Education Endowment Foundation

The Student Grouping Study

Evaluation report

April 2026

Jeremy Hodgen, Becky Taylor, Antonina Tereshchenko, Jake Anders, Laurie Jacques, Maria Cockerill, Rosa Kwok, David Sirl, Hester Burn, and Nicola Bretscher

UCL Institute of Education

The Education Endowment Foundation (EEF) is an independent charity dedicated to breaking the link between family income and education achievement. We support schools, colleges, and early years settings to improve teaching and learning for 2–19-year-olds through better use of evidence.

We do this by:

- **Summarising evidence**. Reviewing the best available evidence on teaching and learning and presenting in an accessible way.

- **Finding new evidence.** Funding independent evaluations of programmes and approaches that aim to raise the attainment of children and young people from socio-economically disadvantaged backgrounds. Putting evidence to use.

- **Putting evidence to use**. Supporting education practitioners, as well as policymakers and other organisations, to use evidence in ways that improve teaching and learning.

We were set-up in 2011 by the Sutton Trust partnership with Impetus with a founding £125m grant from the Department for Education. In 2022, we were reendowed with an additional £137m from government, allowing us to continue our work until at least 2032.

For more information about the EEF or this report please contact:

The Education Endowment Foundation
5th Floor, Millbank Tower,
21–24 Millbank,
London,
SW1P 4QP

0207 802 1653

info@eefoundation.org.uk

www.educationendowmentfoundation.org.uk

Table of contents

About the evaluator 3

Acknowledgements 3

Executive summary 4

Introduction 6

Methods 15

Impact evaluation results 44

IPE results 54

Cost 68

Conclusion 69

References 73

Appendix A: The EEF cost rating 80

Appendix B: Security classification of evaluation findings 81

Further appendices: 83

About the evaluator

The Student Grouping Study was conducted by a team from University College London (UCL) Institute of Education, Brunel University, and Queens University Belfast, led by Professor Jeremy Hodgen and Professor Becky Taylor.

For further information, please see contact details below:

Contact details:

Professor Becky Taylor
UCL Institute of Education,
20 Bedford Way,
London,
WC1H 0AL

Email: **becky.taylor@ucl.ac.uk**

Acknowledgements

We are extremely grateful to the students and teachers from all the schools that participated in this research.

We would like to thank our colleagues Pete Wright and Piers Saunders, who helped with validation of the lesson observation instrument and assisted with case study data collection.

During this research we received invaluable research and administrative assistance from a number of people, including: Sadie Briggs; Maria Gil; Bohan Liu; Qian Liu; Laura Male; Jo Morgan; Katie Pepper; and a team of test administrators.

We have been expertly supported throughout this research by our study advisory group, consisting of: Rob Coe; Stefan Speckesser; Fatima Husain; Bronwen Maxwell; and Richard Dorsett.

We would also like to thank all the evaluation and programme managers at the Education Endowment Foundation who have worked with us during this research since the project started in 2019, especially Lottie Norton and Jamila Boughelaf.

This work was undertaken in the Office for National Statistics Secure Research Service using National Pupil Database data owned by the Department for Education in England (reference DR240920.01). The DfE does not accept responsibility for any inferences or conclusions derived from the DfE Matched Dataset by third parties. Use of the ONS SRS does not imply the endorsement of the ONS.

Executive summary

The project

The Student Grouping Study aimed to compare directly the impact of being taught mathematics in mixed attainment groups or in attainment sets on mathematics attainment and self-confidence, for Year 7 and Year 8 students (age 11–13). Of particular interest were outcomes for disadvantaged students, and outcomes for students with low-prior attainment.

This was a naturalistic quasi-experimental study that investigated the effects of setting by attainment and mixed attainment grouping in mathematics, comparing schools with established grouping practices. The evaluation used a matched design, meaning that two groups of schools were deliberately made as similar as possible on a broad range of observable characteristics before comparing the outcomes. Schools were recruited on the basis of their usual practices in grouping for mathematics teaching. 28 mixed attainment schools were recruited, followed by 69 matched schools that used setting by attainment.
The study took place between September 2022 and July 2024. There were significant delays in accessing data from the National Pupil Database, which meant this report is published later than anticipated. The implementation and process evaluation (IPE) consisted of surveys in all participating schools, of heads of mathematics, mathematics teachers, school leaders, and students. Additionally, 12 schools were randomly selected for in-depth qualitative case studies.

Table 1: Key conclusions

Key conclusions	
1.	Students in schools using mixed attainment grouping made one month's less progress in mathematics, on average, compared to students in schools using setting by attainment. This is our best estimate of impact, which has a low/moderate security rating.
2.	Among students eligible for free school meals (FSM), those in school using mixed attainment grouping made similar progress in mathematics, on average, compared to those in schools using setting by attainment. These results may have lower security than the overall findings because of the smaller number of students.
3.	Among low-prior attaining students, those in schools using mixed attainment grouping made similar progress in mathematics, on average, compared to students in schools using setting by attainment. These results may have lower security than the overall findings because of the smaller number of students.
4.	Among high-prior attaining students, those in schools using mixed attainment grouping made, on average, two fewer months' progress in mathematics compared to students in schools using setting by attainment. These results may have lower security than the overall findings because of the smaller number of students.
5.	The IPE lesson observations indicated that schools' teaching practices are broadly similar across schools that used mixed attainment grouping and schools that set by attainment and that overall practices in most case study schools appear to be driven by national policy around mastery. In general, despite well-intentioned policies in mixed attainment schools around equity and challenge for high attainers, only setting schools appear to be challenging high-prior attaining students.

EEF security rating

The headline finding has a low to moderate security rating. The primary outcome was well powered. The samples of schools teaching mathematics in mixed attainment classes and those teaching mathematics in sets were well-matched on attainment and most other key variables at student and school levels, although there was some imbalance in the Office for Standards in Education, Children's Services and Skills (Ofsted) rating between schools. Around 29% of students were excluded from the final analysis due to students dropping out of the study over time. In addition, the security of the findings was further lowered, because only a small proportion of schools teach mathematics in mixed attainment classes in both Year 7 and Year 8, which may mean that there were unmeasured differences that may have influenced the overall finding.

Additional findings

The impact analysis findings show that, for students overall, those in schools using mixed attainment grouping made one month's less progress in mathematics, compared with those in schools using setting by attainment. This is our best estimate of impact, which has a low/moderate security rating. As with any study, there is always some uncertainty around the result: the possible difference for students in schools using mixed attainment grouping ranges from two months' less progress to no difference in progress. This finding is broadly consistent with the EEF Toolkit (EEF, 2021) and the wider literature (Steenbergen-Hu *et al.*, 2016). For students with lower prior attainment, those in schools using mixed attainment classes made similar progress to those in schools using setting by attainment.

There was some evidence of a reduction in the mathematics attainment gap between low- and high-prior attaining students in schools using mixed attainment grouping. This appears to be largely driven by lower progress among the high-prior attaining students, rather than greater progress by the low-prior attaining group. These results are broadly consistent with existing research, although the EEF Toolkit (EEF, 2021) meta-analysis suggests that the magnitude of these effects is more balanced: while a small negative effect of mixed attainment would be expected for high-prior attainers, a small positive effect would be expected for low-prior attainers. Much of the existing international evidence on attainment grouping is dated, of low quality and focuses on tracking, a form of attainment grouping used in the United States, rather than subject-specific setting by attainment that is common in England.

For the subgroup of students in receipt of FSM, there was no difference in attainment between those in schools using setting by attainment and those in schools using mixed attainment groups, which questions prior evidence that students from disadvantaged backgrounds are disadvantaged by the practice of setting. In line with the attainment results, the impact analysis also found negative effects on mathematics self-confidence for students in mixed attainment schools compared to those in schools using setting. These negative effects were small for the student group as a whole, and for students in receipt of FSM, but moderate for the low-prior attaining students. Through lesson observations in IPE case study schools, we found that lessons from mixed attainment schools tended to be more similar to the bottom set than to top set lessons, especially in terms of pace. Although there were opportunities for challenging extension activities for high-prior attaining students, these opportunities frequently involved mathematics that was not directly related to the lesson content and were rarely discussed in class. In contrast, top sets provided more of a challenge through a faster pace and a greater emphasis on GCSE.

The EEF Toolkit (EEF, 2021) judges the current evidence on the effects of attainment grouping on student progress to be very limited. This study strengthens this evidence base by providing more robust evidence on the impact of mixed attainment in the context of the English educational system, including those from disadvantaged backgrounds.

Cost

A small amount of additional continuous training may be required in implementing mixed attainment grouping in mathematics when compared with setting, as only a minority of mathematics teachers have prior experience of mixed attainment grouping. This additional requirement was estimated as no more than one day of internal training per year.

Impact

Table 2: Summary of impact on primary outcome for students taught in mixed attainment classes compared to those taught in setted classes

Outcome / Group	Effect size (95% confidence interval)	Estimated months' progress	EEF security rating	No of pupils	P Value	EEF cost rating
Mathematics achievement (all students)	-0.05 (-0.14, 0.03)	-1	🔒🔒 (2 of 5)	14,877	0.232	—
Mathematics achievement (low-prior attainers)	0.04 (-0.05, 0.13)	0	🔒🔒 (2 of 5)	4,378	0.330	—
Mathematics achievement (FSM)	0.01 (-0.08, 0.10)	0	🔒🔒 (2 of 5)	2,883	0.860	—
Mathematics achievement (high-prior attainers)	-0.14 (-0.25, -0.03)	-2		5,360	0.017	—

Introduction

Background

The question of whether setting or mixed attainment grouping is a more effective strategy for teaching in secondary schools remains a contested issue in England. Mixed attainment grouping is the practice of grouping students for teaching in such a way that there is a wide range of prior attainment in each teaching group. Some schools choose to accompany mixed attainment groups with a nurture group of students who are deemed not ready for secondary school.

By contrast, setting is the practice of grouping students for teaching by their prior attainment in the subject being taught, with the intention that within each teaching group, students have a similar, narrow, range of prior attainment. Other forms of grouping are also used in schools in England, including streaming, where students are grouped by 'ability' for all subjects, and within-class grouping, where students are placed in attainment groups within the classroom. In the USA the term 'tracking' is often used, which refers to practices similar to streaming.

