The association between within-class ability grouping and children's achievement in Year 2, Year 5 and Year 9

Evaluation Study Plan

Evaluator (institution): UCL Institute of Education

Principal investigator: John Jerrim

Education Endowment Foundation

Template Last Reviewed: October 2019

PROJECT TITLE	The association between within-class ability grouping and children's achievement in Year 2, Year 5 and Year 9.
EVALUATOR (INSTITUTION)	UCL
PRINCIPAL INVESTIGATOR(S)	John Jerrim
STUDY PLAN AUTHOR(S)	John Jerrim
STUDY DESIGN	Regression analysis
PUPIL AGE RANGE AND KEY STAGE	Year 2; Year 5; Year 9
NUMBER OF SCHOOLS	Year 2 = N/A. Year 5 ≈ 147; Year 9 ≈ 143.
NUMBER OF PUPILS	Year 2 Maximum 8,876[1]. Year 5 ≈ 3,600; Year 9 ≈ 4,300.
PRIMARY OUTCOME MEASURE AND SOURCE	Year 2 = Progress in Maths and BAS II Word reading test scores (MCS) Year 5 = TIMSS mathematics test scores (TIMSS database) Year 9 = TIMSS mathematics test scores (TIMSS database)
SECONDARY OUTCOME MEASURE AND SOURCE	Year 2 = Enjoyment of mathematics (MCS database) Year 5 = Mathematics self-confidence (TIMSS database) Year 9 = Mathematics self-confidence (TIMSS database)

Study Plan version history

VERSION	DATE	REASON FOR REVISION
1.0 [*original*]	October 2019	
2.0	November 2019	Updated following advisory board meeting

[1] A total of 8,876 teachers across England, Wales, Northern Ireland and Scotland completed the age 7 teacher survey. This is therefore the maximum number of observations available for analysis. The final sample size will be smaller due to the sample selection rules imposed.

2.1	December 2019	Updated following EEF comments.

Table of contents

Contents

Background and study rationale

In almost every education system across the world, children are separated into different 'ability' (or achievement) groups at some stage. Yet this is done at different ages and in different ways. In some countries, children are divided into different schools based upon their achievement (between-school tracking), with a domestic example being the continuing prevalence of grammar schools within certain parts of the UK. Alternatively, schools may not use children's prior achievement as an entry criteria, but then divide pupils into different ability groups within school. This can take the form of streaming (dividing children into the same ability group for all subjects), setting (dividing children into ability groups on a subject-by-subject basis) or within-class ability grouping (e.g. children of similar ability being placed on the same table together). Of course, teachers and schools can (and do) use a combination of these approaches, so that children are first 'set' for a given subject, and then further divided by their ability within their class. Yet, despite the fundamental importance of this issue, little is currently known about *why* some teachers decide to group children by ability within their classes and the association that this has with their subsequent academic achievement. This is particularly true within the context of the English education system; a nation where within-school ability grouping is extensive (Education Datalab 2019). The aim of this report is to start to address this gap in the literature, providing new evidence on the use of within-class ability grouping within Year 2, Year 5 and Year 9 classes within England, and the association that this has with children's skills (particularly in mathematics).

Why might teachers and schools choose to group pupils by ability? Proponents of ability grouping suggest that it allows teachers to set different tasks to different groups of pupils, which most appropriately matches their existing knowledge and skills (Coe et al 2008). This then enables teachers to stretch the highest-achieving pupils with more challenging material, while lower-achieving children develop firmer foundations in the basics to ensure that they are not left behind (Francis et al 2017). It may also enable teachers to deploy their resources in what they believe to be an effective manner (e.g. to allocate class teaching assistants to provide extra support to the groups in most need). Within-class ability grouping may also make the workload of teachers more manageable, including in the management of disruptive pupils (Smith and Sutherland 2003).

In contrast, those who argue against the use of ability grouping suggest that there is a lack of empirical evidence to support the claims made above (Taylor et al 2017). Rather, it unhelpfully leads to some children being labelled as low-achievers, leading them to develop negative attitudes towards school or in a particular subject (Archer et al 2018), with their academic achievement suffering as a result. This may be supported by evidence from the peer-effects literature, suggesting that low-achieving pupils tend to make more academic progress when they are placed in classes with higher-achieving students (Burke and Sass 2013).

Despite the above, existing evidence on the impact of ability grouping upon pupils' outcomes remains limited. Indeed, as the Education Endowment Foundation (EEF) toolkit notes, the existing evidence on the impact of within-class ability grouping is 'limited'. There is a particular lack of studies comparing the impact of within-class ability group grouping to within-class grouping of pupils not based upon attainment.

This project will use data from the Millennium Cohort Study (Year 2 in 2008/09; n ≈ 8,000) and the Trends in Mathematics and Science Study (Year 5 and Year 9 in 2015; n ≈ 4,000 for each) to investigate the association between within-class ability grouping, children's achievement (primarily in mathematics) and their academic self-confidence. It will use regression analysis, including a rich set of controls for children's demographic backgrounds, prior achievement, teacher, class and school characteristics. To interpret estimates from this method as causal, a 'selection upon observables' assumption must hold (i.e. that models we estimate include all factors that influence both the decision of teachers to use within-class ability grouping and children's outcomes). As this is unlikely to be the case, it is more appropriate to treat the estimates presented in this report as conditional associations, rather than capturing cause and effect.

Impact evaluation

Research questions

1. To what extent do Year 2, Year 5 and Year 9 teachers in England use within-class ability grouping? How does this vary by school, teacher, classroom and pupil characteristics?
2. Is within-class ability grouping associated with higher levels of achievement?
3. Is being placed in a higher within-class ability group associated with greater academic progress? (Year 2 pupils only).
4. Is there an association between within-class ability grouping and children's academic self-confidence (Year 5 and Year 9) and enjoyment of mathematics (Year 2)?
5. Does the association between within-class ability grouping and children's achievement differ depending upon children's family background, prior academic achievement and whether the school sets/streams pupils?

Table 1: Design overview

Design		Regression analysis
Unit of analysis (school, pupils)		Pupils
Number of Units to be included in analysis (Intervention, Comparison)		Year 2 Maximum 8,876[2]. Year 5 ≈ 3,600; Year 9 ≈ 4,300. ('Treatment' / 'control' split will not be known until the analysis begins).
Primary outcome	Variable	Year 2 = Maths attainment Year 5 = Maths attainment Year 9 = Maths attainment
	measure (instrument, scale, source)	Year 2= Progress in Maths test, 0 – 28, MCS, variable = MATHS7SC[3]. Year 5/9= TIMSS mathematics test, 0 -1,000; TIMSS; variable = SMMAT01- SMMAT05
Secondary outcome(s)	variable(s)	Year 2 = enjoyment of mathematics. Year 5 / 9 = confidence in mathematics.
	measure(s) (instrument, scale, source)	Year 2 = Single question asked in the MCS – "how much do you enjoy number work?", Three-point scale (converted to binary), MCS questionnaire, variable = dcsc0023 scores,10 - 222, MCS; variable = DCWRAB00.

[2] A total of 8,876 teachers across England, Wales, Northern Ireland and Scotland completed the age 7 teacher survey. This is therefore the maximum number of observations available for analysis. The final sample size will be smaller due to the sample selection rules imposed.
[3] For further details see https://cls.ucl.ac.uk/wp-content/uploads/2017/06/Data-Note-20131_MCS-Test-Scores_Roxanne-Connelly-revised.pdf

		Year 5 / 9 = Students Confidence in Mathematics scale, 3-14, TIMSS questionnaire
Baseline for primary outcome	**variable**	Year 2 = The following prior achievement measures will all be included as separate covariates in the model: Naming vocabulary; pattern construction; picture similarities and Foundation Stage Profile scores (where available). Year 5 = The following prior achievement measures will be included as separate covariates in the model. Key Stage 1 scores in mathematics and English. Year 9 = The following prior achievement measures will be included as separate covariates in the model. Key Stage 2 scores in mathematics and English.
	measure (instrument, scale, source)	<u>**Year 2**</u> Naming vocabulary. BAS II sub-scale at age 5, 20-80, MCS, variable = cdnvtscr. Pattern construction scores. BAS II sub-scale at age 5,20-80, MCS, variable = cdpctscr. Picture similarities. BAS II sub-scale at age 5,20-80, MCS, variable = cdpstscr. Foundation Stage Profile communication, language and literacy, 0-36, MCS, variable = cll. Foundation Stage Profile mathematical development scores, 0-27, MCS, variable = md. <u>**Year 5**</u> Key Stage 1 maths (3-39) and reading/writing (3-27) as two covariates (NPD), variables = KS1_MATPOINTS and KS1_READWRITPOINTS <u>**Year 9**</u> Key Stage 2 maths (0-100) and reading (0-50) scores as two covariates (NPD), variables KS2_MATTOTMRK and KS2_READMRK
Baseline for secondary outcome	**Variable**	As above for primary outcome.
	measure (instrument, scale, source)	

Participants

MCS

The Millennium Cohort Study (MCS) is a rich, nationally-representative longitudinal study of UK children. A stratified, clustered survey design was used, with geographic areas (electoral wards) selected as the primary sampling unit, and then households with newly born children randomly selected from within (see Plewis 2004 for further details). Six sweeps have been conducted between 2000/01 and 2015, when children were nine months, 3, 5, 7, 11 and 14 years old. Parents, children and their teachers have been interviewed within the various sweeps. In total, 18,819 cohort members participated in the first survey, when children were nine months old (11,695 in England). This project draws upon data from the age 5 and age 7 survey sweeps, conducted in 2006 and 2008. By the time of the age 7 survey, 13,797 children (8,882 in England) remained. The focus of this paper is the prevalence and 'impact' of within-class ability grouping at age 7, with this information provided by their Year 2 teachers (who as part of the MCS also completed a questionnaire). A total of 8,876 teachers completed this questionnaire (5,627 in England)[4]. Moreover, 618 children in England and 270 in Wales completed the MCS fourth wave survey when they were in Year 3, and so have been excluded for the analysis. This part of the analysis will draw upon data from across the whole of the UK, where possible.

