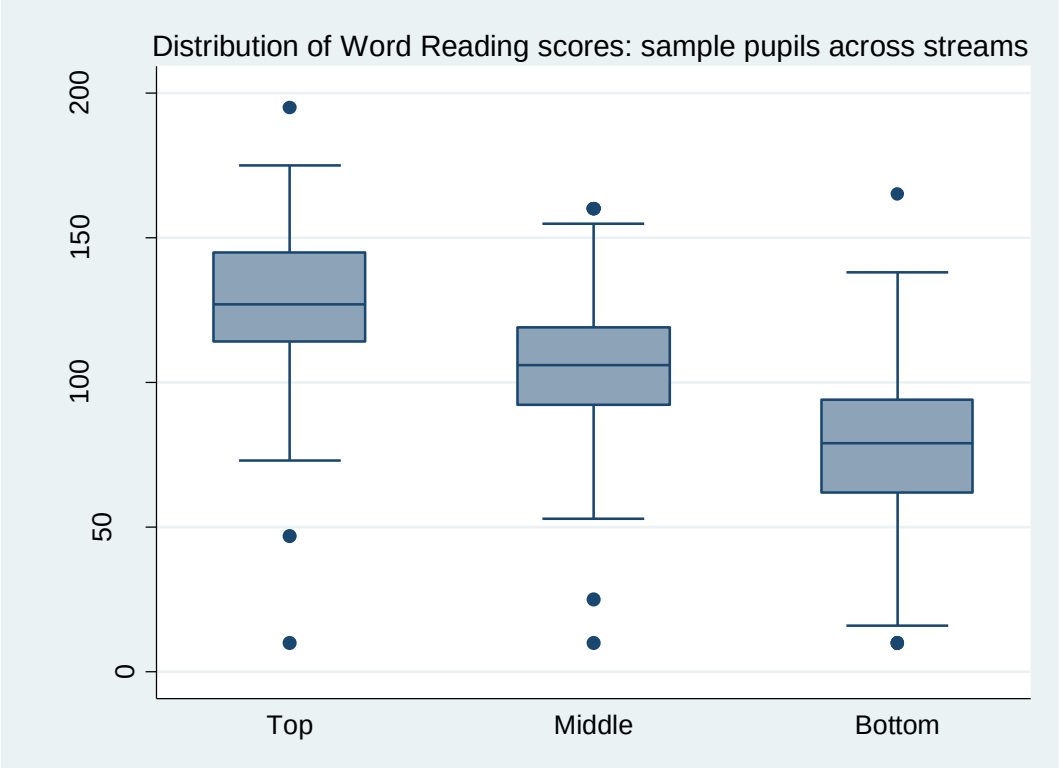
Figures 1, 2 and 3, below, illustrate the distribution of scores on the three cognitive tests for pupils situated in each stream, in the sample with survey-reported teacher judgments.

Figure 1:



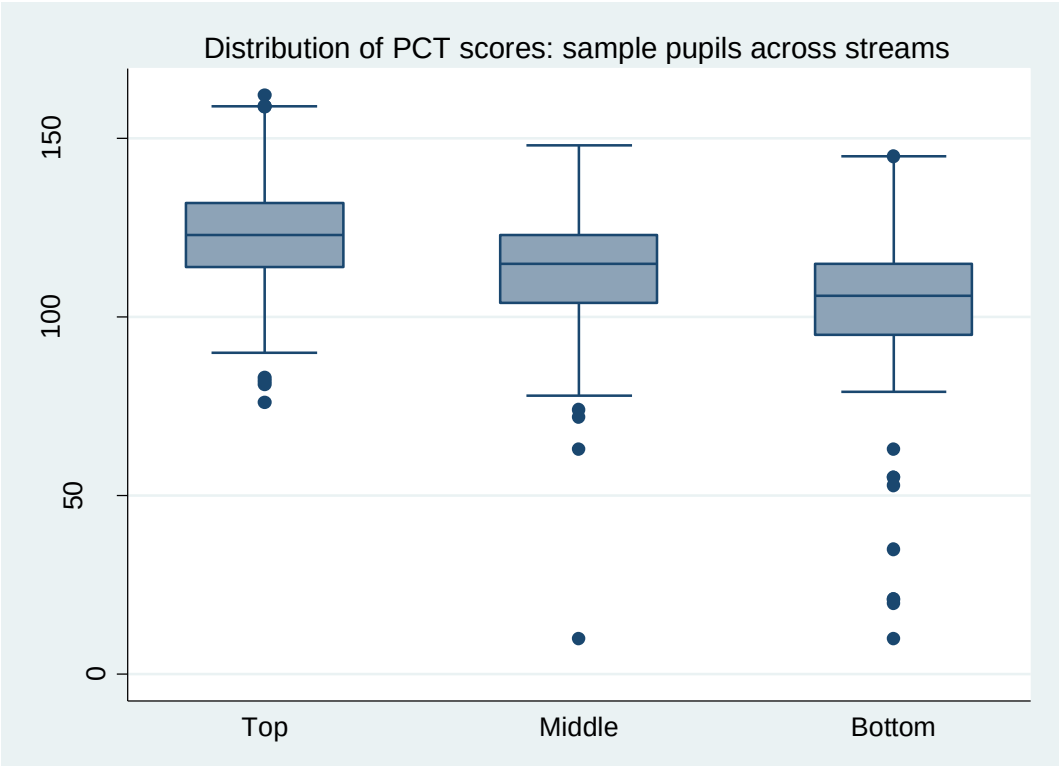
n = 840; Mean for all pupils = 18.2. Line represents median, box represents 25th and 75th percentiles (Q1 and Q3, respectively), whiskers represent Q3+1.5(Q3-Q1) / Q1-1.5*(Q3-Q1).

Figure 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).