The Student Grouping Study addresses a gap in the evidence on setting and mixed attainment grouping and student outcomes (EEF, 2021). Research over several decades shows a small negative effect of setting on the attainment of low attainers and a small positive effect on the attainment of high attainers (e.g. Steenbergen-Hu *et al.,* 2016). Similarly, Terrin and Triventi's (2023) meta-analysis examining effects at the system level finds tracking and setting to be associated with increased inequity of outcomes. While research on the efficacy of mixed attainment is limited (Francis *et al.,* 2017), there is evidence that mixed attainment grouping has a beneficial effect on the learners' self-confidence (e.g. Boaler *et al.,* 2000) and attainment (Rui, 2009) and research from the United States has shown dramatic effects of mixed attainment teaching on low and average attainers' completion of pre-college mathematics courses (Burris *et al.,* 2006; White *et al.,* 1996). Others contend that mixed attainment grouping makes greater demands on teachers and may thus only be effective in schools with highly effective teachers (Delisle, 2015). Yet, many practitioners are steadfastly and vociferously committed to one approach or the other (e.g. Old and Reddy, 2015), while others are actively seeking advice about attainment grouping, wanting to use the best, evidence-informed practice. A Department for Education (DfE)-commissioned report examining how schools support the attainment of disadvantaged students found that more than a third of primary and secondary schools surveyed had 'introduced or improved' setting as a way of raising attainment for disadvantaged students (DfE, 2015).

Much of the existing research on between-class attainment grouping has been conducted in the USA and is more than 25 years old. Additionally, Steenbergen-Hu *et al.'s* (2016) second-order meta-analysis finds that none of the existing meta-analyses is of high quality. High-quality research directly comparing setting and mixed attainment grouping has not yet been conducted in England and there was an urgent need for such research. This study examined the effects on student attainment, attitudes, and other non-cognitive outcomes, some of which are likely to be predictive of important long-term outcomes, such as sustained improvements in attainment or participation in higher education.

Our literature review showed that students in lower-attaining sets tend to be allocated teachers with lower qualifications (and see Francis *et al.,* 2019) and to be exposed to a more restricted curriculum as evidenced by lower teacher expectations (and see Mazenod *et al.,* 2018), restricted teacher pedagogy and a lower quality curriculum (Oakes, 1985; Hodgen *et al.,* 2021). Thus, they have less opportunity to learn (OTL)[1] and are offered teaching of low quality. However, Dunne *et al.'s* (2007, 2011) studies of the teaching and learning of students in low-attaining sets in English secondary schools suggests that some schools take active steps to avoid and additionally to mitigate against these effects by, for example, reducing class sizes in lower sets.

Our previous study, Best Practice in Grouping Students (BPGS), investigated the effects of implementing two interventions, which were designed to enable schools to implement good practice in grouping by attainment and in mixed attainment

[1] OTL generally refers to educational inputs and instructional processes within a school setting necessary for helping students to achieve the intended outcomes. Our approach to OTL focuses on instructional time and content coverage (Kurz, 2011; Suter, 2017).

grouping. BPGS adopted a fully powered experimental randomised controlled trial (RCT) design for the ‘Best Practice in Setting’ intervention and a feasibility trial for the ‘Best Practice in Mixed Attainment’ using an under-powered pilot RCT design. The interventions consisted of instructions for grouping students and allocating teachers to groups, and training for teachers on classroom expectations and pedagogy.

The ‘Best Practice in Setting’ trial found no effect for the intervention compared to a ‘business as usual’ control group (Roy *et al.,* 2018a). Our mixed-methods process evaluation (largely conducted by the project team) identified this lack of effect as being due to low fidelity in applying the practices required (see Taylor *et al.,* 2017; Taylor *et al.,* 2018). For example, we found that practical issues such as timetabling impeded optimal and accurate set allocation practice (Taylor *et al.,* 2018). A number of schools in the control group also practised aspects of the intervention, for example, by having three or four set levels, or by using prior attainment to allocate students to sets (Taylor *et al.,* 2018). These issues with compliance overall reduced the difference between the intervention and control groups and thus the detectable impact of the intervention.

The feasibility trial, ‘Best Practice in Mixed Attainment’, sought to learn more about good mixed attainment practice (an under-researched and under-reported area), and to test the feasibility of application of an intervention in this regard. While we learned much from this (and demonstrated feasibility of application), the small-scale nature of the sample, compounded by the mixed circumstances of the schools recruited to the control and intervention group, meant that outcome measures could not be extrapolated. Furthermore, issues with compliance resulted in reduced differences between the intervention and control groups. The majority of schools in the control group were also practising mixed attainment grouping, and there was non-compliance with mixed attainment grouping in the intervention group in some schools, who returned to setting in the second year of the intervention.

Both trials were also affected by schools’ differing understandings of the definitions of attainment grouping practices, resulting in attrition and non-compliance. For example, schools in the ‘Best Practice in Setting’ trial confused setting and streaming, while schools in the ‘Best Practice in Mixed Attainment’ trial confused mixed attainment grouping and setting where there were few set levels (Taylor *et al.,* 2017). Schools in both trials also used a wide variety of pedagogic practices (described by teachers in the setting and mixing schools, and observed by the research team in mixed attainment schools), raising questions about the characterisation of distinctive ‘setting’ or ‘mixed attainment’ pedagogies (Taylor *et al.,* 2020).

These two prior trials resolved several outstanding questions in the literature, provided new findings about why setting negatively impacts low attainers, and identified why this situation is unlikely to improve in spite of good intentions/evidence-informed interventions designed to do so. However, they were not designed to compare directly mixed attainment practice with ‘ability grouping’. This left an additional, fundamental question unanswered. This question is: Which has a greater impact on progress and attainment—setting or mixed attainment practice?

The Student Grouping Study was designed to:

- Provide direct and robust evidence relevant to schools in England comparing the effects of setting and mixed attainment teaching.

- Provide evidence of the effects of setting and mixed attainment teaching relative to other approaches for addressing low attainment or disadvantage. This was deemed to be particularly important if the study were to find little or no difference in the impact on student attainment between the two approaches. To investigate the experiences of low-attaining students we compared their experiences to those of high-attaining students.

- Provide detailed evidence of how grouping practices are implemented in different contexts. The study aimed to characterise effective and equitable practices (i.e. practices that enable all students to make progress and have positive mathematical experiences regardless of prior attainment or background) so that schools can understand how to implement setting and mixed attainment in order to make the greatest impact on student outcomes in mathematics. This included developing a measure of OTL. In addition, findings regarding implementation of grouping were deemed likely to be useful for the design of grouping interventions that address schools’ needs and are thus likely to be acceptable to schools.

- Provide evidence relevant to a wide range of secondary school subjects, because there is particularly vigorous debate about the impact of attainment grouping in mathematics, where setting is the dominant practice.

Intervention

The study used a matched design in a natural context, to explore the difference in student outcomes between two approaches to organising students for teaching: grouping by subject ability (or setting); and mixed attainment grouping. Schools were recruited based on their existing practices and there was no experimental intervention. Our description of the practices being compared follows the Template for Intervention Description and Replication (TIDieR) framework (Hoffmann *et al.,* 2014).

Name: The Student Grouping Study. Investigating the effects of setting and mixed attainment grouping with a naturalistic matched approach comparing schools with established grouping practices.

Why: When students are taught in attainment sets, the literature suggests that students in lower-attaining groups make less progress compared with their higher-attaining peers (Steenbergen-Hu *et al.,* 2016;EEF*,* 2021).

From the literature and from our own research (Taylor *et al.,* 2019, 2020), we have established that 'Setting' comprises a range of practices which all make use of measures of 'ability' or attainment to group students for teaching in a specific subject, but with local variation in the exact sources of data used to allocate students to sets, the number of set 'levels', the distribution of students across 'levels' and the amount of movement between sets after initial allocation. Conversely, the grouping strategy 'Mixed attainment grouping' comprises a range of practices in which the general principle is to achieve a broad range of prior attainment or 'ability' in each teaching group.

Our experience with 'Best Practice in Mixed Attainment' demonstrated that it is not feasible to compare the impact of setting and mixed attainment grouping using randomisation (Roy *et al.,* 2018b; and see Taylor *et al.,* 2017). We therefore, employed a naturalistic matched design for The Student Grouping Study, where schools retained their established grouping practices.

We conjectured that OTL, teacher quality, class size, and student attitudes (such as liking for school and engagement) may act as mediating or explanatory factors in the relationship between attainment grouping and the impact on key outcomes recognised in the literature: attainment and self-confidence (e.g. Baumert *et al.,* 2010; Dunne *et al.,* 2011; Francis *et al.,* 2017, 2020. We hypothesised that these outcomes are also likely to impact on students' orientation to future participation (cf. Archer *et al.,* 2012).

We conjectured that moderating factors are likely to include those relating to the student (socio-economic status [SES], ethnicity, and sex), the school (characteristics of the whole intake including prior attainment, leadership ethos, and resources), and the teacher (beliefs and attitudes).

Who/Where: We focused on mathematics grouping and teaching for Year 7 and Year 8 students in English state secondary schools. Year 7 and Year 8 were chosen as the first two years of secondary school, and a new phase in students' education. Mathematics was chosen as the subject focus because mathematics teachers have tended to be among the most loyal adherents to setting as a grouping practice (Reid *et al.,* 1981; Taylor *et al.,* 2018). We therefore, consider that mathematics is a useful 'test case' for the feasibility, or not, of mixed attainment grouping that would be potentially convincing to school leaders and teachers and leaders of other subjects, such as science or modern foreign languages.

What/How: Schools were recruited according to their usual grouping practices in mathematics and continued to teach students using their usual resources and strategies. Schools were recruited so that, as far as possible, the two groups of schools were matched on prior attainment of students at Key Stage 2.