TIMSS 2015

The Trends in Mathematics and Science Study (TIMSS) was conducted in May/June 2015. The focus of this project is the data collected for England. A sample of pupils from two year groups in England participated in TIMSS (Year 5 and Year 9), with this including a mathematics and science test, a background questionnaire answered by pupils, a questionnaire answered by teachers and a questionnaire answered by headteachers. Year 5 pupils complete a 72-minute paper-based test, roughly split evenly between science and mathematics. The length of the test for Year 9 pupils was slightly longer (90-minutes).

TIMSS tests participants' knowledge, understanding and application of an international curricula, including aspects such as 'number', 'geometric shapes and measures' and 'data display' in mathematics. The fact that TIMSS attempts to measure children's knowledge and skills of an *internationally determined* curricula means that not all of the test questions covered in the TIMSS test are also taught within any given country's national curricula. However, this is less of an issue in the case of England, with around 90 percent of TIMSS mathematics questions covered by the national curriculum. The test makes extensive use of multiple-choice items, with example questions available from https://timssandpirls.bc.edu/timss2015/downloads/T15_FW_AppB.pdf.

In England, a total of 147 primary schools and 143 secondary schools from England were randomly selected to participate in the study. If schools refused to participate, then a substitute school could take their place. Within each of these schools, either one of two classes were then randomly chosen to participate. The final response rates were 98% for schools and 98% for pupils (Year 5), with analogous figures being 97% (schools) and 97% (pupils) for the Year 9 TIMSS study. This yielded a final total sample size of 8,820 pupils (4,006 in Year 5 and 4,814 in Year 9).

[4] The MCS has also been linked to the National Pupil Database (NPD) in England and Wales. This data is only available for young people in England and Wales, and only for those for whom administrative education data has been successfully linked.

The following criteria will be used to exclude a limited number of observations from the analysis, where key pieces of information are missing:

- Children will be excluded if their teacher has not responded to the mixed/same ability grouping questions.
- Children will be excluded if data is not available about their prior achievement (Key Stage 1 scores missing for Year 5 pupils and Key Stage 2 scores missing for Year 9 pupils).

Outcome measures and other data

Baseline (prior achievement) measures

Throughout the analysis, multiple prior achievement measures will be included within the regression model as separate covariates. This will be done for two reasons. First, skills in different academic areas may be associated with skill development in mathematics. For instance, children who have stronger reading skills may improve their mathematics skills at a faster rate than those with weaker reading skills (over and above their prior achievement in mathematics). Secondly, all prior achievement measures are likely to be measured with error, and hence can only partially account for potential confounding within the subject that they are meant to be measuring. The inclusion of additional controls for prior achievement will help to soak up such residual confounding resulting from baseline skill mismeasurement.

MCS

The following prior achievement measures will be included as separate covariates in the model. Further details about each measure is available in Connelly (2013).

- Naming vocabulary measured at age 5 using the BAS II sub-scale (variable name = cdnvtscr). This is designed to measure children's expressive verbal ability. In the test, the child is shown a series of pictures of objects and asked to name them.
- Pattern construction measured at age 5 using the BAS II sub-scale (variable name = cdpctscr). This is designed to measure children's spatial problem solving. In the test, the child is asked to replicate a design using patterned squares.
- Picture similarities measured at age 5 using the BAS II sub-scale (variable name = cdpstscr). This tests measures children's non-verbal reasoning. In the test, the child is shown a row of four pictures and is asked to identify a further congruent picture.
- Foundation Stage Profile communication, language and literacy scores. This is available for children in England only. (variable name = cll).
- Foundation Stage Profile mathematical development scores. This is available for children in England only (variable name =md).

TIMSS Year 5

The following prior achievement measures will be included as separate covariates in the model.

- Key Stage 1 maths scores, entered in the model as a set of dummy variable (variable name = KS1_MATPOINTS).
- The average of Key Stage 1 reading/writing scores (variable name = KS1_READWRITPOINTS)

TIMSS Year 9

The following prior achievement measures will be included as separate covariates in the model.

- Key Stage 2 maths scores (variable name = KS2_MATTOTMRK).
- Key Stage 2 reading scores (variable name = KS2_READMRK).

Primary outcome

MCS

There are two potential sources of information about cohort member's academic achievement around age seven. The first is their Key Stage 1 scores in mathematics, which have been linked into the MCS from the NPD. These have the advantage of being widely used and understood measures of national achievement that are routinely used in secondary academic research. They are, however, only available for cohort members living in England and Wales. Moreover, within these two countries, some children are missing Key Stage 1 data (e.g. due to non-consent or NPD linkage not being possible) and only broad levels (rather than fine-grained scores) are available within the MCS. These data are also only available via the UK Data Service secure lab, with a likely six-month time lag likely in order to gain access.

The second is cohort member's scores on a maths test taken as part of the MCS survey:

- The Progress in Mathematics test. This test was designed to measure children's mathematic skills and knowledge. It involved them completing a series of basic calculations, covering numbers, shapes measurement and data handling. A specially designed version of the PiM was developed for the MCS, which required children to complete fewer test items than the full PiM test. All children answered an initial set of seven test questions. They were then assigned to a set of four further easy, medium or hard questions, based upon their performance on the initial seven questions. A Rasch model has then been used to derive ability scores for each cohort member. Although the test has been used in previous EEF evaluations (https://educationendowmentfoundation.org.uk/public/files/Support/EEF_Research_Papers/Research_Paper_1_-_Properties_of_commercial_tests.pdf) we are unaware of any previous work that has investigated the psychometric properties of the specific version used in the MCS.

This test has the advantage of being available for a larger number of cohort members and were administered and marked by independent members of the MCS survey team. On the other hand, it is a relatively short assessment and is 'low-stakes' (pupils, teachers and schools have nothing riding upon the results). The correlation between scores on these tests and Key Stage 2 scores for MCS cohort members is around 0.6.

Based upon the above, performance in the Progress in Mathematics test will be used as the primary outcome measures. Throughout the analysis, scores will be standardised to mean zero and standard deviation one (using the standard deviation calculated using the full sample with data available). All estimates will hence be presented in terms of an effect size.

TIMSS

TIMSS uses a multiple matrix-sampling approach to the test design, meaning that each child is randomly assigned a subset of all test items. A complex latent regression model is then used by the survey organisers to derive estimates of pupils' achievement (known in the psychometrics literature as 'plausible values'), based upon how children responded to the test question and their background characteristics. These plausible values are essentially multiple imputations of children's latent ability in science and mathematics, and capture the uncertainty in estimates of children's achievement in these two subject areas. Unless otherwise stated (e.g. in table noes), I follow recommended practise in the use of these plausible values, estimating models five times (once using each plausible value) and then combining these into the final estimate following Rubin's Rules. For further details about the plausible values and the TIMSS test design, see Jerrim et al (2017) and Martin, Mullis and Hooper (2016)[5].

The primary outcome will be children's TIMSS scores in mathematics. I focus upon mathematics – rather than science – due to this subject being assessed for all pupils within the Key Stage 1 and Key Stage 2 tests (thus, critically, providing measures of prior achievement used within the analysis models, as outlined below).

Secondary outcomes

<u>MCS</u>

The MCS age 7 survey completed by children includes a range of questions. This study will make use of the following which was answered using a three-point scale (1. Don't like it; 2. like it a little; 3. like it a lot):

- How much do you enjoy number work? (primary outcome)

Responses to these two questions will be converted into binary format (0 = don't like it and like it a little; 1 = like it a lot).