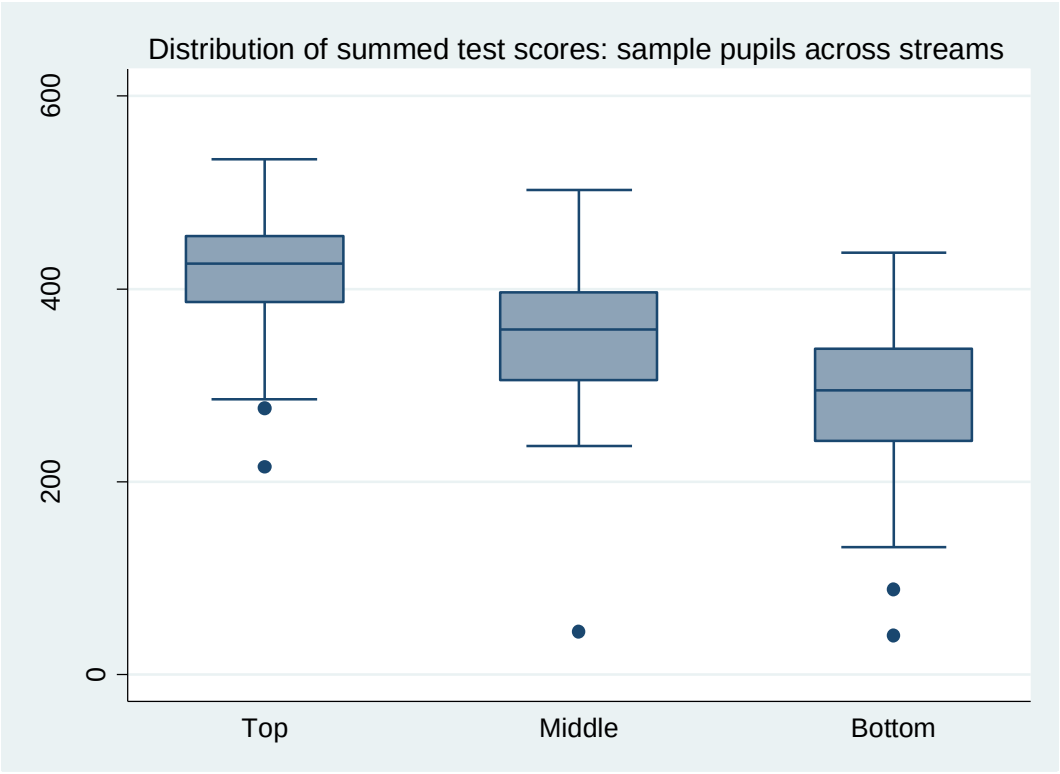
Figure 3:



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 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 B presents the equivalent information for pupils in the sub-sample with KS1 scores, and the same patterns hold for this group.

Figure 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

As discussed in the introduction, there are inequalities according to pupil and family characteristic 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 isolate any independent effect of streaming.

Pupil and family characteristics

Table 2 illustrates distributions across streams according to key individual-level characteristics within the sample with teacher survey judgments, 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.

Table 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

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

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

*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 4 shows the distribution across streams of teacher-assessed SDQ scorings at age seven, measured during the same survey within which judgments 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 4: Mean score on each scale of age seven teacher-completed SDQ test*

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

*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.

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 judgments made and conveyed by other staff within their school. In addition, prior attainment / judgments may have been influential in determining the stream to which a child is allocated. Table 5 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 measured and developing two years previously.

Table 5: Mean total FSP score at age five*

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

*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 6 shows a strong relationship in the sample between being diagnosed with 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 judgments.

Table 6: Teacher report of whether pupil has ever been recognised with SEN: percentage with each response in each stream*

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

*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 7 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 7: Percentage of sample pupils in each stream taught by teachers with each characteristic

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

*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 (survey-reported / KS1-assessed). Controls are added through cumulative model specifications, and Table 8, below, describes each specification for analyses where survey-reported judgments form the outcomes. Table 9 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).

Table 8: Cumulative specifications for models with survey-reported teacher judgments as outcomes

Specification	Predictors	Outcome
One	Stream placement	Survey-reported teacher judgments of 'ability and attainment,' summed (range 5-35; 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)
	Maths test score*	
	Reading test score^	
	PCT score* ^	
	Age at cognitive tests	
	Age at teacher survey	
Two adds…	Pupil gender	
	Pupil months 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	

***Omitted for reading test score outcome ^Omitted for maths test score outcome**

Table 9: Cumulative specifications for models with Key Stage One assessments as outcomes

Specification	Predictors	Outcome
One	Stream placement	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)
	Maths test score*	
	Reading test score^	
	PCT score* ^	
	Age at cognitive tests	
	Month of birth	
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	

***Omitted for reading test score outcome ^Omitted for maths test score outcome**

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 judgment / KS1 assessment 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 may also 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.)

Results: Stream placement and survey-reported teacher judgments

Table 10 presents key results for each model specification, for analysis where the outcome is summed survey-reported teacher judgment (see Annex C for all model coefficients). It shows an enduring relationship between pupils' stream placements and their teachers' judgments of their 'ability and attainment.' Even at specification 6 (controlling for cognitive test scores, pupil, teacher and family characteristics, previous parent and current teacher perceptions of pupil behaviour, FSP score, whether the child has SEN) being in the top stream is associated with overall teacher judgments of 'ability and attainment' (scale 5-35) 2.7 points higher ($p < .001$), and being in the bottom stream associated with judgments -1.7 points lower ($p < .001$).

A sensitivity check was carried out to examine whether removing teachers' judgments regarding less 'academic' subjects from the overall summed judgment of 'ability and attainment' affected findings. Annex D (Table 16) indicates entirely consistent results using this alternative outcome.

Table 10: 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 Ability 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: 5-35
^^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 C for all coefficients.

Table 11 shows that results also hold when teacher judgment of reading ability is considered in isolation (conditional upon children's reading ability test score and all non-cognitive test covariates), as well as when maths ability is considered alone (conditional upon maths cognitive test score). Judgments of both reading and maths ability, like summed overall teacher judgments, 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 and are otherwise similar.

Table 11: Differences in survey-reported teacher judgments of level of reading / maths 'ability and attainment' according to pupils' stream placement (specification six)^ ^^

	Reading judgment	Maths judgment
Top stream	1.227***	1.278***
	(0.153)	(0.155)
(Middle stream)	0	0
	(.)	(.)
Bottom stream	-0.805***	-1.212***
	(0.165)	(0.177)
Word Reading Test score	0.0347***	
	(0.002)	
Maths test score		0.0719***
		(0.011)
Cut 1: Constant	-10.69***	-11.07**
	(3.004)	(3.365)
Cut 2: Constant	-8.531**	-9.072**
	(3.000)	(3.349)
Cut 3: Constant	-6.205*	-6.742*
	(3.001)	(3.376)
Cut 4: Constant	-4.108	-4.817
	(3.012)	(3.386)
N	850	851

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 judgments 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. See Annex E for all coefficients.

Results: Stream placement and KS1 scores

Table 12, below, presents key results for each specification of the model where KS1 Average Points Score represents teacher assessment and judgment (see Annex F for estimates for all modelled covariates). Findings are congruent with those using survey-reported teacher judgments. 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 higher than those in the middle stream ($p < .001$), and children in the bottom stream scores 1.3 points lower ($p < .001$).

Table 12: 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 F for all coefficients.

Again, as with the survey-reported teacher judgments, 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 is controlled for (reading / maths) – 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 who are similar according to other covariates.

Table 13 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 <.001), while children in the bottom stream are less likely (p <.001).

Table 13: Differences in Key Stage One reading / maths level according to pupils' stream placement (specification six)^

	Reading level	**Maths level**
Top stream	0.898***	0.711***
	(0.219)	(0.194)
(Middle stream)	0	0
	(.)	(.)
Bottom stream	-0.467*	-1.038***
	(0.202)	(0.201)
Word Reading Test score	0.0496***	
	(0.005)	
Maths test score		0.106***
		(0.012)
Cut 1: Constant	-6.162	-7.276*
	(4.586)	(3.602)
Cut 2: Constant	-3.348	-5.486
	(4.550)	(3.565)
Cut 3: Constant	-2.029	-4.132
	(4.548)	(3.560)
Cut 4: Constant	-0.224	-2.628
	(4.572)	(3.564)
N	440	465

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. See Annex G for all coefficients.