We conjectured that schools' decisions to set or mix would be influenced by a number of factors, including prior attainment of the cohort, student characteristics, capacity to implement change, and local/regional influences. See the logic model for school decisions on grouping students (Figure 1) and analysis of which factors influence schools' decisions around grouping practices (see Appendix C.3).

When and How much: Variation in the experiences of students with differing levels of prior attainment were explored thoroughly through the implementation and process evaluation (IPE), with particular attention to students with low-prior attainment and those from disadvantaged backgrounds.

Tailoring: Variations in grouping practices and pedagogy within and between schools were explored thoroughly in the IPE. In particular, we explored variation in teacher quality and in OTL.

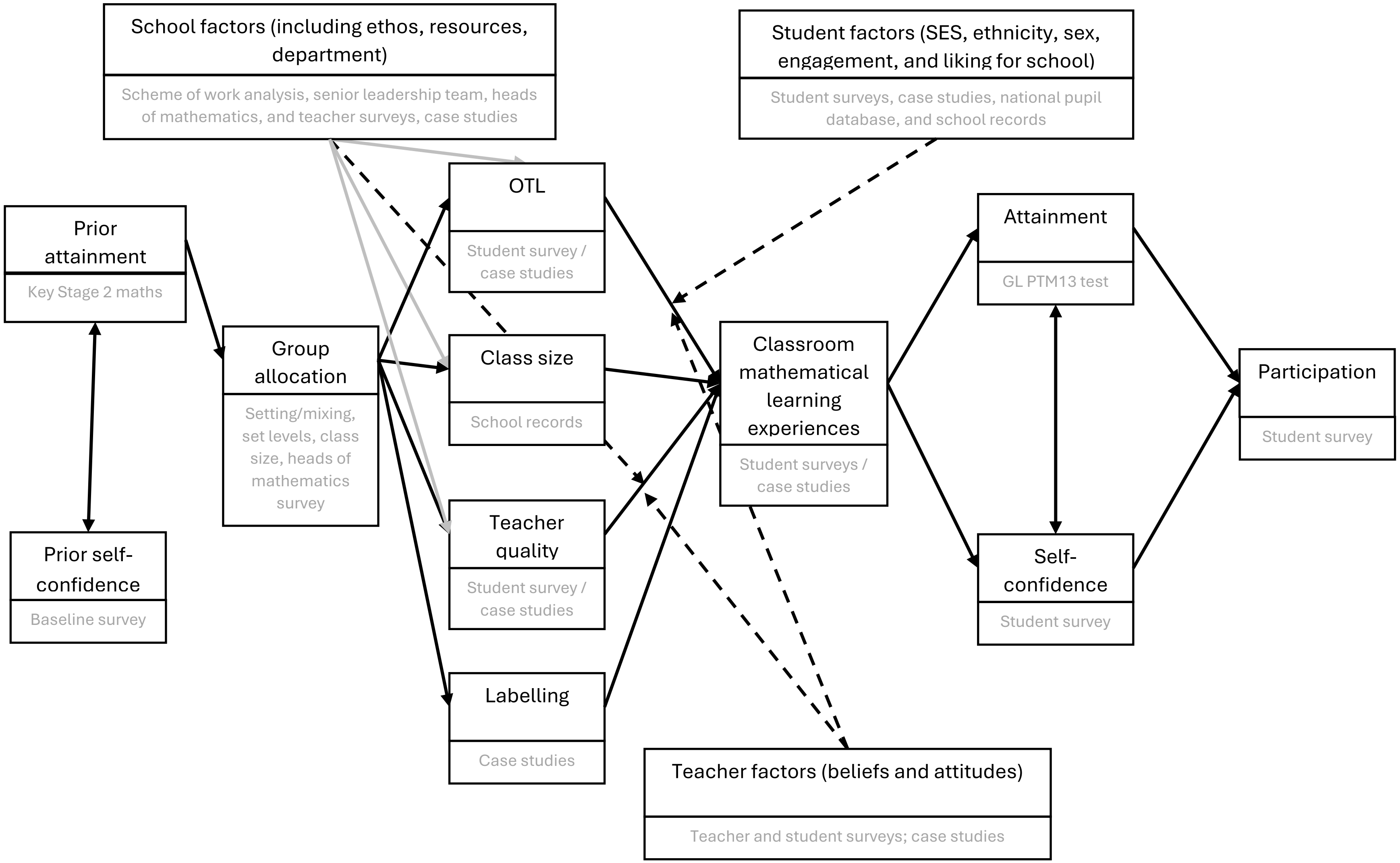
Logic model

The logic model for the study can be found in Figure 1 and includes inputs, proximal outputs, primary and secondary outcomes, mediators and moderators, as well as how each of these was measured during the trial.

We expected to find that mixed attainment grouping and setting would result in different outputs, in terms of OTL, teacher quality, class size, and labelling. We anticipated that these outputs would affect the short-term outcome of students' classroom mathematical learning experiences, and the longer-term outcomes of attainment and self-confidence in mathematics. We also expected the long-term outcome of participation in mathematics to be affected. We anticipated that students in schools that use mixed attainment grouping would have more similar mathematical learning experiences to one another, regardless of prior attainment. In other words, students with high- and low-prior attainment should have similar experiences of learning mathematics, with regard to the curriculum taught, quality of teaching, class size, and experience of labelling. We expected that students in schools that use setting were likely to have different experiences of learning mathematics, based on their prior attainment. As outlined above, research suggests that students in high-attaining sets would be taught a richer curriculum, have access to better teaching, and would benefit from positive labels. They would also be likely to be taught in larger classes. Students in low-attaining sets would be likely to be taught a more restricted curriculum, have access to restricted pedagogies and lower quality teaching and experience negative labelling. They would be likely to be taught in smaller classes but with access to fewer mathematical role models.

We expected that outputs and short-term outcomes would be moderated by school, student, and teacher factors. School factors include the school's ethos or vision, mathematics department goals, and resources. For example, we anticipated that schools with a focus on equity or social justice might try to give students a more similar educational experience, regardless of prior attainment or background. Student factors include disadvantage, ethnicity, sex, engagement, and liking for school. Prior research suggests that students from disadvantaged backgrounds, from certain minority ethnic groups and girls would be more likely to be placed in lower-attaining sets (Archer *et al.,* 2018; Muijs and Dunne, 2010). Teacher factors include teachers' beliefs and attitudes about students learning mathematics and are likely to shape the way that they teach students in different groups and with different prior attainment.

Figure 1: Logic model for The Student Grouping Study. Solid lines indicate causal relationships; dotted lines indicate moderators; grey lines are included to indicate that contextual school factors that directly impact on teacher quality, OTL, and class size



Evaluation objectives

The objective was to evaluate the effects of setting and mixed attainment grouping on student attainment and attitudes. The study addressed the following research questions:[2]

1. What difference is there, if any, in the attainment of all students over Years 7 and 8 attending schools that use mixed attainment grouping for mathematics, and all students over Years 7 and 8 attending a similar group of schools that use setting? (a) for all students; (b) for students receiving free school meals (FSM).
2. What difference is there, if any, in the attainment of low-attaining students over Years 7 and 8 attending schools that use mixed attainment grouping for mathematics, and low-attaining students over Years 7 and 8 attending a similar group of schools that use setting? (a) for all low-attaining students; (b) for low-attaining students receiving FSM.
3. What difference is there, if any, in the mathematics self-confidence of all students over Years 7 and 8 attending schools that use mixed attainment grouping for mathematics, and all students over Years 7 and 8 attending a similar group of schools that use setting? (a) for all students; (b) for students receiving FSM.
4. What difference is there, if any, in the mathematics self-confidence of low-attaining students over Years 7 and 8 attending schools that use mixed attainment grouping for mathematics, and low-attaining students over Years 7 and 8 attending a similar group of schools that use setting? (a) for all low-attaining students; (b) for low-attaining students receiving FSM.
5. To what extent do (a) OTL; and (b) teacher quality explain any differential outcomes for different grouping practices, for different set allocations or for students at different attainment levels?

A further objective was to analyse the implementation of setting and mixed attainment (IPE):

6. (a) To what extent do schools' specific current grouping practices vary within the groups using setting and mixed attainment? (b) What are the reasons for these practices?
7. How do students with lower prior attainment experience different grouping practices? What are the beneficial and detrimental effects of these experiences associated with setting and mixed attainment.
8. (a) What pedagogic practices do teachers use in Years 7 and 8 mathematics lessons, for different grouping practices? (b) To what extent, if any, are these influenced by the prior attainment of students?
9. What factors associated with the specific school context influence grouping practices?

In addition, the IPE aimed to provide further information to supplement the findings of research question 5 by examining the factors that shape OTL and teaching quality and the extent to which these differ for different grouping practices, for different set allocations and for students at different attainment levels.

Note: For the purposes of this study, low-attaining students were defined as the lowest-attaining tertile at entry to Year 7. This broadly corresponds to those students who do not achieve at the expected level at Key Stage 2 (as per the DfE National Pupil Database [NPD]). High-attaining students were defined as the highest attaining tertile.

The study plan (Hodgen *et al.,* 2024) and Statistical Analysis Plan (Anders *et al.,* 2024) can be accessed via the Education Endowment Foundation (EEF) website. A detailed discussion of threats to validity can be found in Appendix C.22.

[2] Research questions have been re-ordered from how they appear in the Statistical Analysis Plan (Anders *et al.,* 2024). Specifically, research question 1 is research question 2 in the statistical analysis plan, and research question 3 is research question 4 in the statistical analysis plan and vice versa.

Impact of COVID on the study

The Student Grouping Study was originally started in 2019, with schools recruited between January and October and participants being students in Year 7 starting secondary school in September 2019. The study had to be halted and then restarted because the COVID pandemic significantly impacted grouping and teaching practices in English secondary schools (Taylor *et al.,* 2024). All student data collected as part of the original The Student Grouping Study was deleted and the study was restarted in 2022 when school practices had returned to usual.

Ethics and evaluation registration

Full ethical approval was given by the Institute of Education (IOE) Research Ethics Committee at University College London (UCL) (reference REC 1139).