<u>TIMSS</u>

As part of the background questionnaire, Year 5 and Year 9 pupils are asked a range of questions about their attitudes towards mathematics. This includes the following nine questions, asked on a four-point scale ('agree a lot' to 'disagree a lot'), capturing their self-confidence in mathematics:

- *I usually do well in mathematics*
- *Mathematics is harder for me than for many of my classmates*
- *I am just not good at mathematics*
- *I learn things quickly in mathematics*
- *Mathematics makes me nervous*
- *I am good at working out difficult mathematics problems*
- *My teacher tells me I am good at mathematics*
- *Mathematics is harder for me than any other subject*
- *Mathematics makes me confused*

The survey organisers have used pupils' responses to these questions to create the Students Confidence in Mathematics scale, which is available within the public use TIMSS database (variable = ASBGSCM / BSBGSCM). Cronbach's alpha for this scale within England is reported to be 0.87 for Year 5 pupils and 0.89 for Year 9 pupils. This scale will be used as a secondary outcome within our analysis.

[5] Note that TIMSS samples pupils within classrooms and not teachers per se. This means that the population of interest in TIMSS – who inferences refer to – are pupils rather than teachers.

Selection Mechanism

Why might teachers choose to group pupils within their class by ability?

Little is currently known about why some teachers choose to teach children within same ability groups and others choose to teach within mixed ability groups (Francis, Taylor and Tereshchenko 2019). The first research question has hence been designed to provide new evidence on this issue in terms of observable child, teacher and class characteristics, in what will be an exploratory analysis. Apriori, I nevertheless expect the following two factors to be key.

First, teachers choosing to group children by ability is likely to be a reaction to the heterogeneity of their class. If it is very diverse, including children from a wide range of backgrounds and abilities, then some form of within-class ability grouping may be a natural response. This then relates to other school policies, such as setting and streaming; classes are likely to be less diverse in those schools where some form of higher-level ability grouping has already taken place. Likewise, it is also likely to be related to the diversity of a school's intake; ability grouping (whether within classes, between classes, or both) may be more appealing to teachers and school leaders when the student body is more diverse.

Second, within-class ability grouping may be linked to individual teacher characteristics. For instance, more experienced teachers, or those with a specialisation in the subject they teach, may be more willing or confident to teach pupils in a certain way (e.g. to teach using mixed-ability groups). On the other hand, teachers may have been trained to use a certain approach within their initial training, and have continued to use this approach throughout their career. Relatedly, it could be school policy (or the advice of a mentor) that pupils should be grouped together in particular ways.

What determines the within-class ability group to which children are assigned?

Within classes where children are grouped by ability, what is likely to determine their allocation to the top, middle or bottom group? In a perfectly meritocratic world, group allocation would be solely determined by two factors: (a) the child's own academic ability; (b) the academic ability of their class peers. Hence their *absolute* and *relative* academic ability would be the only thing that matters.

Yet, in reality, other factors are likely to determine group allocation. For instance, it may be that teachers' *perceptions* of a child's ability (rather than their objectively measured ability) is the more important factor. This could be influenced by their views of the child's motivation, ethnicity and socio-economic background (Muijs and Dunne 2010). Likewise, there could be other influences upon the group the child is assigned to, such as the 'pushiness' of parents or the preferences of a particular child (e.g. whether they would prefer to sit with their friends). Indeed, much research evidence (though mostly in the context of between-school or between-class tracking) has suggested that ability-group allocation is determined by factors outside of children's (measured) academic achievement, with those from lower socio-economic status and ethnic minority backgrounds being more likely to be allocated to lower-sets (Connolly et al 2019; Muijs and Dunne 2010).

Selection of the comparison group

MCS

The age 7 MCS teacher survey gathered a range of information about cohort member's primary school teachers, including their background characteristics, their views of the cohort member's abilities[6], their teaching style and the organisation of their class. This encompassed details about setting/streaming in the school, how class time is allocated to whole class teaching, group work and individual work, and – critically – detailed information about within-class ability grouping. The latter was captured as follows.

To begin, teachers were provided with definitions of within-class ability grouping and within-subject ability grouping (the precise wording is provided in Appendix A). They were then asked:

- '*In this child's class, is there within-class ability grouping*' (yes / no).

To those teachers who selected yes, they were then asked:

- '*How many within-class ability groups are there*?' (open field)
- '*Which group is this child in*?' (highest, middle, lowest)

All teachers were then asked analogous questions specifically about within-subject ability groupings in literacy and mathematics. For example:

- '*In this child's class, are there within-class subject groups for literacy*' (yes / no).
- '*How many within-class subject groups are there for literacy*' (open field)
- '*Which group is this child in for literacy*' (highest, middle, lowest)

The MCS survey organisers have used teachers' responses to these questions to derive a set of variables, which have been deposited as part of the MCS dataset. These are listed below, and are the key covariates of interest within our analysis:

- wicagn = The within-class ability group of the child in mathematics (not grouped; top group; middle group; bottom group).
- wicagl = The within-class ability group of the child in literacy (not grouped; top group; middle group; bottom group).

Specifically, based upon these variables, the paper will investigate the 'impact' of being (a) grouped within class versus not being grouped within class and (b) for those who are grouped within class, the 'impact' of being allocated to the top, middle and bottom group.

TIMSS

As part of the TIMSS questionnaire, teachers were asked:

"*In teaching mathematics to the students in this class, how often do you ask pupils to do the following*"

- Work problems together in the whole class with direct guidance from me
- Work in mixed ability groups
- Work in same ability groups

With four possible response options (1. Every or almost every lesson; 2. About half the lessons; 3. Some lessons; 4. Never). Note that, to avoid small sample sizes, the bottom two categories ('never' and 'some lessons') will be combined into a single group[7]. The primary

[6] Teachers only answered this with respect to the individual child; not for every child in the class.
[7] This decision was made based upon some initial pre-specified descriptive analysis of the TIMSS 2015 data for England.

exposure of interest in the analysis is the frequency with which mixed-ability grouping is used. The 'effect' of same-ability grouping upon children's achievement can also be compared to the 'effect' of (a) mixed-ability grouping and (b) whole class teaching. For further details, see the methodology section below.

Primary analysis

Recognising that little is currently known about the 'selection mechanisms' (i.e. why some teachers choose to teach in ability groups while others do not) the analysis will begin by looking at the characteristics of pupils who are exposed to within-class ability grouping and those who are not (as well as documenting the characteristics of teachers who are most likely to organise their class in this way).

The analysis will then attempt to answer the following primary research question:

- Is within-class ability grouping associated with higher levels of achievement?

Primary analysis of the MCS (Year 2 pupils)

First, I will report raw Progress in Maths test score outcomes for the 'treatment' (within-class ability grouping used) and 'control' (within-class ability grouping not used) groups. Second, I will estimate an OLS regression model of the form:

$$A_{ij} = \alpha + \beta . G_{ij} + \partial . C_{ij} + \delta . P_{ij} + \gamma . S_{ij} + \Delta . T_{ij} + \tau . C_{ij} + \varepsilon_{ij} \quad (1)$$

Where:

i = Child i.

j = Teacher / class j.

A_{ij} = The achievement of child i in the age 7 Progress in Mathematics test. This has been standardised (using the full age 7 MCS sample) to mean zero and standard deviation one.

G_{ij} = A dummy variable capturing whether ability grouping is used in the child's class (1) or not (0).

C_{ij} = A vector of children's demographic background characteristics, including gender, permanent family income, age in months, ethnicity, language spoken at home and maternal and paternal education.

P_{ij} = Prior achievement of child i, as measured by performance in the MCS age 5 reading test, pattern construction scores, picture similarity scores and children's Foundation Stage Profile scores (where available)[8].

S_{ij} = Whether the teacher reports that setting / streaming is used for the subject in the child's school, and the set to which the child has been assigned (bottom, middle, top).

T_{ij} = A vector of teacher characteristics. This includes their gender, years of teaching experience, qualifications, years working at the school.

C_{ij} = A vector of class characteristics, as reported by the teacher. This includes class size, whether the class contains mixed year groups, percent of children in class with SEN, percent of children in class with EAL, number of days taught by a supply teacher, class time spent on

[8] Foundation Stage Profile scores are only available for children in England. They are therefore treated as missing data (and hence imputed) for children in other parts of the UK.

literacy / numeracy per week, whether class gets regular support from a teaching assistant, special needs or other teacher, percentage of lesson time devoted to group work per week.

ε_{ij} = The error term. Standard errors will be calculated having accounted for the stratification and clustering in the MCS survey design, following the recommendations of the MCS documentation (see Ketende and Jones 2011).

The parameter of interest from this model is β; this illustrates the difference in (standardised) Progress in Maths scores between children who have been grouped by ability within their class to those who have not been grouped by ability (conditional upon other variables included in the model). Multiple imputation using chained equations is used to account for missing covariate data, with the age 7 response weight (variable = DDOVW200) applied. Standard errors will be reported, having been adjusted for the stratification and clustering within the MCS survey design – as it recommended within the technical documentation (see Ketende and Jones 2011). All estimates will be presented in terms of an effect size.