Discussion

This research set out to explore whether teacher judgments 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 judgments of pupils' academic ability and attainment. This holds both for survey-reported teacher perceptions, and for assessments of performance at KS1. The hypothesis that teacher judgments 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. Moreover, analyses in this paper 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 therefore a danger that the increasing use of streaming among primary school pupils will only perpetuate or widen attainment gaps.

Alternative and additional explanations

Findings in this paper indicate a cross-domain and pan-situational relationship between stream placement and teacher judgments of pupils. They show that otherwise similar sample children who score equivalently on cognitive tests taken in the same year that the teacher makes their judgment are judged differently, depending on their placement. However, as well as supporting the hypothesis that stream placement influences teacher perceptions and assessments, findings here may be interpreted as suggesting a number of alternative explanations for this association.

It is possible that, in the period between cognitive testing and teacher survey / KS1 assessment, pupils' actual performances (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 judgments. As discussed in the introduction, 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.

Though these possibilities cannot be eliminated using the data available, two key points can be noted. Firstly, the time lags between cognitive test completion and teacher judgments are short (particularly for survey-reported assessments, at 2-4 months on average), suggesting that a discrepancy between judgment of attainment 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 earlier test performances 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 that findings in this paper 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 themselves indicates the immediacy, strength and enduring influence of the practice of streaming.

Conclusions and policy recommendations

Whichever explanation for results in this paper is preferred, streaming appears to have a durable effect on a range of teacher judgments 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, 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 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, there is an arguable imperative for instigation of this data collection to 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? So far, the work presented in this paper supports the latter.

References

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

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.* Available online at: http://gtcni.openrepository.com/gtcni/bitstream/2428/26653/1/RS_9-2_report_160508_Structures_for_teaching_learning.pdf (accessed 16 February 2014).

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

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.

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.* Available online at: http://www.bris.ac.uk/cmpo/publications/papers/2009/wp221.pdf (accessed 16 February 2013).

Campbell, T. (2013a) In-school ability-grouping and the month of birth effect: Preliminary evidence from the Millennium Cohort Study. Available online at: http://www.cls.ioe.ac.uk/shared/get-file.ashx?itemtype=document&id=1618 (accessed 16 February 2013).

Campbell, T. (2013b) Stereotyped at seven? Biases in teacher judgments of pupils' ability and attainment. Available online at: http://www.cls.ioe.ac.uk/shared/get-file.ashx?itemtype=document&id=1715 (accessed 16 February 2014).

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.* Available online at: http://image.guardian.co.uk/sys-files/Education/documents/2007/11/20/newopps.pdf (accessed 16 February 2014).

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.

Department for Children, Schools and Families (2008) *21st Century Schools: A World Class Education for Every Child.* Available online at: http://webarchive.nationalarchives.gov.uk/20130401151715/https://www.education.gov.uk/publications/eOrderingDownload/DCSF-01044-2008.pdf (accessed 16 February 2014).

Department for Education (1992) *White Paper: Choice and Diversity.* Available online at: http://www.educationengland.org.uk/documents/wp1992/choice-and-diversity.html (accessed 16 February 2014).

Department for Education (2010) *The Importance of Teaching - The Schools White Paper 2010.* Available online at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/175429/CM-7980.pdf (accessed 16 February 2014).

Department for Education and Skills (2005) *Higher Standards, Better Schools For All: More choice for parents and pupils.* Available online at: http://dera.ioe.ac.uk/5496/1/DfES-Schools%20White%20Paper.pdf (accessed 16 February 2014).

Dunne, M., Humphreys, S., Sebba, J., Dyson, A., Gallannaugh, F., & Muijs, D. (2007) *Effective Teaching and Learning for Pupils in Low Attaining Groups.* Available online at: http://webarchive.nationalarchives.gov.uk/20130401151655/https://www.education.gov.uk/publications/eOrderingDownload/DCSF-RR011.pdf (accessed 16 February 2014).

Good, T. L. (1987) Two Decades of Research on Teacher Expectations: Findings and Future Directions. *Journal of Teacher Education,* 38, 32-47.

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.

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. Available online at: http://www.cls.ioe.ac.uk/shared/get-file.ashx?id=598&itemtype=document (accessed 16 February 2014).

Huang, Y., & Gatenby, R. (2010) *Millennium Cohort Study Sweep 4 Teacher Survey Technical Report.* Available online at: http://www.cls.ioe.ac.uk/shared/get-file.ashx?id=489&itemtype=document (accessed 16 February 2014).

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.* Available online at: http://www.cls.ioe.ac.uk/shared/get-file.ashx?id=1341&itemtype=document (accessed 16 February 2013).

Johnson, J. and Rosenberg, R. (2013) *A Guide to the Linked Education Administrative Datasets: Millennium Cohort Study.* Available online at: http://www.cls.ioe.ac.uk/shared/get-file.ashx?id=1342&itemtype=document (accessed 16 February 2014).

Kutnick, P., Sebba, J., Blatchford, P., Galton, M., & Thorp, J. (2005) *The effects of pupil grouping: Literature review.* Available online at: http://webarchive.nationalarchives.gov.uk/20130401151655/https://www.education.gov.uk/publications/eOrderingDownload/RR688.pdf (accessed 16 February 2014).

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.* Available online at: http://webarchive.nationalarchives.gov.uk/20130401151715/https://www.education.gov.uk/publications/eOrderingDownload/RR796.pdf (accessed 16 February 2014).

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).* Available online at: http://www.cls.ioe.ac.uk/shared/get-file.ashx?id=1749&itemtype=document (accessed 16 February 2014).

OECD (2012) *Equity and Quality in Education: Supporting Disadvantaged Students and Schools.* Available online at: http://www.keepeek.com/Digital-Asset-Management/oecd/education/equity-and-quality-in-education_9789264130852-en_ (accessed 16 February 2014)

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.

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.

Shih, M., Pittinsky, T. L., and Trahan, A. (2005) *Domain specific effects of stereotypes on performance*. Working paper no. RWP05-026. Available online at: http://www.cs.cmu.edu/~cfrieze/courses/Shih.pdf (accessed 16 February 2014).

Steele, C. M., and 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).* Available online at: http://webarchive.nationalarchives.gov.uk/20130401151715/https://www.education.gov.uk/publications/eOrderingDownload/DCSF-RR002.pdf (accessed 16 Feb 2014)

Sutton Trust / Educational Endowment Foundation (2014) *Teaching and Learning Toolkit.* Available online at:

http://educationendowmentfoundation.org.uk/uploads/toolkit/EEF_Teaching_and_learning_toolkit_Feb_2014.pdf (accessed 16 February 2014).

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.

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

Wiliam, 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.