Headteachers agreed to participate in the study by signing a Memorandum of Understanding (MoU) (see Appendix C.1). Full information about the research was provided to parents/carers of all students, students themselves prior to the collection of unique pupil numbers (UPNs) from participating schools (see Appendix C.2). Participating students, or their parents/carers, were given the opportunity to withdraw their data from the research (for ethics purposes; see below for a discussion of data protection considerations). This covered NPD matching, collection of school data, participation in surveys and focus groups, and data archiving. Additionally, opt-in consent was obtained from all students and teachers prior to participation in surveys and focus groups. Parents/carers were informed in advance if their child's class was to be observed. Only students whose parent/carer had given consent were selected as focal students. During interviews and focus groups, participants were reminded that their data would be kept confidential and pseudonymised prior to analysis and reporting. They were told that their participation was voluntary and that they could choose not to answer questions if they so wished, or they could choose to leave the interview early. Permission was obtained from participants for recording and transcription. Finally, participants were advised that the school's safeguarding officer would be informed in the event of a disclosure of harm or potential harm.

Data protection

UCL was the data controller for this project. Schools were data controllers for their own data. Queen's University Belfast, Brunel University, Alphagraphics, AlphaPlus, GL Assessment, and DAMSL Consulting Ltd were data processors for this project. Data sharing between UCL and all other data controllers/processors was governed by data sharing agreements. The project was fully approved for compliance with data protection regulations including the General Data Protection Regulation (GDPR) by UCL's data protection team (registration number: Z6364106/2018/11/03 social research).

Students and their parents/carers, and teachers, were informed of the proposed data processing and given an opportunity to object to this, and withdraw their, or their child's, data. The information provided to parents/carers, students, and teachers included, in clear and plain non-technical language, the details and purposes of data processing including categories of personal data, compliance with GDPR, how to object, and contact details of the organisation.

Link to UCL Data Protection Policy: www.ucl.ac.uk/drupal/legal-services/sites/legal-services/files/migrated-files/DataProtectionPolicy1016.pdf

All personal data collected or obtained as part of this project were treated as 'Highly Restricted' under UCL Data Protection classification guidance. Personal data (student names, UPNs, dates of birth, FSM eligibility, sex, national test results, class, and teacher, as well as teacher names and survey data) were stored, processed, and analysed on the UCL Data Safe Haven (DSH), the technical infrastructure that UCL has built specifically to host sensitive research data.

Qualitative data were pseudonymised. Once pseudonymised they were stored in a secure folder on the UCL network within a project folder only accessible to project team members (using appropriate access control methods), and the pseudonymisation key stored on the DSH. Fieldnotes and audio recordings were stored in a locked filing cabinet within a locked office at UCL to which only the research team had access.

All data transfers between collaborators on this project were carried out using a secure remote connection (e.g. virtual private network [VPN]) between the institutions' networks. In addition, the data was encrypted before sharing using a password shared between parties by separate communication.

Data from the NPD were accessed through the Office for National Statistics (ONS) Secure Research Service (SRS) at an approved Safe Setting. Student-level data were shared securely to the ONS system using end-to-end encryption. Access to the ONS system was solely by ONS approved researchers.

Schools submitted students' personal data to UCL via the DSH's direct data transfer portal. Schools were provided with clear guidance on securely submitting and protecting this data. Online surveys for teachers and students were administered through UCL's REDCap survey system whereby data is uploaded directly to the DSH in an encrypted form.

A risk assessment was conducted for the storage, processing, and transfer of all personal data for the project. All team members undertook regular annual data security training.

The DSH environment is certified to ISO27001 with the British Standards Institution (BSI) – certificate number: IS612909. The hosting is on a thin client system (DSH) with dual factor authentication. This is a multi-user system with permission-based access control. The DSH is subject to penetration testing on an ongoing basis. The DSH has its own firewall separating it from the UCL corporate network and the UCL network has a corporate firewall with a default deny policy for inbound connections. The DSH remote access mechanism is protected by a Secure Sockets Layer (SSL) certificate issued by Terena as well as DualShield dual factor authentication, which couples an Active Directory password with token-based authentication. Connections are AES256 encrypted. Data is transferred into the DSH system via a secure gateway technology, which uses SSL/Transport Layer Security (TLS) with data retained via policy and systems that prevent data leakage.

Project team

The project team was based at UCL IOE and led by Professor Hodgen and Professor Taylor.

- **Professor Jeremy Hodgen (JH).** Project leadership and strategic management, contribution to all aspects of study.
- **Professor Becky Taylor (BT).** Project leadership and management; recruitment and retention, instrument design and analysis.
- **Dr Antonina Tereshchenko (AT) (Brunel University).** Implementation analysis and case studies.
- **Professor Jake Anders (JA).** Additional statistical, design, and methodological expertise and advice: Statistical design and recruitment (matching); quantitative analysis.
- **Dr Maria Cockerill (MC) (Queens University Belfast).** School recruitment, implementation analysis, and case studies.
- **Dr Rosa Kwok (RK).** Research fellow. Developed and validated the teacher and student surveys.
- **Dr Laurie Jacques (LJ).** Research fellow. Provided expertise in qualitative data collection and assisted with carrying out the qualitative component of the implementation analysis and subsequent coding and preliminary analysis of the data.
- **Dr Nicola Bretscher (NB).** Lecturer in mathematics education at IOE. Contributed to data preparation and to the impact analysis.
- **Dr David Sirl (DS) (DAMSL Consulting Ltd).** Conducted the impact analysis.
- **Dr Hester Burn (HB).** Research fellow. Contributed to the impact analysis.

- Administrators provided day-to-day support for the project, including supporting recruitment and data collection.

- Test administrators recruited by UCL and by AlphaPlus supported the delivery of mathematics assessments in schools.

Methods

Evaluation design

Table 3: Evaluation design

Evaluation design		Naturalistic study with matched design
Unit of analysis		Students clustered in schools
No. of units included in analysis (mixed attainment, setted)		89 (mixed attainment 27, setted 62)
Primary outcome	Variable	Attainment in mathematics
	Measure (instrument, scale, source)	GL Assessment Progress Test in Mathematics 13, (69–141)
Secondary outcome(s)	Variable(s)	Self-confidence in mathematics (student self-report)
	Measure(s) (instrument, scale, source)	Survey developed by the research team (for more details see Appendix C.4), (7–35)
Baseline for primary outcome	Variable	Attainment in mathematics
	Measure (instrument, scale, source)	Key Stage 2 mathematics raw score 2022 (NPD), (0–110)
Baseline for secondary outcome(s)	Variable	Self-confidence in mathematics
	Measure (instrument, scale, source)	Survey developed by the research team, as for secondary outcome

Sample size

Minimum detectable effect size (MDES) was estimated for the sample as a whole and for FSM students. Calculations were carried out using the R package PowerUpR. Unfortunately, in standard statistical software including R and Stata, there is no package, routine, or function available to estimate the MDES for the 2-level model specified in the primary analysis section below in the context of a matched sample. Hence, we based these calculations on the closest model specified in PowerUpR: a 3-level fixed effects blocked cluster random assignment design (with treatment at level 2) to account for clustering (of students within schools) and blocking (schools within matched groups at level 3).

These calculations assumed a pre-/post-test correlation of 0.75 (at the student level). On the basis of previous studies (e.g. Hodgen *et al.,* 2019), we expected the pre-/post-test correlation between the Key Stage 2 and Progress Test Mathematics (PTM) to be at least 0.5 and, given several additional student-level covariates were to be added to the model alongside the Key Stage 2 score , we anticipated a substantial increase to the correlation. We assumed the school-level correlation to be 50% of the student level (see e.g. the Key Stage 1 to Key Stage 2 clustered correlations in Allen *et al.,* 2018, Table 2, p. 6), although in practice the school-level correlation was much higher. The school-level intracluster correlation coefficient (ICC) of 0.15 was based on our experience from previous studies (e.g. Evaluation of SMART Spaces Revision Programme, Hodgen *et al.,* 2023). See Table 9 for these calculations alongside the MDES calculations based on the analysed sample and the actual pre-/post-test correlations and ICCs.

Participant selection

Schools were recruited by members of the research team between January 2022 and October 2022. Participants were all students starting Year 7 in September 2022 in recruited schools. They participated in the study until they finished Year 8, in July 2024. Students in the mixed attainment group by then had received two full school years of the 'treatment', i.e. mixed attainment grouping in mathematics.

Two groups of schools were recruited: one group that was already teaching mathematics to Year 7 and Year 8 students in mixed attainment groups (treatment group); and one group that was already teaching mathematics to Year 7 and Year 8 students in groups set by attainment (comparison group).

Schools were eligible to participate in the 'Mixed attainment' treatment group if they met the following criteria:

- State-funded secondary school in England, not selective by 'ability'.
- Currently (or intending to) teach mathematics to Year 7 and Year 8 students in mixed attainment classes. We define mixed attainment classes as those in which the range of attainment in each class broadly reflects the full range of attainment in the year group for that subject. Schools were additionally permitted to have a 'nurture group', in which the very lowest-attaining students were taught separately.[3]

Schools were eligible to participate in the 'Setting' comparison group if they met the following criteria:

- State-funded secondary school in England, not selective by 'ability'.
- Currently (or intending to) teach mathematics to Year 7 and Year 8 students in three or more attainment sets. We define attainment sets as classes in which students are grouped by their attainment in a subject and taught together for that subject.

Schools using streaming were not eligible to participate. We define streaming as the allocation of students to groups for teaching in all subjects, based on a notion of general ability.

Recruitment of mixed attainment schools

Schools were recruited to the 'Mixed Attainment' (treatment) group first. Recruitment started on 11 January 2022. We began by contacting all schools in the mixed attainment group recruited in 2019, aside from those schools that had previously indicated that they had either changed their grouping practices or no longer wished to take part in the study. Of these schools, 17 agreed to take part in the study. Following this, it was necessary to recruit further mixed attainment schools to achieve an acceptable level of power. We estimated that there were between 120 and 190 eligible schools in England teaching mathematics in mixed attainment groups to Year 7 and Year 8 (Taylor *et al.,* 2020). As we did not know the identity of all these schools in advance, recruitment took place through simultaneous processes of publicising the study in education media, using existing contacts and networks to advertise the study to schools, and cold-calling schools to ask about grouping practices. As a result, a further 15 schools were identified as teaching mathematics to Year 7 and Year 8 students in mixed attainment classes and agreed to participate in the study by returning a completed MoU, resulting in a mixed attainment group of 32 schools.