<u>TIMSS</u>

The analysis will begin by presenting a series of descriptive statistics, illustrating the characteristics of pupils who are exposed to (a) same ability grouping; (b) mixed ability grouping and (c) whole class teaching within their mathematics classes. Cross-tabulations will also be presented, documenting the association between mixed-ability grouping, same-ability grouping and whole class teaching used in mathematics classes in England. Some initial pre-specified descriptive statistics, conducted at the teacher level, can be found in Appendix B[9].

Next, the following OLS regression model will be estimated, comparing the "effect" of same-ability grouping to the "effect" of whole class teaching:

$$A_{ijk} = \alpha + \beta.SA_{ij} + \gamma.WC_{ij} + \emptyset.P_i + \delta.Ch_i + \nabla.C_{ij} + \varphi.T_j + \omega.Set_k + \theta.Sch_k + \varepsilon_{ijk} \quad (2)$$

Where:

i = child i

j = teacher j

k = school k

A_{ijk} = The TIMSS mathematics plausible values, capturing children's achievement in this subject. These will each be standardised to mean zero and standard deviation one within England (using the full sample standard deviation). All model estimates will therefore be presented in terms of effect sizes.

SA_{ij} = The variable capturing the frequency that same-ability teaching is used within the child's mathematics class.

WC_{ij} = The variable capturing the frequency that whole-class teaching is used within the child's mathematics class.

P_i= Measures of children's prior achievement. For Year 5 pupils, this will be their scores in their Key Stage 1 English and mathematics test. For Year 9 pupils, this will be their Key Stage 2 English and mathematics test scores.

[9] It was agreed with the study advisory board that this limited initial analysis should be conducted in order to help inform the contents of this study plan.

Ch_i = A vector capturing children's background characteristics. This will include gender, FSM status, home economic resources index, immigrant status, absence from school, age in months.

T_j = A vector of teacher characteristics that may confound the relationship between the use of within-class ability grouping and children's achievement. This will include key measures of observable teacher quality that have previously been identified in the academic literature, including (a) subject specialism in mathematics and (b) number of years teaching experience. It will also include characteristics of the class: (c) minutes per week spent teaching mathematics to the class; (d) average prior maths achievement levels of the class (based upon Key Stage scores); (e) standard deviation of the prior maths achievement of the class[10].

Set_k = A dummy variable measured at the school level for whether it is school policy to set/stream pupils into different mathematics classes.

Sch_k = A school-level control for headteacher reports of the percent of pupils who come from advantaged/disadvantaged backgrounds

The parameters of interest from this model are β and γ. The former (β) illustrates whether more frequent use of same-ability grouping by teachers is associated with greater academic progress made in mathematics. The latter (γ) has a similar interpretation with respect whole class teaching. By comparing β to γ, we can establish which approach is more strongly associated with the progress children make in mathematics.

An analogous model will then be estimated – including the same set of controls – to compare the impact of same-ability grouping to mixed-ability grouping:

$$A_{ijk} = \alpha + \beta.SA_{ij} + \gamma.MA_{ij} + \emptyset.P_i + \delta.Ch_i + \nabla.C_{ij} + \varphi.T_j + \omega.Set_k + \theta.Sch_k + \varepsilon_{ijk} \quad (3)$$

Where:

MA_{ij} = The variable capturing the frequency that mixed-ability teaching is used within the child's mathematics class.

A comparison of the β and γ parameters will now reveal whether same-ability or mixed-ability grouping is more strongly associated with mathematics achievement.

To account for the complex TIMSS survey design, the recommended practise of the survey organisers will be followed. Specifically, the final TIMSS pupil and replication weights will be applied via the Stata package 'pv' (Macdonald 2008). This accounts for the hierarchical nature of the data, including the nesting of pupils within classes and schools, making the appropriate adjustment to the estimated standard errors. Likewise, following recommended practise in the use of plausible values, all models will be estimated five times (once using each plausible value) with final parameter estimates and standard errors combined according to Rubin's multiple imputation rules (Rubin 1987).

<u>Multiple hypothesis testing</u>

A Bonferroni correction will be used to adjust for multiple hypothesis testing. As three independent datasets will be used, with three primary outcomes (mathematics achievement in each of the three datasets) a Bonferroni correction factor of three will be applied.

[10] The average and standard deviation of prior achievement of the class will be based upon Key Stage 1 maths scores (Year 5) and Key Stage 2 maths scores (Year 9).

Robustness checks

MCS

Within certain elements of the analysis, a lack of 'common support' may be a concern. For instance, when investigating whether ability group allocation matters, it is possible that there are no sufficiently comparable children in the top and bottom groups. The main advantage of PSM over regression is that common support can be enforced on the analytic sample, thus ensuring that there are comparable children in the 'treatment' (e.g. top ability group) and control (e.g. bottom ability group) groups. Propensity score matching (PSM) is therefore used as an alternative to test the robustness of our regression results.

The PSM models will match children using one-to-one nearest neighbour matching, with the caliper set to 0.05. The first-stage selection model will include the same covariates as outlined for the regression models in equation (1). Multiple imputation will not be used to account for missing covariate data; listwise deletion or 'missing' dummy flags will be used instead. Post-estimation, balance of the covariates included in the selection model will be compared between each of the different groups. Average mathematics test scores for the matched sample will then be presented, with differences between the groups taken as the 'treatment effect'.

Finally, due to a delay with the fieldwork, the MCS questionnaire was sent to teachers relatively late in the academic year. Consequently, some teachers completed the questionnaire after the child had progressed for Year 2 to Year 3, which could introduce problems with recall. The primary analysis model presented in equation (1) will therefore be re-estimated using the sub-sample of teachers who completed the questionnaire before the start of the new academic year (i.e. before September 2008).

TIMSS

To test the robustness of the results, the models presented in (2) and (3) will be re-estimated including some additional teacher and school level controls. The motivation for not including these controls in the headline model specification is due to some concerns of overfitting the data (given the limited number of teachers in TIMSS).

Specifically, the following additional variables will be included:

- Class size.
- Teachers confidence in teaching mathematics.
- School shortages in mathematics
- Language spoken at home.

The second robustness test will be to use multiple imputation to account for the limited amount of missing covariate data, which mostly stems from non-linkage to the NPD or teacher non-response to the background questionnaire. This will be implemented using multiple imputations using chained equations in Stata, with the random seed set to a value of 5000, to produce five imputed datasets. The analysis will then be re-run with these imputed data, using the first plausible value in mathematics as the outcome, and making a Huber-White adjustment to the estimated standard errors at the school level.

Further analyses

Secondary outcome analyses

MCS

The MCS age 7 survey completed by children includes a range of questions. This study will make use of the following which was answered using a three-point scale (1. Don't like it; 2. like it a little; 3. like it a lot):

- How much do you enjoy number work? (primary outcome)

Responses to this question will be converted into binary format (0 = don't like it and like it a little; 1 = like it a lot). The same analysis model will then be estimated as for the primary analysis, but will be estimated using logistic regression rather than ordinary least squares. Specifically, the following model will be estimated:

$$E_{ij} = \alpha + \beta.G_{ij} + \partial.C_{ij} + \delta.P_{ij} + \gamma.S_{ij} + \Delta.T_{ij} + \tau.C_{ij} + \varepsilon_{ij} \quad (4)$$

Where:

E_{ij} = Whether the child enjoys doing number work a lot (1) or like it a little/don't like it (0)

All other variables included are as defined for equation (1).

The parameter of interest (β) will illustrate whether children who have been grouped by ability within their class are more likely to enjoy number work than those who have not been grouped by ability within their class (conditional upon other variables included in the model). Results from this model will be reported as odds-ratios.

Moreover, it is also possible to answer the following secondary outcome research question in the MCS:

- Is being placed in a higher within-class ability group associated with greater academic progress? (Year 2 pupils only).

To answer this question, the sample will be restricted to children where within-class ability grouping is used. An OLS regression model will be estimated, specified as:

$$A_{ij} = \alpha + \beta.G_{ij} + \partial.C_{ij} + \delta.P_{ij} + \gamma.S_{ij} + \Delta.T_{ij} + \tau.C_{ij} + \varepsilon_{ij} \quad (5)$$

Where:

G_{ij} = A dummy variable capturing the child's within-class ability group; low, middle (reference) or high.

And all other variables are specified as in equation (1). The β estimates will now capture whether children assigned to higher within-class ability groups have higher levels of achievement than children assigned to lower within-class ability groups – conditional upon the covariates (including the numerous measures of prior achievement) included in the model.