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

Table 14 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 14: Percentage of sample^ pupils who are streamed and coefficients from probit regression of whether streamed / not where each characteristic is simultaneously included as predictor^^

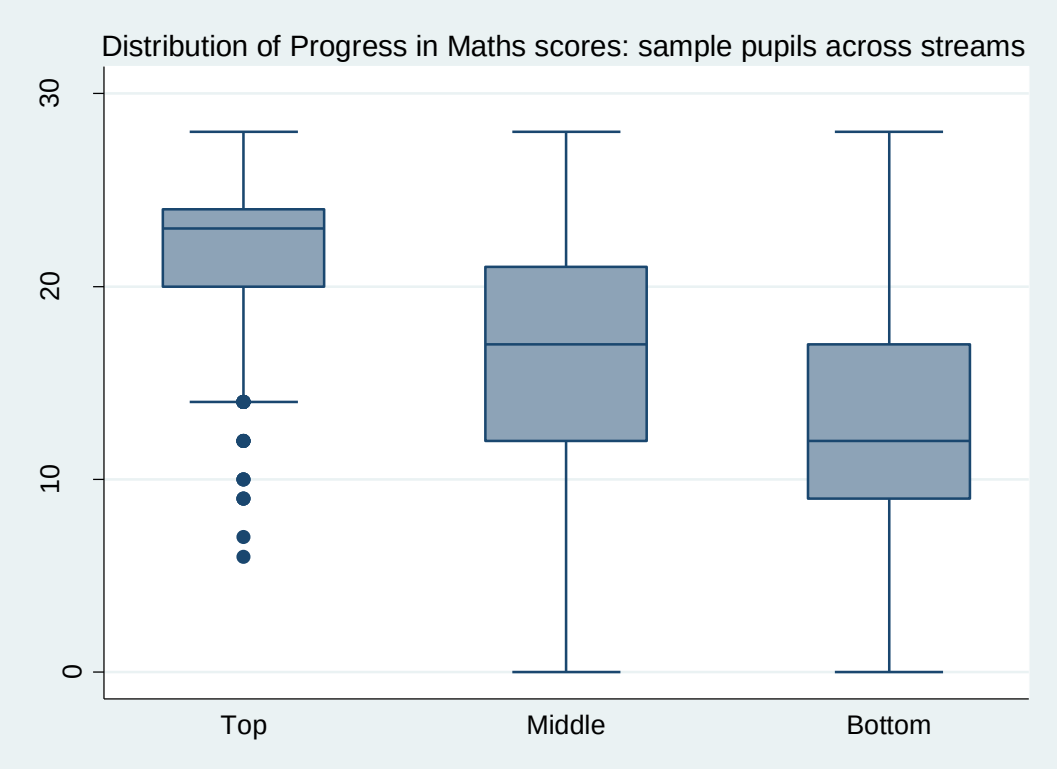
	Percentage streamed (a)	**Probit regression coefficient (b)**
All sample pupils (n = 4999 / 4951)	17.6	
Boys (n = 2508)	17.8	(reference)
Girls (n = 2491)	17.2	.022 (.045)
White (4000)	17.1	(reference)
Mixed ethnicity (169)	20.3	.143 (.138)
Indian (148)	24.7	.295 (.172)*
Pakistani / Bangladeshi (363)	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 – 4951 [vs 4999] cases in total are included in the model)

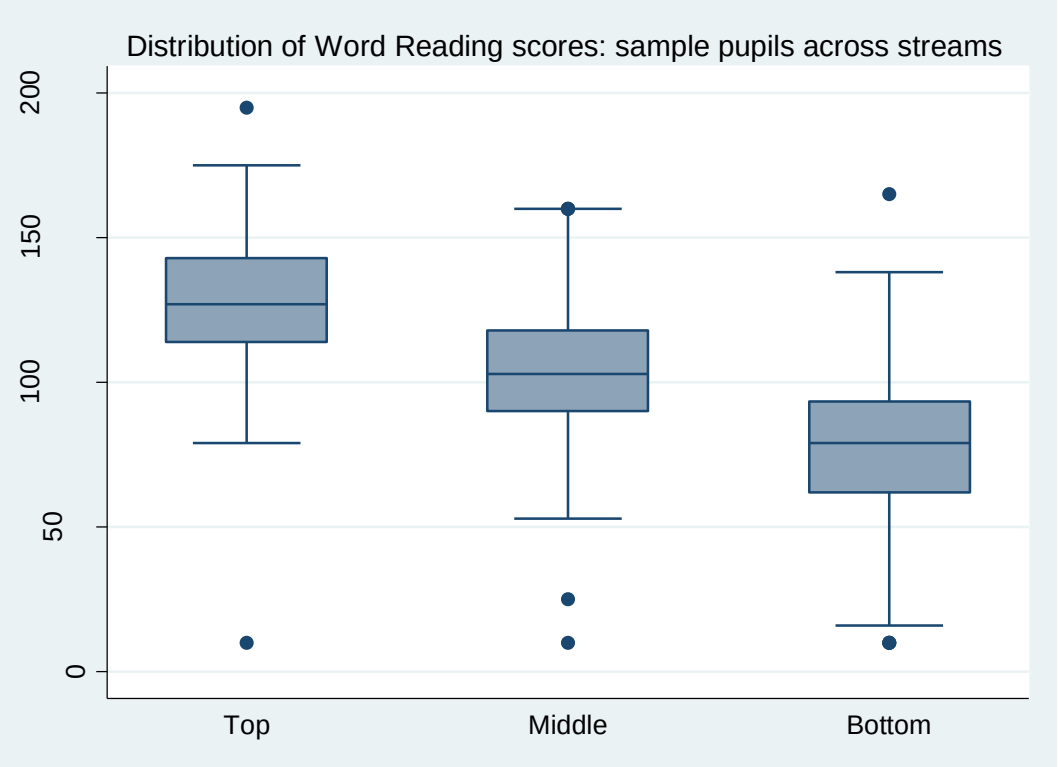
Annex B: Distribution across streams of test scores for KS1 sample

Figure 5:



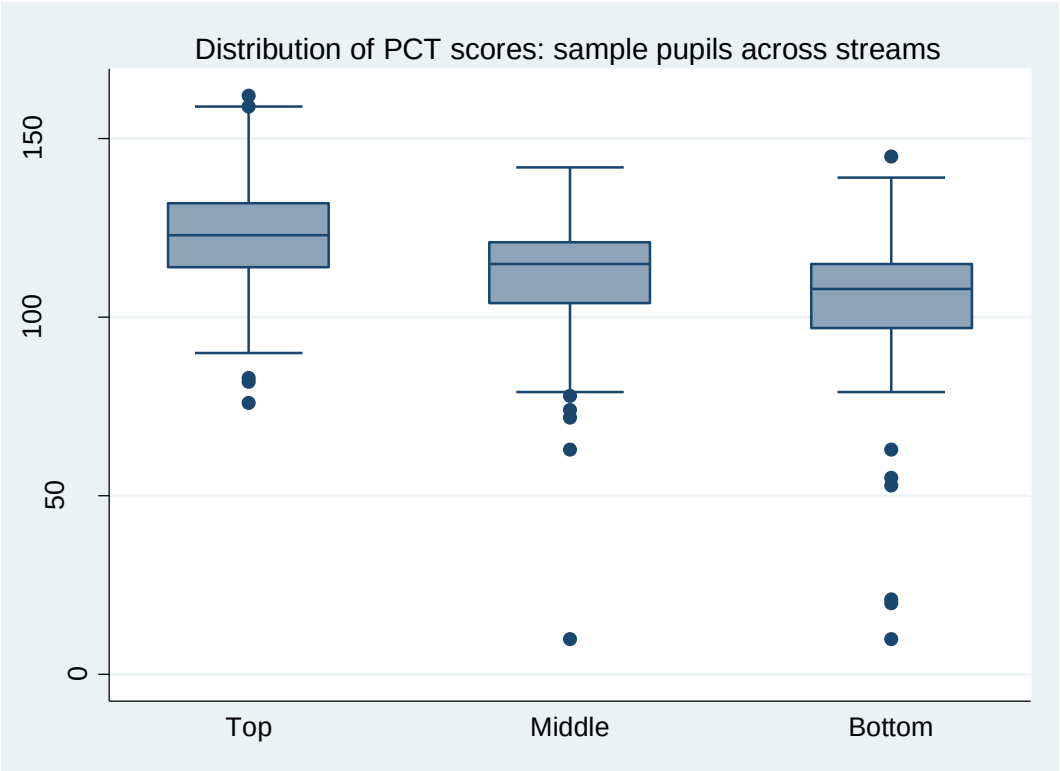
n = 644; Mean for all pupils = 18.2. Line represents median, box represents 25th and 75th percentiles (Q1 and Q3, respectively), whiskers represent Q3+1.5(Q3-Q1) / Q1-1.5*(Q3-Q1).

Figure 6:



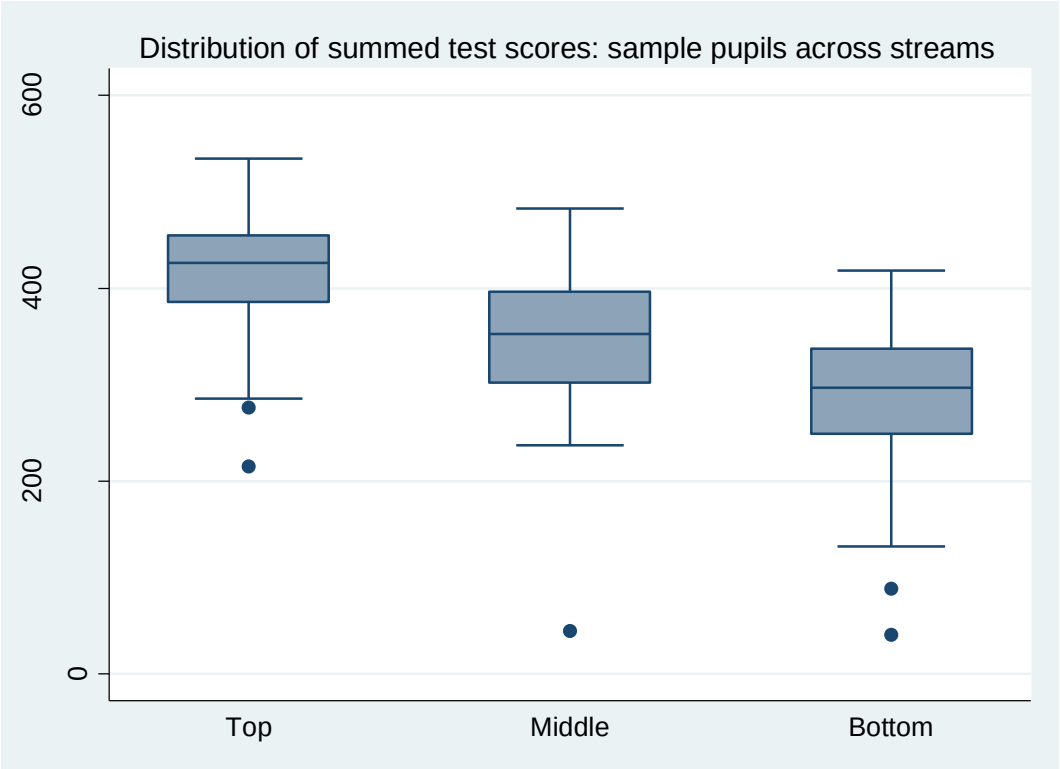
n = 644; Mean for all pupils = 108.9. Line represents median, box represents 25th and 75th percentiles (Q1 and Q3, respectively), whiskers represent Q3+1.5(Q3-Q1) / Q1-1.5*(Q3-Q1).

Figure 7:



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 8:



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 C: Full model for summed survey-reported teacher judgments

Table 15: 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 Ability 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	0.251	0.191	0.148	0.166	0.191
	(0.462)	(0.443)	(0.418)	(0.408)	(0.410)	(0.410)
92-93 months	0.533	0.175	0.0909	0.163	0.200	0.230
	(0.405)	(0.330)	(0.343)	(0.342)	(0.343)	(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	0.442	0.555	0.447	0.513