---

[3] Schools were permitted to operate a 'nurture group' and still be included in the mixed attainment group, provided that this was the only grouping by prior attainment used mathematics. Nurture groups are typically small groups of students who are not deemed 'secondary school ready' and so are taught separately to their peers in order to aid transition from primary to secondary school (Cooper and Whitebread, 2007; Mazenod *et al.,* 2018).

Matching

After recruitment of mixed attainment schools, statistical matching (prioritising a high degree of balance in terms of the average and distribution of schools' intakes) was carried out to identify schools that were similar to the mixed attainment group of schools aside from their attainment grouping policy. This exercise was conducted on 16 June 2022 using the set of all state secondary schools in England (excluding the recruited mixed attainment schools) and, as a result, a pool of potential comparison schools was identified.

This matching was carried out using the R (R Core Team, 2019) package MatchIt (Ho *et al.,* 2011). We used propensity score matching (PSM) to identify matched schools. PSM is our preferred method, because the PSM model takes account of all the observed characteristics of schools that we believe to be important.

In order to match the schools, we identified a range of school-level variables that we hypothesised both influenced a school's decision to adopt the particular grouping practice (i.e. mixed attainment or setting) and influenced the outcome (i.e. student attainment in mathematics at the end of Year 8), but were not influenced by this outcome (Caliendo and Kopeinig, 2008). (See Appendix C.3 for an overview of this identification process.)

The approaches to matching for the primary analysis of this study were limited because of the need to identify fixed treatment and control samples on which to collect outcome and implementation data. We considered the advantages and disadvantages of two broad approaches to matching that would allow this: PSM; and coarsened exact matching (CEM). We anticipated that both approaches would involve the loss of recruited schools through the matching process, either through the imposition of common support in PSM or the failure to identify matched comparison schools within the tolerances chosen for each variable in CEM. PSM was preferred because of the large number of variables identified for matching. A simulation exercise (see Appendix C.21) was conducted that provided evidence for the feasibility of the approach and indicated that identifying 25 setted schools as potential matches for each recruited mixed attainment school would result in a sufficient response.

It was envisaged that the matching would be conducted on the full set of variables identified for matching. This proved difficult to achieve, largely due to the small number of mixed attainment schools nationally and to the unusual nature of these schools. As a result, we conducted the matching exercise using key, mainly attainment, variables, because these were judged to play the most significant role. (See Appendix C.20 for details of the model.)

PSM produced a subset of 25 unique potential matched schools for each of the 32 mixed attainment schools, a list with a total of 800 schools, which was used as the basis for recruitment. In the event (see below), matched schools were recruited for 30 of the 32 mixed attainment schools. This was judged to be an acceptable level attrition to allow for reasonably strong grounds for generalisation to the broader population of mixed attainment schools. Common support was implemented by dropping the unmatched mixed attainment schools.

Recruitment of setting schools

Schools were then recruited to the setting (comparison) group. Recruitment took place between 17 June 2022 and 25 November 2022.

All 800 matched schools were approached by email to take part in the study. The hope was to recruit three matched comparison schools for each recruited mixed attainment school. Of the schools that responded to these invitations, 80 were judged eligible and agreed to take part in the study, a response rate of 10% of the total schools approached (which is below the positive response rate of 20% as estimated in the study plan Hodgen *et al.,* 2024, but above the 7.3% rate achieved in the initial 2019 recruitment exercise). The number of matched setting schools recruited for each mixed attainment school varied from two to seven. Fifty of the schools approached were matched to two mixed attainment schools that did not supply student data and so were ineligible.

Although 30 mixed attainment and 80 matched setting schools had agreed to participate, by the point that recruitment ended some of these had not shared student data with us. Two mixed attainment schools and seven setting schools did not share student data, meaning that there was some further loss of matched schools. Four setting schools were lost, as they were matched to the two mixed attainment schools that did not share data. One mixed attainment school was lost, as it was matched only to setting schools that did not share data.

The final set of 27 mixed attainment schools and 69 matched comparison (setting) schools as at 25 November 2022 (the date recruitment closed, aka pseudo-randomisation date) is used as the school-level analysis sample.

In the participant flow diagram (see page 44) we have only included as recruited those schools that shared data with us. We have reported that 29 mixed attainment schools were recruited, and 750 matched setting schools approached for recruitment. Schools that did not share data, and their matched schools, are considered not to have been recruited to the study.

Given the variable number of recruited matched comparison schools for each mixed attainment school, we used weighting to reduce the importance for estimation of students in schools where we recruited multiple matched comparator schools for each mixed attainment school; the sample was weighted (to reflect the number of matched comparator schools recruited corresponding to each mixed attainment school and to reflect the number of students in each school) such that students in each mixed attainment school have the same level of importance as all the schools that were identified and recruited as its matched comparators.

As described above, we were unable to recruit a match for two of the mixed attainment schools, and one mixed attainment school and four setting schools were excluded because their matched schools did not share data. These seven schools are therefore, excluded from the balance tables and the power calculations for the primary analysis. However, we retained these seven schools in the study for the investigation of implementation as well as both the student level and sensitivity analyses. Hence, following recruitment we intended to carry out our primary analysis with 97 schools, 28 mixed attainment and 69 setting. These are in 28 matched groups as set out in Table 4.

Table 4: Matched groups of schools in recruited sample of schools with primary data

Matched group	No. of mixed attainment schools	No. of setted schools	Total schools
1	1	5	6
2	1	2	3
3	1	2	3
4	1	2	3
5	1	1	2
6	1	3	4
7	1	1	2
8	1	1	2
9	1	1	2
10	1	1	2
11	1	2	3
12	1	2	3
13	1	4	5
14	1	1	2
15	1	2	3
16	1	6	7
17	1	2	3
18	1	3	4
19	1	1	2
20	1	4	5
21	1	5	6
22	1	3	4
23	1	5	6
24	1	2	3
25	1	1	2
26	1	2	3
27	1	3	4
28	1	2	3
Total	**28**	**69**	**97**

Outcome measures

Baseline measures

The baseline measures were prior attainment in mathematics using Key Stage 2 test raw scores from 2023, obtained from the NPD, and self-confidence in mathematics, measured using survey items[4] developed for the BPGS project (see Roy *et al.,* 2018a, p. 16). The baseline survey was administered in the Autumn Term of Year 7 by school staff as a machine-read, paper survey.

Figure 2 and Figure 3 below present the distributions of Key Stage 2 mathematics scores. The distributions for students in mixed attainment and setting schools are similar. Both distributions are typical for Key Stage 2 National Test Scores and, as would be expected, there was some evidence of a ceiling effect. However, the measure has good statistical properties.[5]

Figure 2: Distribution of Key Stage 2 National Test Scores (KS2_MATMRK) in mathematics for students in mixed attainment (blue) and setted (red) schools

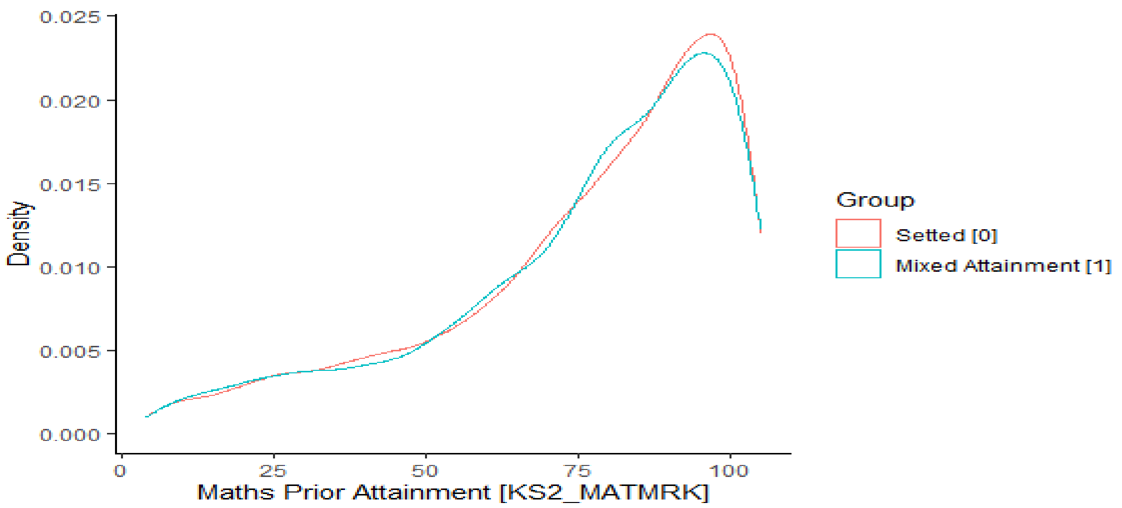
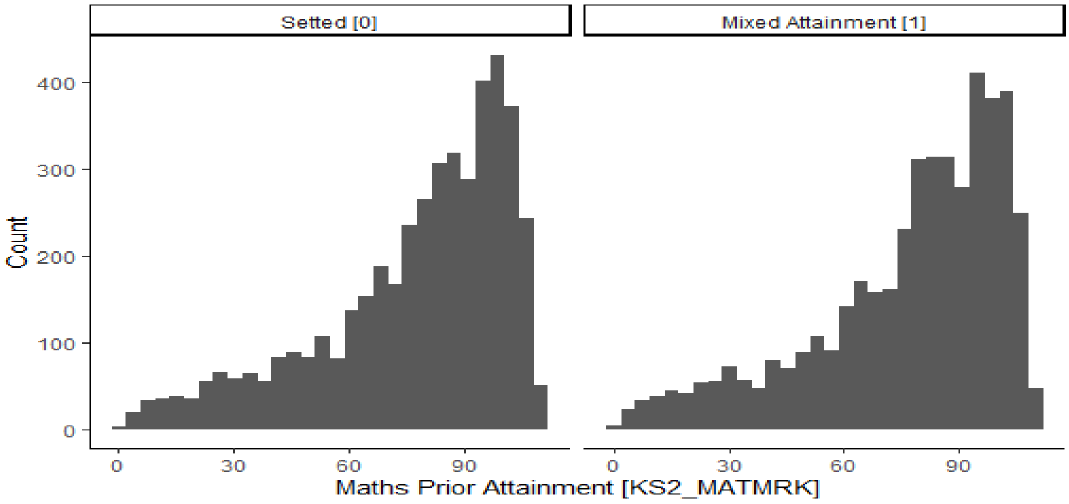