This will be supplemented using a child fixed-effects model, capturing within-child, between-subject (English versus mathematics) ability group allocation. To begin, the sample will be restricted to children where within-class ability grouping was used. These models will exploit the fact that children may be allocated to different ability groups in different subjects (e.g. the middle group in English versus the top group in mathematics). This between-subject variation potentially allows one to get closer to a 'causal' effect than regression/matching techniques alone. Specifically, such models implicitly control for all factors that are constant within a child (e.g. innate ability, socio-economic background, school), with the focus now children's ***relative*** performance in English versus mathematics (and how this relates to their ability group allocation). The specification of these fixed-effects models are as follows:

$$A_{ij} = \alpha + \beta.G_{ij} + \gamma.S_{ij} + \delta.P_{ij} + \Delta.T_{ij} + u_j + \varepsilon_i \quad (6)$$

Where:

i = subject i (mathematics or English).

j = child j.

A_{ij} = Achievement test score in subject i of child j.

G_{ij} = Within-class ability group (low, middle, high) in subject i of child j.

S_{ij} = A vector of dummy variables capturing the set / stream (low, middle, high) in subject i of child.

P_{ij} = Prior achievement of child i in subject j, as measured by scores in the Foundation Stage profile[11].

T_{ij} = Amount of class time devoted to subject i in child's j class.

u_j = Child fixed-effect

ε_i = Random (within-child) error term.

The β parameter now captures how ability group allocation in mathematics/English is related to relative performance in English/mathematics at age 7. The model also includes controls for factors that potentially vary within subject for each child, such as set allocation and prior achievement in English and mathematics. Note that, when estimating these models, multiple imputation is not used (listwise deletion is applied instead) and the survey weight is not applied.

<u>TIMSS</u>

As part of the background questionnaire, Year 5 and Year 9 pupils are asked a range of questions about their attitudes towards mathematics. This includes the following nine questions, asked on a four-point scale ('agree a lot' to 'disagree a lot'), capturing their self-confidence in mathematics:

- *I usually do well in mathematics*
- *Mathematics is harder for me than for many of my classmates*
- *I am just not good at mathematics*
- *I learn things quickly in mathematics*
- *Mathematics makes me nervous*
- *I am good at working out difficult mathematics problems*
- *My teacher tells me I am good at mathematics*
- *Mathematics is harder for me than any other subject*
- *Mathematics makes me confused*

The survey organisers have used pupils' responses to these questions to create the Students Confidence in Mathematics scale, which is available within the publicly available TIMSS

[11] Only foundation stage profile scores are included in this model. This is because the model now focuses upon *within-subject* variation, and hence any controls must be subject specific. The foundation stage profile is the only direct source of prior achievement information in mathematics.

database (variable = ASBGSCM / BSBGSCM). Cronbach's alpha for this scale within England is reported to be 0.87 for Year 5 pupils and 0.89 for Year 9 pupils. This scale will be used as a secondary outcome within our analysis.

Sub-group analysis

MCS

There are three sources of heterogeneity that we are interested in: differences by socio-economic background, differences by prior achievement and differences by whether other forms of ability grouping (i.e. setting / streaming) are used within the school. These sub-groups will be defined as follows:

- Family background. Permanent family income will be the primary measure of family background used in the MCS analysis. This will be defined as follows. First, an average will be taken of the income measures collected in the first five MCS waves (using, for each child, all waves where data is available). This permanent income measure will then be divided into thirds (tertiles), to create low, average and high-income groups.
- Prior achievement. Children's scores on the age 5 MCS cognitive tests (naming vocabulary, pattern similarities and picture similarities) will be used to define children's prior achievement. To begin, scores on each of these tests will be standardised to mean 0 and standard deviation 1. The average will then be calculated for each child across these three test scores. This measure will then be divided into tertiles to create low (bottom third), average (middle third) and high (top third) prior achievement groups.
- Setting / streaming used in the school. As part of the MCS, teachers were asked about whether setting / streaming were used in their school and for what subjects. The survey organisers have derived a set of variables to indicate whether children are set/streamed in mathematics (variable = strnum). This variable will be used to identify sub-groups of children who have been grouped into different classes based upon their prior ability/achievement.

To explore sub-group effects, the same model as presented in (1) will be re-estimated for: (a) high-income pupils; (b) low-income pupils; (c) children with low prior achievement; (d) children with high prior achievement; (e) children who have been set/streamed; (f) children who have not been set/streamed.

TIMSS

Separate results will be presented for the following sub-groups:

- Children from a socio-economically disadvantaged versus non-disadvantaged background. This will be operationalised in TIMSS as the performance of children who are FSM eligible (disadvantaged pupils) and those who are not FSM eligible (non-disadvantaged pupils).
- Children will be divided into low, average and high prior achievement groups, based upon tertiles of their Key Stage 1 (Year 5) or their Key Stage 2 (Year 9) mathematics scores.
- The TIMSS Year 5 sample will be divided into schools where setting/streaming is used and those where it is not. This will enable investigation of the association between within-class ability grouping and children's achievement in settings where they have already been separated by ability into different classes (or not). Note that this analysis will not be conducted for the Year 9 TIMSS sample, as almost all secondary schools in England set / stream pupils in mathematics.

The main analysis model presented above will hence be estimated for the following sub-groups: (a) FSM; (b) non-FSM; (c) low prior achievement; (d) high prior achievement; (e) setting used in school; (f) setting not used in school.

Ethics

The ethics form for this project has been submitted to the UCL Institute of Education ethics committee. It has been provisionally approved, subject to minor adjustments.

Data protection

The main issue with respect to data protection is access to the TIMSS data linked to the NPD. The form submitted to the Department for Education is available upon request with further details.

Personnel

- John Jerrim, UCL Institute of Education.

Risks

The only risk to the project surrounds access to the NPD-TIMSS data. Although access to these data have been provided previously for academic research (e.g. https://epi.org.uk/wp-content/uploads/2018/01/World-Class-Standards-In-Primary-Report.pdf) this was before the introduction of the ONS Secure Research Service. The staff within the Department for Education International Statistics team has also changed. There is hence a risk with respect to the TIMSS-NPD data access, particularly the uncertainty how quickly this will be provided by the Department for Education. This is an area where support from the EEF, given its relationship with the Department for Education, may be needed.

The only thing that can be done to partially mitigate this risk is to commit to using the public use TIMSS data as a last resort. This can be freely downloaded straight from the TIMSS website, but will be missing some key variables (children's Key Stage 1, Key Stage 2 and FSM status variables). Although the analysis presented are still possible with the public use TIMSS data, the strength of the evidence resulting from the project will be weaker.

Timeline

Date	Activity	Staff responsible/ leading
12/11/2019	Advisory board meeting	Complete
30/11/2019	Complete TIMSS-NPD application	JJ
30/11/2019	Complete ethics form	Provisionally approved
19/12/2019	Complete study plan	JJ
30/03/2020	First draft of MCS analysis +interim report	JJ
30/06/2020	Final draft of MCS analysis + interim report	JJ

30/09/2020	First draft of TIMSS-NPD analysis + report	JJ
31/12/2020	Final draft of TIMSS-NPD analysis + report	JJ

References

Archer, L., Francis, B. , Miller, S. , Taylor, B. , Tereshchenko, A. , Mazenod, A. , Pepper, D. and Travers, M. (2018), The symbolic violence of setting: A Bourdieusian analysis of mixed methods data on secondary students' views about setting. *British Educational Research Journal*, 44: 119-140. doi:10.1002/berj.3321

Burke, M. & Sass, T. Classroom Peer Effects and Student Achievement. *Journal of Labor Economics* 31(1): 51-82.

Coe, R.; Jones, K.; Searle, J.; Kokotsaki, D.; Kosnin, A. and Skinner, P. 2008. 'Evidence on the effects of selective educational systems'. Accessed 04/10/2019 from https://www.suttontrust.com/wp-content/uploads/2008/10/SuttonTrustFullReportFinal11.pdf

Connolly, P. , Taylor, B. , Francis, B. , Archer, L. , Hodgen, J. , Mazenod, A. and Tereshchenko, A. (2019), The misallocation of students to academic sets in maths: A study of secondary schools in England. British Educational Research Journal 45: 873-897. doi:10.1002/berj.3530

Connolly, R. (2013). Interpreting test scores. Millennium Cohort Study Data Note 2013/1. Accessed 25/10/2019 from https://cls.ucl.ac.uk/wp-content/uploads/2017/06/Data-Note-20131_MCS-Test-Scores_Roxanne-Connelly-revised.pdf

Education Datalab. 2019. England's schools segregate by ability more than almost every other country in the world. Accessed 07/10/2019 from https://ffteducationdatalab.org.uk/2019/09/englands-schools-segregate-by-ability-more-than-almost-every-other-country-in-the-world/

Francis, R.; Louise Archer, Jeremy Hodgen, David Pepper, Becky Taylor & Mary-Claire Travers (2017) Exploring the relative lack of impact of research on 'ability grouping' in England: a discourse analytic account, Cambridge Journal of Education, 47:1, 1-17, DOI: 10.1080/0305764X.2015.1093095

Francis, R.; Taylor, R. and Tereshchenko, A. 2019. Reassessing ability grouping. Improving practice for equity and attainment. Routledge: Oxon.