		(0.525)	(0.545)	(0.530)	(0.510)	(0.553)
Pakistani / Bangladeshi		-0.843*	-0.730*	-0.588	-0.709+	-0.654+
		(0.362)	(0.369)	(0.390)	(0.383)	(0.389)
Black / Black British		-1.299+	-1.625**	-1.577*	-1.588*	-1.463*
		(0.668)	(0.590)	(0.629)	(0.639)	(0.622)
(Higher-income)		0	0	0	0	0
		(.)	(.)	(.)	(.)	(.)
Low-income		-0.463+	-0.427	-0.395	-0.364	-0.392
		(0.277)	(0.273)	(0.262)	(0.263)	(0.265)
Parent level 1 qual		-0.564	-0.613	-0.563	-0.577	-0.614
		(0.514)	(0.497)	(0.496)	(0.504)	(0.495)
Parent level 2 qual		-0.857+	-0.961*	-0.982*	-0.994*	-0.977*
		(0.452)	(0.449)	(0.450)	(0.453)	(0.454)
Parent level 3 qual		-0.0611	-0.0986	-0.0707	-0.127	-0.0977
		(0.495)	(0.460)	(0.456)	(0.459)	(0.462)
Parent level 4 qual		-0.292	-0.360	-0.390	-0.421	-0.431
		(0.452)	(0.442)	(0.445)	(0.449)	(0.450)
(Parent level 5 qual – ref)		0	0	0	0	0
		(.)	(.)	(.)	(.)	(.)
Parent overseas qual		0.453	0.386	0.435	0.456	0.437
		(0.918)	(0.864)	(0.859)	(0.902)	(0.904)
Parent no qual		-1.410*	-1.353*	-1.240*	-1.217*	-1.205*
		(0.570)	(0.556)	(0.565)	(0.567)	(0.560)
(Age five SDQ emotional – 'normal')			0	0	0	0
			(.)	(.)	(.)	(.)
Age five SDQ emotional – 'borderline'			0.0528	0.114	0.0453	0.135
			(0.468)	(0.464)	(0.468)	(0.484)
Age five SDQ emotional – 'abnormal'			-0.194	-0.152	-0.171	-0.143
			(0.403)	(0.394)	(0.396)	(0.408)
Age five SDQ emotional – missing data			1.779	1.773	1.720	1.675
			(1.244)	(1.105)	(1.092)	(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+ (2.043)	4.072+ (2.170)	3.912+ (2.149)	4.253+ (2.291)
(Age five SDQ pro-social – 'normal')			0 (.)	0 (.)	0 (.)	0 (.)
Age five SDQ pro-social – 'borderline'			0.395 (0.424)	0.544 (0.405)	0.496 (0.401)	0.536 (0.399)
Age five SDQ pro-social – 'abnormal'			0.157 (0.528)	0.116 (0.490)	0.187 (0.498)	0.142 (0.511)
Age five SDQ pro-social – missing data			-3.060** (1.021)	-3.175** (1.035)	-2.601* (1.150)	-2.800* (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** (0.095)	0.270** (0.096)	0.258** (0.095)	0.264** (0.096)
Age seven SDQ hyperactive			-0.154** (0.057)	-0.164** (0.056)	-0.157** (0.055)	-0.159** (0.055)
Age seven SDQ peer			-0.149+ (0.086)	-0.156+ (0.086)	-0.148+ (0.085)	-0.145+ (0.087)
Age seven SDQ pro-social			0.148** (0.054)	0.150** (0.055)	0.154** (0.054)	0.151** (0.056)
(No behaviour difficulties)			0 (.)	0 (.)	0 (.)	0 (.)
Minor behaviour difficulties			0.0112 (0.284)	0.0755 (0.287)	0.130 (0.288)	0.151 (0.292)
Definite behaviour difficulties			-0.742 (0.516)	-0.711 (0.519)	-0.523 (0.533)	-0.522 (0.536)
Severe behaviour difficulties			-2.608** (0.929)	-2.510** (0.919)	-2.390** (0.912)	-2.426** (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*	-0.678+
					(0.356)	(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	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
$\boldsymbol{R^2}$	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: 5-35

Annex D: Summed survey-reported teacher judgments: 'academic' domains only

Table 16: Difference in survey-reported summed teacher judgment of academic 'ability and attainment' according to pupils' stream placement^ ^^

	Spec 1	Spec 2	Spec 3	Spec 4	Spec 5	Spec 6
Top stream	2.779***	2.519***	2.406***	2.327***	2.347***	2.308***
	(0.216)	(0.208)	(0.211)	(0.207)	(0.205)	(0.209)
(Middle stream)	0	0	0	0	0	0
	(.)	(.)	(.)	(.)	(.)	(.)
Bottom stream	-2.304***	-2.103***	-1.859***	-1.806***	-1.633***	-1.640***
	(0.253)	(0.253)	(0.249)	(0.235)	(0.237)	(0.229)
Maths test score	0.0728***	0.0735***	0.0585***	0.0551***	0.0516**	0.0507**
	(0.016)	(0.017)	(0.017)	(0.016)	(0.016)	(0.016)
Word Reading Test score	0.0439***	0.0460***	0.0447***	0.0419***	0.0404***	0.0408***
	(0.004)	(0.004)	(0.003)	(0.004)	(0.004)	(0.004)
Pattern Construction Ability test score	0.0215***	0.0168**	0.0119*	0.0115*	0.0120*	0.0108+
	(0.005)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Constant	4.915	26.75***	28.09***	27.51***	27.42***	27.39***
	(4.354)	(5.837)	(5.807)	(5.734)	(5.647)	(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: 5-35
^^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 E: Full model for teacher survey-reported reading and maths judgments in isolation, specification six

Table 17: Differences in survey-reported teacher judgments of level of reading / maths 'ability and attainment' according to pupils' stream placement^

	Reading judgment	Maths judgment
Top stream	1.227***	1.278***
	(0.153)	(0.155)
(Middle stream)	0	0
	(.)	(.)
Bottom stream	-0.805***	-1.212***
	(0.165)	(0.177)
Word Reading Test score	0.0347***	
	(0.002)	
Maths test score		0.0719***
		(0.011)
Age at tests	-0.117***	-0.103**
	(0.032)	(0.037)
Age at teacher survey: missing data	-0.341	0.0169
	(0.246)	(0.225)
82-87 months	-0.000493	0.359
	(0.214)	(0.226)
88-89 months	-0.208	0.189
	(0.209)	(0.189)
90-91 months	-0.116	0.149
	(0.183)	(0.195)
92-93 months	-0.0381	0.0242
	(0.160)	(0.153)
(94-104 months)	0	0
	(.)	(.)
Boy	-0.177+	0.366***
	(0.096)	(0.097)
(Girl)	0	0
	(.)	(.)
August-born	-1.014**	-1.262***
	(0.366)	(0.346)
July-born	-1.299***	-1.460***
	(0.315)	(0.342)
June-born	-1.113***	-1.160***
	(0.290)	(0.269)
May-born	-0.986***	-0.861**
	(0.250)	(0.271)
April-born	-0.795*	-1.207***
	(0.317)	(0.313)
March-born	-0.507*	-0.776**
	(0.225)	(0.249)
February-born	-0.536*	-0.408

	(0.271)	(0.284)
January-born	-0.579*	-0.412+
	(0.251)	(0.240)
December-born	-0.614*	-0.436*
	(0.239)	(0.219)
November-born	-0.325	-0.577*
	(0.211)	(0.243)
October-born	-0.279	-0.285
	(0.202)	(0.202)
(September-born)	0	0
	(.)	(.)
(White ethnicity)	0	0
	(.)	(.)
Mixed / 'other' / missing data	-0.0864	0.0691
	(0.205)	(0.211)
Indian	0.140	0.125
	(0.214)	(0.245)
Pakistani / Bangladeshi	-0.455**	-0.193
	(0.137)	(0.199)
Black / Black British	-0.597*	-0.853***
	(0.232)	(0.242)
(Higher-income)	0	0
	(.)	(.)
Low-income	-0.241+	-0.110
	(0.126)	(0.120)
Parent level 1 qual	-0.218	0.0132
	(0.234)	(0.276)
Parent level 2 qual	-0.386+	-0.197
	(0.205)	(0.229)
Parent level 3 qual	-0.140	0.130
	(0.206)	(0.237)
Parent level 4 qual	-0.160	0.159
	(0.202)	(0.234)
(Parent level 5 qual – ref)	0	0
	(.)	(.)
Parent overseas qual	0.183	0.145
	(0.332)	(0.319)
Parent no qual	-0.254	-0.0682
	(0.279)	(0.258)
(Age five SDQ emotional – 'normal')	0	0
	(.)	(.)
Age five SDQ emotional – 'borderline'	0.0255	0.387*
	(0.223)	(0.173)
Age five SDQ emotional – 'abnormal'	-0.0259	-0.144
	(0.209)	(0.192)
Age five SDQ emotional – missing data	0.514	1.445**
	(0.322)	(0.540)
(Age five SDQ conduct – 'normal')	0	0
	(.)	(.)
Age five SDQ conduct – 'borderline'	-0.0346	0.0422