Figure 3: Histograms showing the distributions of Key Stage 2 National Test Scores (KS@_MATMRK) for students in the setted and mixed attainment schools



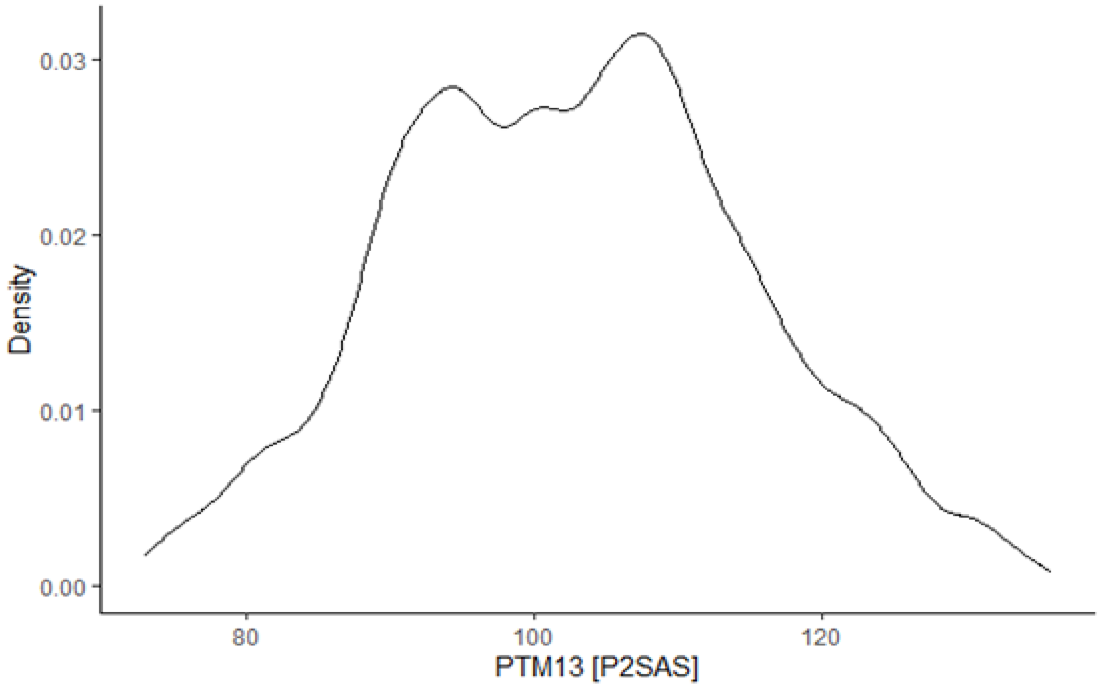
[4] See Francis *et al. (*2017) for further information about the self-confidence survey items.
[5] See: www.gov.uk/government/publications/national-curriculum-test-development-handbook.

Primary outcome

The primary outcome measure was attainment in mathematics, measured using the paper version of the GL Assessment PTM[6] (PTM13), with tests carried out in schools under exam conditions at the end of the second year of the study (students at the end of Year 8). Tests were independently marked by GL Assessment and marks returned securely to the research team.

Figure
4 below presents the overall distributions of the PTM13 test. This does not suggest any skewing or either floor or ceiling effects.

Figure 4: Distribution of primary outcome measure PTM13 (P2SAS) Test Scores for all students



Wherever possible, in order to avoid attrition, and consequent bias, collection of the attainment outcome data was supported by test administrators, who were unaware of the study group to which the school belonged. Schools administered tests themselves in 29 schools (nine mixed attainment schools and 20 setting schools). A financial incentive of £1,000 was offered to both study groups, payable at the end of the second year of the study when data collection was complete.

Secondary outcomes

The secondary outcome measure was students' mathematics self-confidence, measured using the same survey items as at baseline. Figure 5 and Figure 6 below present the distribution of the mathematics self-confidence scores. The distributions for students in mixed attainment and setting schools are similar, although there is some evidence of a floor effect.

---

[6] See www.gl-assessment.co.uk/assessments/progress-test-in-maths/. The instrument is available commercially and so is not able to be included in this evaluation report.

Figure 5: Distribution of mathematics self-confidence scores at baseline (rawMSCbase_scale) for students in mixed attainment (blue) and setted (red) schools

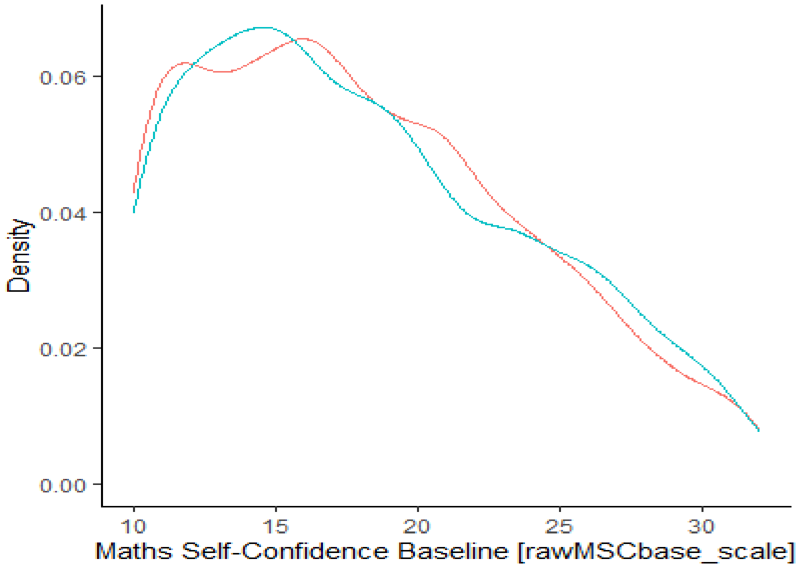
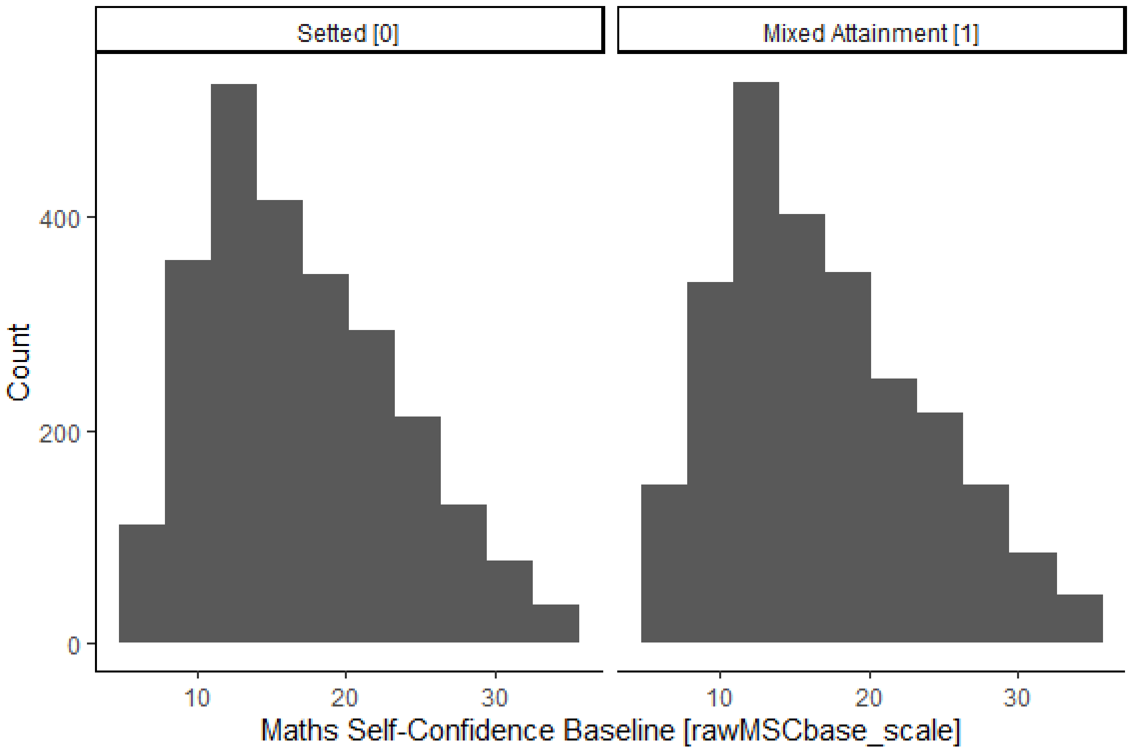


Figure 6: Histograms showing the distributions of mathematics self-confidence scores at baseline (rawMSCbase_scale) for students in setted and mixed attainment schools



Additionally, we used general self-confidence survey items and adapted items from PISA (Programme for International Student Assessment) measuring students' orientation towards further participation in mathematics (OECD, 2013b) as additional secondary outcomes. Endline surveys were completed by students in the Summer Term of Year 8 and were administered online by school staff.