Jerrim, J., Lopez-Agudo, L. A., Marcenaro-Gutierrez, O.D., & Shure, N. (2017). What Happens When Econometrics and Psychometrics Collide? An Example Using PISA Data. *Economics of Education Review, 61,* 51-58.

Ketende, S. and Jones, E. 2011. User Guide to Analysing MCS Data Using Stata. Accessed 07/10/2019 from https://cls.ucl.ac.uk/wp-content/uploads/2017/07/User-Guide-to-Analysing-MCS-Data-using-Stata.pdf

Macdonald, K. 2008. "PV: Stata module to perform estimation with plausible values," Statistical Software Components S456951, Boston College Department of Economics, revised 03 Feb 2019.

Martin, M. O., Mullis, I. V. S., & Hooper, M. (Eds.). (2016). *Methods and Procedures in TIMSS 2015*. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: **http://timssandpirls.bc.edu/publications/timss/2015-methods.html**

Muijs, D. & Dunne, M. (2010) Setting by ability – or is it? A quantitative study of determinants of set placement in English secondary schools, Educational Research, 52:4, 391-407, DOI: 10.1080/00131881.2010.524750

Rubin, D.B. (1987), Multiple Imputation for Nonresponse in Surveys, New York: John Wiley & Sons, Inc

Smith, C. and Sutherland, M. 2003. Setting or mixed ability? Teachers' views of the organisation of pupils for learning. Journal of Research in Special Educational Needs 3(3):141-146.

Taylor, R.; Becky Francis, Louise Archer, Jeremy Hodgen, David Pepper, Antonina Tereshchenko & Mary-Claire Travers (2017) Factors deterring schools from mixed attainment teaching practice, Pedagogy, Culture & Society, 25:3, 327-345, DOI: 10.1080/14681366.2016.1256908

Appendix – Template results tables

Table 1. Sample selection criteria used to define the analytic sample in the MCS

	UK	England
Initially sampled		
Participated at wave 1		
Participated at wave 4 (age 7)		
Teacher survey completed		
Within-class ability grouping data available		
Age 7 test score data available		
Final analytic sample		

Notes: Author’s calculations using the MCS data.

Table 2. Characteristics of children included and excluded from the MCS analytic sample

	In analytic sample		Not in analytic sample	
	n	% / mean	n	% / mean
Socio-economic background				
Family income				
% father hold degree				
% mother hold degree				
Prior cognitive ability				
Total FSP score (England only)				
Bracken school readiness score				
Age 5 pattern construction score				
Age 5 reading score				
Demographics				
% White				
% Male				

Notes: Raw number of observations (n) reported. Missing data has been excluded. MCS age 7 UK weight (variable = DDOVW200) applied.

Table 3. Characteristic of children included and excluded from the TIMSS analytic sample

	Year 5		Year 9	
	Included in analytic sample	Excluded from analytic sample	Included in analytic sample	Excluded from analytic sample
Child characteristics				
% male				
% FSM				
% immigrant				
% English not spoken at home				
% regularly absent from school				
Prior achievement				
Key Stage 1 / 2 maths score				
Key Stage 1 / 2 reading score				
Teacher characteristics				
% Subject specialism in maths				
Years of teaching experience				
Class characteristics				
Class size				
% students with language issues				
Minutes per week teaching class				
Average KS1/2 maths scores of class				
Standard deviation of class KS1/2 maths scores				
School characteristics				
Pupils set for mathematics				
% of pupils from disadvantaged backgrounds				
Resource shortages affecting maths instruction scale				
Observations				

Table 4. The characteristics of Year 2 children and teachers who use within-class ability grouping in mathematics (evidence from the MCS)

	Mathematics		English	
Within-class ability grouping used?	**Yes**	**No**	**Yes**	**No**
Child background				
Average family income				
% father hold degree				
% mother hold degree				
% White				
% Male				
Child's prior achievement				
Total FSP score (England only)				
Bracken school readiness score				
Age 5 pattern construction score				
Age 5 reading score				
Teacher characteristics				
Years of teaching experience				
Hold PGCE				
Hold other postgraduate degree				
Class characteristics				
Children set / streamed				
Class size				
Mixed year groups				
% EAL				
% of class time spent working in groups				
Teaching assistant in classroom				

Notes: Missing data has been excluded. MCS age 7 UK weight variable = DDOVW200) applied. Figures refer to either the mean (continuous variables) or the percentage (binary variables).

Table 5. The characteristics of Year 2 children within low, middle and high within-class ability groups (evidence from the MCS)

	Mathematics			English		
Within-class ability group?	**Low**	**Middle**	**High**	**Low**	**Middle**	**High**
Child background						
Average family income						
% father hold degree						
% mother hold degree						
% White						
% Male						
Child's prior achievement						
Total FSP score (England only)						
Bracken school readiness score						
Age 5 pattern construction score						
Age 5 reading score						

Notes: Missing data has been excluded. MCS age 7 UK weight (variable = DDOVW200) applied. Figures refer to either the mean (continuous variables) or the percentage (binary variables).

Table 6. OLS regression estimates of the relationship between whether Year 2 teachers use within-class ability grouping and children's achievement

(a) Primary outcomes (achievement)

	Mathematics scores		
	Effect	**SE**	**N**
Within-class ability grouping used?			
No (Reference group)			
Yes			

(b) Secondary outcomes

	Enjoy number work			Reading scores		
	Effect	**SE**	**n**	**Effect**	**SE**	**n**
Within-class ability grouping used?						
No (Reference group)						
Yes						

Notes: MCS age 7 UK weight (variable = DDOVW200) applied, with clustering and stratification in the MCS survey design used to adjust the estimated standard errors. Multiple imputation used to account for missing covariate data. Models include the following controls: gender, permanent family income, ethnicity, age in months, language spoken at home and maternal and paternal education, foundation stage profile English and maths scores, MCS cognitive test scores at age 5, teacher gender, years of teaching experience, teacher qualifications, years teacher working at the school, whether children set/streamed in the subject, class size, whether the class contains mixed year groups, percent of children in class with SEN, percent of children in class with EAL, number of days taught by a supply teacher, class time spent on literacy / numeracy per week, whether class gets regular support from a teaching assistant, special needs or other teacher, percentage of lesson time devoted to group work per week

Table 7. OLS regression estimates of the relationship between within-class ability group assignment (low, middle, high) and children's achievement

(a) Primary outcomes (achievement)

	Mathematics		
	Effect	SE	n
Within-class ability group			
Low			
Middle (reference group)			
High			

(b) Secondary outcomes

	Enjoy number work			Reading scores		
	Effect	SE	n	Effect	SE	n
Within-class ability group						
Low						
Middle (reference group)						
High						

Notes: MCS age 7 UK weight (variable = DDOVW200) applied, with clustering and stratification in the MCS survey design used to adjust the estimated standard errors. Multiple imputation used to account for missing covariate data. Models include the following controls: gender, permanent family income, ethnicity, age in months, language spoken at home and maternal and paternal education, foundation stage profile English and maths scores, MCS cognitive test scores at age 5, teacher gender, years of teaching experience, teacher qualifications, years teacher working at the school, whether children set/streamed in the subject, class size, whether the class contains mixed year groups, percent of children in class with SEN, percent of children in class with EAL, number of days taught by a supply teacher, class time spent on literacy / numeracy per week, whether class gets regular support from a teaching assistant, special needs or other teacher, percentage of lesson time devoted to group work per week

Table 8. Child fixed-effects estimates of the relationship between within-class ability group assignment (low, middle, high) and children's achievement

	Effect	SE	N
Within-class ability group			
Low			
Middle (reference group)			
High			

Notes: Model includes child fixed-effect, with the 'effect' capturing how difference in set allocation between subjects is associated with relative performance in English and mathematics. Estimates restricted to England only. Any observations where group allocation is the same for English and mathematics (e.g. top set for both) have been excluded (as there is no between-subject variation in within-class ability group allocation to exploit. The final model sample size is XXX. Controls included for foundation stage profile scores in English / mathematics, whether setting used in English/mathematics, English/mathematics set the child was allocated to, the amount of class time allocated to English/mathematics.