	(0.164)	(0.175)
Age five SDQ conduct – 'abnormal'	-0.104 (0.173)	-0.304^{+} (0.156)
Age five SDQ conduct – missing data	-0.534 (0.589)	-3.430** (1.063)
(Age five SDQ hyperactive – 'normal')	0 (.)	0 (.)
Age five SDQ hyperactive – 'borderline'	0.0869 (0.167)	0.0873 (0.181)
Age five SDQ hyperactive – 'abnormal'	0.130 (0.196)	0.197 (0.201)
Age five SDQ hyperactive – missing data	0.772** (0.268)	0.807^{+} (0.430)
(Age five SDQ peer – 'normal')	0 (.)	0 (.)
Age five SDQ peer – 'borderline'	-0.163 (0.167)	-0.0343 (0.160)
Age five SDQ peer – 'abnormal'	0.161 (0.187)	0.00254 (0.188)
Age five SDQ peer – missing data	1.137* (0.497)	1.020 (0.850)
(Age five SDQ pro-social – 'normal')	0 (.)	0 (.)
Age five SDQ pro-social – 'borderline'	0.183 (0.228)	0.278^{+} (0.167)
Age five SDQ pro-social – 'abnormal'	0.571* (0.261)	-0.0942 (0.250)
Age five SDQ pro-social – missing data	-1.847*** (0.341)	0.129 (0.692)
Age seven SDQ emotional	-0.00673 (0.028)	0.00491 (0.026)
Age seven SDQ conduct	-0.00429 (0.046)	0.0742^{+} (0.040)
Age seven SDQ hyperactive	-0.0208 (0.030)	-0.0284 (0.026)
Age seven SDQ peer	-0.0580 (0.035)	-0.00124 (0.038)
Age seven SDQ pro-social	0.0226 (0.027)	0.0181 (0.024)
(No behaviour difficulties)	0 (.)	0 (.)
Minor behaviour difficulties	0.277 (0.175)	-0.0205 (0.144)
Definite behaviour difficulties	0.0331 (0.289)	-0.273 (0.230)
Severe behaviour difficulties	-0.409	-0.793*

	(0.389)	(0.388)
(FSP score – bottom quintile)	0 (.)	0 (.)
FSP score – second quintile	0.244 (0.165)	0.318* (0.158)
FSP score – third quintile	0.335+ (0.198)	0.638*** (0.147)
FSP score – fourth quintile	0.203 (0.214)	0.562** (0.182)
FSP score – top quintile	0.563* (0.232)	0.867*** (0.208)
FSP score – missing data	0.299 (0.238)	0.459* (0.201)
Recognised SEN	-0.457** (0.155)	-0.300+ (0.156)
(No SEN / do not know)	0 (.)	0 (.)
(Female teacher)	0 (.)	0 (.)
Male teacher	0.0664 (0.180)	-0.0285 (0.248)
Teacher gender missing data	0.346 (0.233)	0.208 (0.240)
Teacher years taught: missing data	-0.369 (0.235)	-0.412 (0.270)
Teacher years taught: 24-48 years	0.340 (0.329)	-0.573+ (0.319)
Teacher years taught: 14-23 years	-0.0398 (0.238)	-0.317 (0.274)
Teacher years taught: 8-13 years	0.329 (0.240)	-0.633** (0.214)
Teacher years taught: 4-7 years	-0.126 (0.224)	-0.610** (0.231)
(Teacher years taught: 1-3 years – ref)	0 (.)	0 (.)
Teacher years at school: missing data	0.173 (0.232)	0.108 (0.248)
Teacher years at school: 8-48 years	-0.0193 (0.209)	0.335 (0.227)
Teacher years at school: 4-7 years	0.342+ (0.202)	0.546** (0.191)
(Teacher years at school: 1-3 years – ref)	0 (.)	0 (.)
Cut 1: Constant	-10.69*** (3.004)	-11.07** (3.365)
Cut 2: Constant	-8.531** (3.000)	-9.072** (3.349)
Cut 3: Constant	-6.205* (3.001)	-6.742* (3.376)

Cut 4: Constant	-4.108 (3.012)	-4.817 (3.386)
N	850	851

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 judgments of reading / maths ability; range: 1-35

Annex F: Full model for Key Stage One Average Points Score outcome

Table 18: 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 Ability 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 (0.548)	-0.300 (0.515)	-0.339 (0.487)	-0.237 (0.488)	-0.286 (0.479)	-0.307 (0.482)
January-born	-0.694 (0.436)	-0.622 (0.430)	-0.607 (0.385)	-0.473 (0.374)	-0.467 (0.375)	-0.504 (0.374)
December-born	-0.305 (0.383)	-0.254 (0.369)	-0.403 (0.332)	-0.327 (0.318)	-0.343 (0.318)	-0.288 (0.312)
November-born	-0.269 (0.379)	-0.242 (0.363)	-0.359 (0.347)	-0.329 (0.343)	-0.349 (0.350)	-0.342 (0.342)
October-born	0.112 (0.352)	0.224 (0.356)	0.0597 (0.339)	0.110 (0.341)	0.115 (0.345)	0.117 (0.337)
(September-born)	0 (.)	0 (.)	0 (.)	0 (.)	0 (.)	0 (.)
Boy		-0.186 (0.146)	-0.0948 (0.138)	-0.0666 (0.137)	-0.0549 (0.140)	-0.0539 (0.140)
(Girl)		0 (.)	0 (.)	0 (.)	0 (.)	0 (.)
(White ethnicity)		0 (.)	0 (.)	0 (.)	0 (.)	0 (.)
Mixed / 'other' / missing data		0.548+ (0.283)	0.524* (0.256)	0.558* (0.264)	0.494+ (0.262)	0.527+ (0.275)
Indian		0.710* (0.343)	0.465 (0.355)	0.520 (0.335)	0.439 (0.328)	0.424 (0.336)
Pakistani / Bangladeshi		-0.154 (0.271)	-0.132 (0.298)	-0.0917 (0.315)	-0.179 (0.304)	-0.127 (0.314)
Black / Black British		-0.421 (0.463)	-0.610 (0.466)	-0.541 (0.504)	-0.586 (0.489)	-0.573 (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
$\boldsymbol{R^2}$	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 G: Full model for Key Stage One reading / maths levels outcomes, specification six

Table 19: Probability of moving up a Key Stage One reading / maths level according to pupils' stream placement^