Statistical analysis

Primary analysis

Using this weighted student-level sample, we estimated the following linear regression model (Model A) to address research question 2 (note that, as discussed in the **study plan**, we do not use the EEF's standard approach as it is unclear whether this is appropriate within the context of a matching framework; Hodgen *et al.,* 2024):

$$y_{ij} = \alpha + \beta_1 Mixed_j + \boldsymbol{X_{ij}} + \varepsilon_{ij}$$

where y is the outcome variable for student i in school j, α is an intercept term, β_1 recovers the average difference in outcome performance associated with being in a mixed attainment school rather than a comparable setting school, $\boldsymbol{X_{ij}}$ is a vector of student- and school-level control variable characteristics (aiming to further reduce bias on top of that reduced through the matching approach and to increase the precision of our estimate of β_1), and ε is an individual-level error term. We note that the matching process was conducted at school level using PSM and some of the attainment-related matching variables are dated. Hence, the inclusion of these additional covariates is considered a reasonable approach to reducing bias. Standard errors were calculated taking into account the school-level clustering; as noted above this is as an alternative to use of school-level random effects in the model but, which does not require assumptions regarding the distribution of these school-level effects. The analysis was conducted on an intention-to-treat (ITT) basis.

The vector of student- and school-level control variables ($\boldsymbol{X_{ij}}$) consisted of the following covariates:

- individual Key Stage 2 prior attainment;
- school average Key Stage 2 attainment of intake;
- school low-prior attainment proportion;
- school high-prior attainment proportion;
- school cohort number of students;
- individual FSM;
- school FSM proportion;
- individual English as an Additional Language (EAL) status;
- school EAL proportion;
- school academy status;
- individual Income Deprivation Affecting Children Index (IDACI);
- school composition IDACI quintile;
- Ofsted (Office for Standards in Education, Children's Services and Skills) grade; and
- urban/rural classification.

A core part of this project's aim was to explore the potential distributional changes (research question 1). We carried this out using the following model (Model B), using the same weighted student-level sample as for the primary analysis model described above.

$$y_{ij} = \alpha + \beta_2 Mixed_j + \beta_3 HighAtt_i + \beta_4 LowAtt_i + \beta_5 Mixed_j * HighAtt_i + \beta_6 Mixed_j * LowAtt_i + \boldsymbol{X_{ij}} + \varepsilon_{ij}$$

where, in addition to the elements already defined in the primary analysis model above, β_2 recovers the average difference in outcome performance associated with being in a mixed attainment school, rather than a comparable setting school, among average attainment students (i.e. those who are neither 'Low Attainment' nor 'High Attainment'); β_3 recovers the difference in outcome performance for higher attainers in setting schools, β_4 recovers the difference in outcome performance for low attainers in setting schools, β_5 recovers the difference in outcome performance among high attainers in mixed attainment schools compared to high attainers in setting schools, β_6 recovers the difference in outcome performance among low attainers in mixed attainment schools compared to low attainers in setting schools. As such, $\beta_5 - \beta_6$ recovers the difference in differences between high attainer and low attainer groups associated with being in a mixed attainment school, rather than a comparable setting school. The effect for high attainers is included because our previous work indicates that the 'attainment gap' between low and high attainers is more pronounced than that between low and 'average', or middle, attainers. [7,8]

Secondary analysis

The secondary outcome analysis used the same approaches to modelling as in the main analysis to address research questions 3 and 4 comparing the effects on student self-confidence in mathematics, using a model analogous to Model A for research question 4 (all students) and a model analogous to Model B for research question 3 (low attainers). There are two separate self-confidence measures, general self-confidence and mathematics self-confidence. Each is a 7-item scale, with each item scored on a 5-point Likert scale and the scale score being the total score for the 7 items. The scales can be found in Appendix C.4.

Imbalance at baseline

Given the likely importance of inequality in grouping, it is particularly important to consider balance in measures of centrality, spread, and skewness (1st, 2nd, and 3rd order moments). We tested balances between groups on a fixed set of variables for means, medians, standard deviations (SDs), and skewness at both school and student levels, and for the subgroup of students eligible for FSM. As in Table 2, we report the comparison of means for the analysed sample with actual student-level data in a similar way to that required for a standard RCT trial by the EEF, with standardised differences using Glass's delta (arithmetically the same, but conceptually different to effect sizes in this setting; Table 11). (See Appendix C.19.) Unstandardised differences in means, medians, SDs, and skewness are also reported. We also plot overlapping kernel density plots of these characteristics between the treatment (mixed attainment) and matched comparison (setting) groups to give an overall impression of the different distributions. As discussed above under the 'Matching' and 'Recruitment of setting schools' sections, we compared our treated sample with the following samples and discuss any differences, particularly with all English schools in order to consider whether there are any factors that may be particularly associated with schools that adopt mixed attainment grouping:

- all English schools;
- pool of potential comparators identified by matching; and
- recruited comparison sample.

Analysis in the presence of non-compliance

If non-compliance was found to be present, we planned to explore compliance using a Complier Average Causal Effect (CACE)/instrumental variables analysis. All schools complied with the eligibility criteria and did not change their grouping practices, which had been a commitment in the MoU. Hence, this analysis was not conducted.

---

[7] Note that in our previous study we compared groups of students based on set membership (top set compared with the middle set, and bottom set compared with the middle set), whereas in this research we compare students based on prior attainment tertiles.

Compliance was defined according to whether the schools met the eligibility criteria throughout the period of research, i.e. during the academic years 2022/2023 and 2023/2024. We are aware that a small number of schools changed their grouping practices prior to the beginning of the academic year, 2022/2023. The eligibility criteria were:

- **For mixed attainment schools.** They teach mathematics to Year 7 and Year 8 students in mixed attainment classes. Mixed attainment classes are defined as those in which the range of attainment in each class broadly reflects the full range of attainment in the year group for that subject. Mixed attainment schools may additionally have, or introduce, a 'nurture group', in which the very lowest-attaining students are taught separately. Hence, we assessed compliance by examining whether mixed attainment schools have, in effect, one or more high-attaining 'sets', where the average attainment of the class is above the 75th percentile for the school as a whole.

- **For setting schools.** They teach mathematics to Year 7 and Year 8 students in three or more attainment sets. Attainment sets are defined as classes in which students are grouped by their attainment in a subject and taught together for that subject. Since this study focuses on the effects of grouping by subject attainment compared to mixed attainment grouping, schools that introduced streaming would have been considered non-compliant.

School compliance was assessed on the basis of a simple question in the head of mathematics surveys.

Missing data analysis

We summarised the extent of any missing data in the primary and secondary outcomes, and in the model associated with the analysis. Where possible, reasons for any missing data are described to enable a judgement of whether the data is missing at random (MAR).

If more than 10% of school-level or student-level data in the model is missing (based on the finalised matched sample), we implemented the following pre-planned missing data strategy. The strategy was followed separately for main primary analysis model and attainment outcome variable. We first explored whether there was evidence that the missing data is MAR, since this is a pre-requisite for missing data imputation modelling to produce meaningful results. To do this we created an indicator variable for each variable in the impact model specifying whether the data is missing or not. We then used logistic regression to test whether this missing status can be predicted from the variables used for imbalance testing (listed above). Where predictability was confirmed we proceeded to use these same variables to estimate a multiple imputation (MI) model using a fully conditional specification, implemented using MICE in R (rather than Stata MI as in the Statistical Analysis Plan; Anders *et al.,* 2024) to create 20 imputed data sets; we believe this is an appropriate number of imputed datasets given the anticipated level of potential missing data as a result of the administrative data source we are employing (Graham *et al.,* 2007). We then re-estimated the treatment effect and took the average and estimated standard error using Rubin's (2004) combination rules.

Analysis using the dataset produced through either of these missing data strategies was used as a sensitivity analysis i.e. we based confirmation of the effectiveness of the treatment on complete case analysis only but assessed the sensitivity of the estimate to missingness using the estimates from the multiply imputed dataset.

Subgroup analyses

We conducted a subgroup analysis for FSM students. In line with the EEF's guidance (EEF, 2022) for subgroup analysis for RCTs, we first ran separate subgroup analyses identical to the primary analysis model on the sample defined as falling in the FSM subgroup and also for the students with low-prior attainment who also fell into the FSM group. As a robustness check, we estimated a model on the full sample adding a covariate for our subgroup of interest and an interaction between this covariate and the treatment indicator.

Additional analyses

Mediation analysis

In order to investigate research question 5, we conducted an exploratory analysis focused on two potential mediators, OTL and teacher quality[9] as identified in the logic model outlined in the study plan (Hodgen *et al.,* 2024). We developed a bespoke survey instrument for completion by students to measure OTL as described in Appendix C.5. In order to measure teacher quality, we used two dimensions of the teacher quality student survey developed by Evidence-Based Education (2022): *understanding the content*; and *activating hard thinking*. See **study plan** (Hodgen *et al.,* 2024) for further details on the instrument.

To explore the role of OTL and teacher quality in mediating the overall changes in the primary outcome that we observe (research question 5), we estimated mediation models (under assumptions of sequential ignorability and no interaction between treatment and mediator) using the approach proposed by Imai *et al.*, 2010) to carry out appropriate inference. For the sequential ignorability assumption to be met:

> ...two ignorability assumptions are made sequentially. First, given the observed pretreatment confounders, the treatment assignment is assumed to be ignorable, that is, statistically independent of potential outcomes and potential mediators. ...[Second] the mediator is ignorable given the observed treatment and pretreatment confounders. ...In other words,...among those [participants] who share the same treatment status and the same pretreatment characteristics, the mediator can be regarded as if it were randomized. (Imai *et al.,* 2010, pp. 312–313).