Table 9. Sub-group OLS regression estimates of the relationship between whether Year 2 teachers use within-class ability grouping and children's achievement

(a) Primary outcomes (achievement)

	Mathematics		
	Effect	SE	n
Family-income			
Low-income (bottom third)			
High-income (top third)			
Prior achievement			
Low prior achievement (bottom third)			
High prior achievement (top third)			
Setting/streaming used for subject			
No			
Yes			

(b) Secondary outcomes

	Enjoy number work			Reading scores		
	Effect	SE	n	Effect	SE	n
Family-income						
Low-income (bottom third)						
High-income (top third)						
Prior achievement						
Low prior achievement (bottom third)						
High prior achievement (top third)						
Setting/streaming used for subject						
No						
Yes						

Notes: MCS age 7 UK weight (variable = DDOVW200) applied, with clustering and stratification in the MCS survey design used to adjust the estimated standard errors. Multiple imputation used to account for missing covariate data. Models include the following controls: gender, permanent family income, ethnicity, age in months, language spoken at home and maternal and paternal education, foundation stage profile English and maths scores, MCS cognitive test scores at age 5, teacher gender, years of teaching experience, teacher qualifications, years teacher working at the school, whether children set/streamed in the subject, class size, whether the class contains mixed year groups, percent of children in class with SEN, percent of children in class with EAL, number of days taught by a supply teacher, class time spent on literacy / numeracy per week, whether class gets regular support from a teaching assistant, special needs or other teacher, percentage of lesson time devoted to group work per week

Table 10. Sub-group OLS regression estimates of the relationship between within-class ability group assignment (low, middle, high) and children's achievement

(a) Primary outcomes

	Mathematics		
	Effect	SE	N
Low-income			
Low within-class ability group			
High within-class ability group			
High-income			
Low within-class ability group			
High within-class ability group			
Low prior achievement			
Low within-class ability group			
High within-class ability group			
High prior achievement			
Low within-class ability group			
High within-class ability group			
Not set/streamed			
Low within-class ability group			
High within-class ability group			
Set / streamed			
Low within-class ability group			
High within-class ability group			

(b) Secondary outcomes

	Enjoy number work			Reading test scores		
	Effect	SE	N	Effect	SE	N
Low-income						
Low within-class ability group						
High within-class ability group						
High-income						
Low within-class ability group						
High within-class ability group						
Low prior achievement						
Low within-class ability group						
High within-class ability group						
High prior achievement						
Low within-class ability group						
High within-class ability group						
Not set/streamed						
Low within-class ability group						
High within-class ability group						
Set / streamed						
Low within-class ability group						
High within-class ability group						

Notes: MCS age 7 UK weight (variable = DDOVW200) applied, with clustering and stratification in the MCS survey design used to adjust the estimated standard errors. Multiple imputation used to account for missing covariate data. Models include the following controls: gender, permanent family income, ethnicity, age, language spoken at home and maternal and paternal education, foundation stage profile English and maths scores, MCS cognitive test scores at age 5, teacher gender, years of teaching experience, teacher qualifications, years teacher working at the school, whether children set/streamed in the subject, class size, whether the class contains mixed year groups, percent of children in class with SEN, percent of children in class with EAL, number of days taught by a supply teacher, class time spent on literacy / numeracy per week, whether class gets regular support from a teaching assistant, special needs or other teacher, percentage of lesson time devoted to group work per week

Table 11. Descriptive statistics for the TIMSS sample

(a) Year 5

	Same ability groups			Mixed ability groups		
	Never / some	**Half**	**Almost every**	**Never / some**	**Half**	**Almost every**
Child characteristics						
% male						
% FSM						
% immigrant						
% English not spoken at home						
% regularly absent from school						
Prior achievement						
Key Stage 1 maths score						
Key Stage 1 reading score						
Teacher characteristics						
% Subject specialism in maths						
Years of teaching experience						
Class characteristics						
Class size						
% students with language issues						
Minutes per week teaching class						
Average KS1 maths scores of class						
Standard deviation of class KS1 maths scores						
School characteristics						
Pupils set for mathematics						
% of pupils from disadvantaged backgrounds						
Resource shortages affecting maths instruction scale						

Notes: Final TIMSS student weight applied.

(b) Year 9

	Same ability groups			Mixed ability groups		
	Never / some	**Half**	**Almost every**	**Never / some**	**Half**	**Almost every**
Child characteristics						
% male						
% FSM						
% immigrant						
% English not spoken at home						
% regularly absent from school						
Prior achievement						
Key Stage 2 maths score						
Key Stage 2 reading score						
Teacher characteristics						
% Subject specialism in maths						
Years of teaching experience						
Class characteristics						
Class size						
% students with language issues						
Minutes per week teaching class						
Average KS2 maths scores of class						
Standard deviation of class KS2 maths scores						
School characteristics						
Pupils set for mathematics						
% of pupils from disadvantaged backgrounds						
Resource shortages affecting maths instruction scale						

Notes: Final TIMSS student weight applied.

Table 12. Cross-tabulation of the within-class same ability and mixed ability grouping variables. TIMSS dataset.

(a) Year 5

Same ability ↓ / mixed ability →	No lessons	Some lessons	About half of lessons	Every / almost every lesson
No lessons				
Some lessons				
About half of lessons				
Every / almost every lesson				

(b) Year 9

Same ability ↓ / mixed ability →	No lessons	Some lessons	About half of lessons	Every / almost every lesson
No lessons				
Some lessons				
About half of lessons				
Every / almost every lesson				

Notes: Final TIMSS student weight applied. Figures refer to column percentages.

Table 13. OLS regression model estimates of the relationship between within-class mixed-ability and same-ability grouping and children's mathematics outcomes

(a) Primary

	Maths achievement			Self-confidence		
	Effect	**SE**	**n**	**Effect**	**SE**	**n**
Teaching same ability groups						
No/some lessons (reference)						
Half of lessons						
Almost every lesson						
Teaching mixed ability groups						
No/some lessons (reference)						
Half of lessons						
Almost every lesson						
Observations						

(b) Secondary

	Maths achievement			Self-confidence		
	Effect	**SE**	**n**	**Effect**	**SE**	**n**
Teaching same ability groups						
No/some lessons (reference)						
Half of lessons						
Almost every lesson						
Teaching same ability groups						
No/some lessons (reference)						
Half of lessons						
Almost every lesson						
Observations						

Notes: Estimates based upon OLS regression models, applying the final student and replication weights, and averaging across estimates using the five plausible values. Child controls include gender, FSM status, immigrant status, age in months, home economic resources index, absence from school, language spoken at home, prior English and mathematics Key Stage scores. Teacher controls include whether they have a subject specialism in mathematics and number of years teaching experience. Class characteristics include minutes per week spent teaching to the class, average prior maths achievement levels of the class, standard deviation of the prior maths achievement of the class. School controls include whether it is school policy to set/stream pupils in mathematics and the percent of pupils who come from advantaged/disadvantaged backgrounds.

Table 14. Sub-group regression model estimates of the relationship between within-class mixed-ability and same-ability grouping and children's mathematics achievement. Year 5 results.

(a) Mathematics achievement

	FSM		Non-FSM		Low prior achievement		High prior achievement		Setting/streaming used		Setting/streaming not used	
	Effect	SE	Effect	SE	Effect	SE	Effect	SE	Effect	SE	Effect	SE
Teaching same ability groups												
No/some lessons (reference)												
Half of lessons												
Almost every lesson												
Teaching mixed ability groups												
No/some lessons (reference)												
Half of lessons												
Almost every lesson												
Observations												

(b) Self-confidence

	FSM		Non-FSM		Low prior achievement		High prior achievement		Setting/streaming used		Setting/streaming not used	
	Effect	SE	Effect	SE	Effect	SE	Effect	SE	Effect	SE	Effect	SE
Teaching same ability groups												
No/some lessons (reference)												
Half of lessons												
Almost every lesson												
Teaching mixed ability groups												
No/some lessons (reference)												
Half of lessons												
Almost every lesson												
Observations												

Notes: See notes to Table 13 for further details. All estimates presented in terms of an effect size.

Table 15. Sub-group regression model estimates of the relationship between within-class mixed-ability and same-ability grouping and children's mathematics achievement. Year 9 results.

(a) Mathematics achievement

	FSM		Non-FSM		Low prior achievement		High prior achievement		Setting/streaming used		Setting/streaming not used	
	Effect	**SE**	**Effect**	**SE**	**Effect**	**SE**	**Effect**	**SE**	**Effect**	**SE**	**Effect**	**SE**
Teaching same ability groups												
No/some lessons (reference)												
Half of lessons												
Almost every lesson												
Teaching mixed ability groups												
No/some lessons (reference)												
Half of lessons												
Almost every lesson												
Observations												

(b) Self-confidence

	FSM		Non-FSM		Low prior achievement		High prior achievement		Setting/streaming used		Setting/streaming not used	
	Effect	**SE**	**Effect**	**SE**	**Effect**	**SE**	**Effect**	**SE**	**Effect**	**SE**	**Effect**	**SE**
Teaching same ability groups												
No/some lessons (reference)												
Half of lessons												
Almost every lesson												
Teaching mixed ability groups												
No/some lessons (reference)												
Half of lessons												
Almost every lesson												
Observations												

Notes: See notes to Table 13 for further details. All estimates presented in terms of an effect size.