	Reading level	Maths level
Top stream	0.898***	0.711***
	(0.219)	(0.194)
(Middle stream)	0	0
	(.)	(.)
Bottom stream	-0.467*	-1.038***
	(0.202)	(0.201)
Word Reading Test score	0.0496***	
	(0.005)	
Maths test score		0.106***
		(0.012)
Age at tests	-0.0568	-0.0569
	(0.051)	(0.039)
August-born	-0.900	-0.194
	(0.622)	(0.450)
July-born	-1.610**	-0.466
	(0.569)	(0.450)
June-born	-1.177**	0.00761
	(0.437)	(0.385)
May-born	-0.462	-0.207
	(0.446)	(0.369)
April-born	-0.843*	0.0889
	(0.408)	(0.423)
March-born	-0.901*	0.488
	(0.399)	(0.333)
February-born	-0.975*	0.622
	(0.434)	(0.417)
January-born	-0.848*	0.573+
	(0.375)	(0.327)
December-born	-0.805*	0.312
	(0.325)	(0.295)
November-born	-1.249***	-0.151
	(0.370)	(0.285)
October-born	-0.598	0.586+
	(0.384)	(0.319)
(September-born)	0	0
	(.)	(.)
Boy	-0.170	0.455**
	(0.136)	(0.137)
(Girl)	0	0
	(.)	(.)
(White ethnicity)	0	0

	(.)	(.)
Mixed / 'other' / missing data	0.497 (0.318)	0.215 (0.283)
Indian	0.0135 (0.364)	-0.181 (0.394)
Pakistani / Bangladeshi	-0.357 (0.223)	-0.141 (0.263)
Black / Black British	0.407 (0.566)	0.134 (0.466)
(Higher-income)	0 (.)	0 (.)
Low-income	-0.00905 (0.151)	0.0693 (0.157)
Parent level 1 qual	0.849+ (0.487)	-0.482 (0.320)
Parent level 2 qual	0.539 (0.430)	-0.277 (0.283)
Parent level 3 qual	0.676 (0.450)	0.00190 (0.298)
Parent level 4 qual	0.894+ (0.471)	0.0467 (0.305)
(Parent level 5 qual – ref)	0 (.)	0 (.)
Parent overseas qual	1.240* (0.585)	-0.143 (0.410)
Parent no qual	0.334 (0.504)	-0.486 (0.353)
Community mainstream school	0 (.)	0 (.)
Voluntary aided school	0.0895 (0.250)	-0.137 (0.189)
Voluntary controlled / foundation	0.229 (0.205)	-0.0648 (0.208)
(Did not join school in current academic year)	0 (.)	0 (.)
Joined school in current academic year	-0.229 (0.261)	0.302 (0.281)
(Did not join school in last academic year)	0 (.)	0 (.)
Joined school in last academic year	0.431 (0.347)	0.519* (0.252)
(Age five SDQ emotional – 'normal')	0 (.)	0 (.)
Age five SDQ emotional – 'borderline'	-0.359 (0.337)	-0.108 (0.346)
Age five SDQ emotional – 'abnormal'	-0.140 (0.301)	-0.184 (0.239)
Age five SDQ emotional – missing data	1.343 (0.918)	5.005*** (0.771)

(Age five SDQ conduct – 'normal')	0 (.)	0 (.)
Age five SDQ conduct – 'borderline'	-0.476* (0.187)	0.0300 (0.169)
Age five SDQ conduct – 'abnormal'	-0.518* (0.248)	-0.107 (0.189)
Age five SDQ conduct – missing data	-4.194*** (1.239)	-7.512*** (1.103)
(Age five SDQ hyperactive – 'normal')	0 (.)	0 (.)
Age five SDQ hyperactive – 'borderline'	0.0219 (0.221)	0.0651 (0.210)
Age five SDQ hyperactive – 'abnormal'	-0.490* (0.211)	0.00328 (0.180)
Age five SDQ hyperactive – missing data	1.515** (0.493)	1.044+ (0.530)
(Age five SDQ peer – 'normal')	0 (.)	0 (.)
Age five SDQ peer – 'borderline'	0.107 (0.177)	-0.0523 (0.182)
Age five SDQ peer – 'abnormal'	0.0235 (0.262)	-0.0204 (0.284)
Age five SDQ peer – missing data	0.767 (0.542)	1.178+ (0.644)
(Age five SDQ pro-social – 'normal')	0 (.)	0 (.)
Age five SDQ pro-social – 'borderline'	0.257 (0.256)	-0.134 (0.286)
Age five SDQ pro-social – 'abnormal'	0.707* (0.303)	-0.707* (0.355)
Age five SDQ pro-social – missing data	0 (.)	0 (.)
Age seven SDQ emotional	0.0272 (0.033)	0.00722 (0.028)
Age seven SDQ conduct	0.0383 (0.065)	0.00253 (0.054)
Age seven SDQ hyperactive	-0.0366 (0.038)	0.0473 (0.041)
Age seven SDQ peer	-0.0465 (0.055)	0.0527 (0.041)
Age seven SDQ pro-social	-0.0602 (0.041)	-0.0518 (0.039)
(No behaviour difficulties)	0 (.)	0 (.)
Minor behaviour difficulties	0.0200	-0.440**

	(0.189)	(0.158)
Definite behaviour difficulties	-0.105 (0.289)	-0.413+ (0.234)
Severe behaviour difficulties	-0.338 (0.638)	-1.345* (0.543)
(FSP score – bottom quintile)	0 (.)	0 (.)
FSP score – second quintile	0.0790 (0.180)	0.481** (0.171)
FSP score – third quintile	0.258 (0.236)	0.573* (0.251)
FSP score – fourth quintile	0.403 (0.279)	0.705** (0.232)
FSP score – top quintile	0.389 (0.335)	0.749* (0.307)
FSP score – missing data	0.0702 (0.317)	0.309 (0.283)
Recognised SEN	-0.647*** (0.187)	-0.461* (0.205)
(No SEN / do not know)	0 (.)	0 (.)
(Female teacher)	0 (.)	0 (.)
Male teacher	-0.451 (0.340)	0.284 (0.435)
Teacher gender missing data	-0.00504 (0.345)	0.243 (0.340)
Teacher years taught: missing data	-0.704 (0.494)	-0.375 (0.479)
Teacher years taught: 24-48 years	0.124 (0.536)	-0.268 (0.418)
Teacher years taught: 14-23 years	-0.424 (0.434)	-0.214 (0.322)
Teacher years taught: 8-13 years	-0.392 (0.371)	-0.526 (0.319)
Teacher years taught: 4-7 years	-0.558+ (0.322)	-0.756* (0.326)
(Teacher years taught: 1-3 years – ref)	0 (.)	0 (.)
Teacher years at school: missing data	0.495 (0.420)	-0.0795 (0.438)
Teacher years at school: 8-48 years	-0.129 (0.313)	0.489+ (0.268)
Teacher years at school: 4-7 years	0.351 (0.282)	0.333 (0.221)
(Teacher years at school: 1-3 years – ref)	0 (.)	0 (.)
Cut 1: Constant	-6.162 (4.586)	-7.276* (3.602)

Cut 2: Constant	-3.348 (4.550)	-5.486 (3.565)
Cut 3: Constant	-2.029 (4.548)	-4.132 (3.560)
Cut 4: Constant	-0.224 (4.572)	-2.628 (3.564)
N	440	465

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.'