Appendix A. Question used to gather information on within-class ability grouping in the MCS

Class groupings

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

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

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

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

- Q55. In this child's class, is there within-class ability grouping? (Yes / No)
- Q56. How many within-class ability groups are there? (Open text)
- Q57. Which group is this child in? (Highest, middle, lowest)
- Q58. In this child's class, are there within-class subject groups for literacy? (Yes/No)
- Q59. How many within-class subject groups are there for literacy? (Open text)
- Q60. Which group is this child in for literacy? (Highest, middle, lowest)
- Q61. In this child's class, are there within-class subject groups for maths? (Yes/No)
- Q62. How many within-class subject groups are there for maths? (Open text)
- Q63. Which group is this child in for maths? (Highest, middle, lowest)

Appendix B. Pre-specified descriptive statistics

TIMSS includes the following question, which encompasses three statements of interest:

"In teaching mathematics to this class, how often do you ask students to do the following?"

- ATBM03E = Work problems together in the whole class with direct guidance from me
- ATBM03H = Work in mixed ability groups
- ATBM03I = Work in same ability groups

There are four possible response options: Never, some lessons, about half of lessons, almost every lesson.

Table 1 below illustrates the (polychoric) correlation between responses provided by those who teach children mathematics (note – this initial descriptive analysis is conducted at the teacher level with no weight applied). Annex A provides cross-tabulations, presenting the number of teachers in each cells.

There is a moderate negative correlation between same and mixed ability grouping; those who group pupils by ability more frequently are less likely to use same ability grouping.

Table 1. The polychoric correlation between use of different teaching approaches

(a) Year 5

	Same ability grouping	Mixed ability grouping	Whole class teaching
Same ability grouping	-	-	-
Mixed ability grouping	-0.40	-	-
Whole class teaching	0.26	0.24	-

(b) Year 9

	Same ability grouping	Mixed ability grouping	Whole class teaching
Same ability grouping	-	-	-
Mixed ability grouping	-0.31	-	-
Whole class teaching	-0.17	0.15	-

The is a weak positive correlation between both forms of grouping pupil and whole class teaching. Teachers who said they used group work frequently were also (slightly) more likely to say they used whole class teaching. Though this association is clearly weak (correlation coefficient around 0.25).

Overall, the descriptive statistics in Table 1 and Annex A suggest:

- The bottom two categories (never and some) may need to be combined to avoid small / zero cell sizes. (See Appendix A).
- Including same and mixed ability grouping in the same model (as planned) should not be a problem.
- Including same, mixed and whole group teaching in the same model may be possible, though there may be some concerns about some small cell sizes.

Tables 2 (Year 5) and 3 (Year 9) investigate how a set of teacher and school characteristics are associated with the use of within-class ability grouping. The purpose was to try and get a better understanding about potential confounders, given the limited knowledge we have about the 'selection mechanism' (why teachers choose to group pupils within classes by ability).

However, there is little strong evidence for or against the inclusion of most controls. In general, differences seem to be relatively small in most instances, and are unlikely to be major sources of confounding. One potential exception is the school affluence scale; however, as the primary outcome model will control for pupil-level measures of disadvantaged (e.g. FSM and home economic resources) it remains unclear whether even this school-level factor is potentially important.

Given the relatively small teacher and school level sample size in TIMSS, and the lack of clear evidence of potential confounding by observable teacher and school characteristics, it has been decided that a relatively limited set of teacher/ school controls in the main analysis. A wider array of school/teacher controls will then be included in the models as a robustness test.

Table 2. The association between teacher/school characteristics and use of different mathematics teaching practices. Year 5 TIMSS.

	Mixed ability group			Same ability group			Whole class		
	Every	Half	Some / never	Every	Half	Some / never	Every	Half	Some / never
Maths background									
No	18	22	59	43	35	22	36	41	23
Yes	26	22	52	30	30	41	33	44	22
Years of experience									
5 years or less	18	19	63	47	34	19	40	42	18
6 years +	20	24	55	38	35	28	32	42	27
Class size									
25 and below	11	30	60	34	34	32	32	40	28
26+	23	19	58	43	35	22	35	42	22
Pupil(s) with language issues									
No	17	20	62	39	38	23	34	40	27
Yes	21	26	52	46	30	25	36	44	20
Maths per week									
5 hours and less	15	22	63	42	33	24	31	45	24
more than 5 hours	29	22	49	39	35	27	45	33	22
Confidence in teaching maths									
Bottom half	13	20	68	41	33	25	28	46	26
Top half	27	26	47	41	35	24	43	36	20
School affluence									
More affluent	19	14	67	26	38	36	17	52	31

Neither	16	28	56	50	32	18	32	46	22
More disadvantage	25	13	63	50	21	29	52	31	17
Set / stream in maths									
No	20	26	54	39	37	24	29	47	24
Yes	18	16	66	45	26	29	37	38	25
School shortages in maths									
No	15	28	57	36	39	25	24	52	24
Yes	25	13	63	47	25	28	43	32	25

Notes: Analysis at teacher level. No weights applied. Figures refer to row percentages.

Table 3. The association between teacher/school characteristics and use of different mathematics teaching practices. Year 9 TIMSS.

	Mixed ability group			Same ability group			Whole class		
	Every	Half	Some / never	Every	Half	Some / never	Every	Half	Some / never
Maths background									
No	14	28	59	24	24	52	41	28	31
Yes	8	22	71	31	22	47	39	40	22
Years of experience									
5 years or less	10	25	65	27	25	48	38	30	32
6 years +	8	21	71	31	21	48	40	42	18
Class size									
25 and below	6	19	75	27	20	53	37	41	23
26+	10	26	64	33	25	43	41	37	23
Pupil(s) with language issues									
No	8	22	70	30	22	48	37	41	22
Yes	8	25	67	29	25	46	44	31	25
Maths per week									
3 hours and less	9	22	69	31	20	49	46	31	23
more than 3 hours	8	24	67	25	27	48	31	47	22
Confidence in teaching maths									
Bottom half	8	21	71	26	21	53	42	33	25
Top half	9	24	66	34	24	42	36	43	21
School affluence									
More affluent	13	26	62	43	21	36	40	43	17
Neither	7	23	70	6	23	51	44	34	21
More disadvantage	12	17	71	27	22	51	27	46	27
School shortages in maths									
No	9	27	65	35	23	42	46	33	22
Yes	11	18	72	26	22	53	31	47	22

Notes: Analysis at teacher level. No weights applied. Figures refer to row percentages.

Annex A. Cross-tabulations of whole class teaching, same ability grouping and mixed ability grouping in TIMSS (numbers of teachers in England)

Mixed ability versus same ability

(a) Year 5

Mixed ability ↓ / Same ability →	Every / almost every	About half	Some lessons	Never	Total N
Every / almost every	7	7	13	4	**31**
About half	3	33	0	0	**36**
Some lessons	51	14	21	0	**86**
Never	5	1	1	1	**8**
Total N	**66**	**55**	**35**	**5**	**161**

(b) Year 9

Mixed ability ↓ / Same ability →	Every / almost every	About half	Some lessons	Never	Total N
Every / almost every	3	1	6	6	**16**
About half	1	32	8	0	**41**
Some lessons	17	3	51	4	**75**
Never	33	5	3	9	**50**
Total N	**54**	**41**	**68**	**19**	**182**

Mixed ability versus whole class

(a) Year 5

Mixed ability ↓ / Whole class →	Every / almost every	About half	Some lessons	Never	Total N
Every / almost every	17	12	2	0	**31**
About half	9	17	8	2	**36**
Some lessons	29	35	21	1	**86**
Never	1	3	4	0	**8**
Total N	**56**	**67**	**35**	**3**	**161**

(b) Year 9

Mixed ability ↓ / Whole class →	Every / almost every	About half	Some lessons	Never	Total N

Every / almost every	4	6	4	2	**16**
About half	14	18	9	0	**41**
Some lessons	29	29	16	1	**75**
Never	24	16	9	1	**50**
Total N	**71**	**69**	**38**	**4**	**182**

Same ability versus whole class

(a) Year 5

Same ability ↓ / Whole class →	Every / almost every	About half	Some lessons	Never	Total N
Every / almost every	34	21	10	1	66
About half	13	28	12	2	55
Some lessons	6	17	12	0	35
Never	3	1	1	0	5
Total N	**56**	**67**	**35**	**3**	**161**

(b) Year 9

Same ability ↓ / Whole class →	Every / almost every	About half	Some lessons	Never	Total N
Every / almost every	23	19	12	0	**54**
About half	14	21	6	0	**41**
Some lessons	27	26	14	1	**68**
Never	7	3	6	3	**19**
Total N	**71**	**69**	**38**	**4**	**182**