Imai *et al.*'s (2010) approach involves the estimation of the following linked models:

$$OTL_j = \alpha_1 + \beta_7 Mixed_j + \boldsymbol{X_{ij}} + \varepsilon_{ij1}$$

$$Quality_j = \alpha_2 + \beta_8 Mixed_j + \boldsymbol{X_{ij}} + \varepsilon_{ij2}$$

$$y_{ij} = \alpha_3 + \beta_9 Mixed_j + \gamma_1 OTL_j + \gamma_2 Quality_j + \boldsymbol{X_{ij}} + \varepsilon_{ij3}$$

where y is the outcome variable for student i in school j, mediated by OTL and $Quality$ measured in school j; α are model-specific intercept terms, $\boldsymbol{X_{ij}}$ is a vector of student- and school-level control variable characteristics (aiming to further reduce bias on top of that reduced through the matching approach and to increase the precision of our other estimates), and ε are individual-level error terms. From these models the estimate of β_7 recovers the average difference in OTL associated with being in a mixed attainment school rather than a comparable setting school, while β_8 recovers the same but for teacher quality; $\beta_7\gamma_1$ recovers the average difference in our outcome of interest mediated through OTL and $\beta_8\gamma_2$ recovers the same for the difference mediated through teacher quality ('mediated effects'), finally β_9 recovers the average difference that is not mediated by through OTL or teacher quality ('direct effects') (see Imai *et al.,* 2010, pp. 313–314). Underlying this approach is the assumption of independence between treatment assignment and the potential outcomes and mediators (OTL and $Quality$). Although treatment allocation is not at random, we consider this a reasonable assumption given Weidmann and Miratrix's (2021) evidence indicating the effectiveness of a matching approach.[10] We used the R package mediation (Tingley *et al.,* 2014)[11] to implement this approach and conduct a sensitivity analysis. See Appendix C.15 for pseudo-code to illustrate this proposed analysis. We conducted a subgroup analysis for students with low-prior attainment using an analogous modelling process.

---

[9] Our original plan was to develop and validate an instrument to measure teachers' mathematical knowledge, drawing on an approach used in the German COACTIV study (Baumert *et al.,* 2010). However, the results of the validation process strongly indicated multi-dimensionality that did not reflect the Baumert *et al.,* (2010) theoretical model of teacher knowledge that we are using. For further details, see Hodgen *et al.,* (2024, p. 13).

Sensitivity and robustness analyses
We conducted several sensitivity and robustness analyses as follows:

- **Quantile analysis.** As a robustness check on the analysis of distributional change, we used quantile regression to explore distributional changes in performance associated with being in a school with setting compared to being in a mixed attainment school.
- **Comparison of alternative student-level matched samples.** As a sensitivity check on the main analyses, we also carried out a replication of the primary analysis using a sample constructed from an additional student-level matching exercise on all data gathered from mixed attainment and setting schools, including students from the two mixed attainment schools for which matched setting schools were not able to be recruited. This was carried out using the same matching variables considered for inclusion in the school-level matching exercise, along with their student-level counterparts. This approach was not feasible as our primary approach because student-level matching was not possible before the recruitment of schools, which required the matching exercise to have been completed; furthermore, it is appropriate to match at the level of the treatment variation (i.e. at the school level in this case).
- **Self-administered outcome testing.** We conducted a sensitivity analysis to examine whether schools administering outcome testing independently influence outcomes (see Appendix C.16).
- **Rosenbaum bounds**. We had planned to explore the sensitivity of our primary analysis results using the approach discussed by Rosenbaum (2015), as recommended by the EEF (EEF, 2019).

There is an R package developed by Rosenbaum for this purpose for the setting where there are matched sets with variable numbers of controls, which is the case in our analysis (sensitivitymv). This approach is considered appropriate to investigate a difference that is considered statistically significant at a conventional level. Since no such difference was found, this analysis was not conducted.

Estimation of effect sizes

Hedges' g effect size was calculated as follows:

$$\boldsymbol{g = j\left(\frac{\overline{x_1} - \overline{x_2}}{\widehat{s*}}\right)\sqrt{\lambda}}$$

where our conditional estimate of $\overline{\mathbf{x_1}} - \overline{\mathbf{x_2}}$ is recovered from $\boldsymbol{\beta_1}$ in the primary ITT analysis model; $\widehat{s*}$ is estimated from the analysis sample as follows:

$$s^* = \sqrt{\frac{(n_1 - 1)s_1{}^2 + (n_2 - 1)s_2{}^2}{n_1 + n_2 - 2}}$$

where n_1 is the sample size in the control group, n_2 is the sample size in the treatment group, s_1 is the SD of the control group, and s_2 is the SD of the treatment group (all estimates of SD used are unconditional, in line with the EEF's analysis guidance (EEF, 2022) to maximise comparability with other trials); and j and λ are calculated as follows:

$$\boldsymbol{j = 1 - \left(\frac{3}{4h - 1}\right)}$$

$$\boldsymbol{\lambda = 1 - \left(\frac{2\left(\frac{n}{m}\right)p}{n - 1}\right)}$$

where $\boldsymbol{p}$ is the ICC, $\boldsymbol{n}$ is the total sample size (students), and $\boldsymbol{m}$ is the total number of clusters, and $\boldsymbol{h}$ is defined as follows:

$$h = \frac{\left[(n-2) - 2\left(\frac{n}{m} - 1\right)p\right]^2}{(n-2)(1-p)^2 - \frac{n}{m}\left(n - 2\frac{n}{m}\right)p^2 + 2\left(n - 2\frac{n}{m}\right)p(1-p)}$$

Confidence intervals (CIs) for the effect size were calculated by substituting the upper and lower bound CIs from the estimate of β_1 into the above equation (other aspects remaining fixed) and using these as upper and lower CIs for the effect size.

Estimation of ICC

In order to estimate the ICC of the outcome measures at school level we employed an empty variance components model, as follows:

$$Y_{ij} = \alpha + \eta_j + \varepsilon_{ij}$$

where Y_{ij} is the relevant outcome for student i who is in school j, η_j is a school-level random effect, and ε_{ij} is a student-level error term. The school-level random effect is assumed to be normally distributed and uncorrelated with the student-level errors.

The ICC itself was estimated from this model using the following equation:

$$\rho = \frac{var(\eta_j)}{var(\eta_j) + var(\varepsilon_{ij})}$$

Implementation and process evaluation

Research methods

As outlined above, the study took place in a natural context and compared two matched groups of schools that implement contrasting approaches to grouping students for mathematics lessons: mixed attainment; and setting. In contrast to an RCT, the study did not aim to evaluate the impact and implementation of a well-described intervention. Rather, the aim of our IPE was to investigate how these broad approaches are understood and implemented across schools and classes. Our purpose was to document and compare students' mathematical experiences, the pedagogical practices of their teachers, and the contextual factors that influence these experiences and practices with the aim of characterising effective and equitable practices for both mixed attainment and setting, i.e. practices that enable all students to make progress and have positive mathematical experiences regardless of prior attainment or background. The EEF current IPE guidance (EEF, 2022) is designed for tightly defined, manualised interventions and focuses on the extent to which implementation matches the requirements specified by the developer of the intervention. In contrast, this study was a natural experiment and sought to understand the different ways in which schools implement either 'mixed attainment grouping' or 'setting'. Hence, a focus on fidelity of implementation was not appropriate. We therefore adapted the EEF IPE dimensions as described below.

We examined the following dimensions of implementation:

- **Programme differentiation.** Any variation in schools' grouping practices between and within the mixed attainment and setting arms, the reasons underlying schools' choices, (research question 6), including the reasons underlying any non-compliance to the eligibility criteria for each approach and for any changes made during the trial (supplementing the compliance effects analysis).

- **Students' mathematical experiences.** Any differences in students' experiences between and within the mixed attainment and setting arms and between students of different prior attainment, including OTL (research questions 7 and 5).

- **Responsiveness of pedagogy.** Any differences in the pedagogical practices used by teachers and teaching quality between and within the mixed attainment and setting arms and between students of different prior attainment, with a particular focus on the needs of students with low-prior attainment. This included consideration of set movement (research questions 8 and 5).

- **Contextual factors.** Factors that explain any differences in grouping practice and how they were implemented (research question 9). This included whether there were any factors that especially enabled mixed attainment practice within the mixed attainment schools, and whether there are any particular barriers to mixed attainment which are less evident in the types of schools that do mixed attainment.

In addition, the IPE, together with research question 5 of the effects analysis, examined the extent to which the hypothesised logic model provided a plausible explanation of the processes by which the different grouping approaches impacted on student mathematical outcomes and whether, and how, this logic model should be adapted.

Theoretical framework

The theory of change that underpinned this study is outlined above and the logic model is found in Figure 1. Central to this model and to this IPE is the concept of students' mathematical learning experiences and how teaching quality and OTL affect these. Our understanding of classroom mathematical learning experiences was underpinned by the 'Teaching for Robust Understanding (TRU) framework' (Schoenfeld and the Teaching for Robust Understanding Project, 2016), which allowed us to evaluate the quality of a mathematics learning environment from the students' perspective. The framework enabled us to compare opportunities for lower and higher prior attainers with respect to the following five dimensions:

- **Mathematical Content**. The extent to which the mathematics discussed was clear, correct, and well justified (tied to conceptual underpinnings).

- **Cognitive Demand.** The extent to which students' sense-making and 'productive struggle' was promoted and they were engaged in classroom interactions that created and maintained an environment of intellectual challenge.

- **Access.** The extent to which classroom activity structures invited and supported active engagement for a diverse range of students so that they experienced meaningful and equitable access to concepts and mathematical practices.

- **Agency, Ownership, and Identity.** The extent to which students' experiences supported the construction of positive disciplinary identities, for example, opportunities for them to make mathematical conjectures, explanations, and arguments, developing 'voice' (agency and authority) while adhering to mathematical norms (accountability).

- **Formative Assessment.** The extent to which students' reasoning was elicited, challenged, and refined and how the classroom environment was responsive to student thinking.

In the IPE we extended the focus on OTL as time and content by examining OTL as captured through the mathematical content, cognitive demand, and access to mathematics dimensions of the TRU, supplemented with evidence about the third aspect, the quality of teaching, together with evidence about the intended OTL established from school documentation (e.g. schemes of work, policies), and interviews with heads of mathematics.

Teaching quality is challenging to capture. There is evidence that teacher quality is associated with teacher knowledge (Baumert *et al.,* 2010), but we were additionally interested in teachers' pedagogic strategies in the classroom. Teacher responsiveness can be captured to a limited extent through the formative assessment dimension of the TRU. Additionally, we drew on the 12 evidence-based strategies and approaches identified in Hodgen *et al.'s* (2020) secondary meta-analysis as having the potential to improve teaching and learning of mathematics for low-attaining students when used in a balanced manner. These strategies are summarised in Table 5. We note that, for each strategy, the manner in which the strategy is implemented is crucial to its likely effectiveness. Hence, we used the TRU in order to differentiate the quality of how these strategies are implemented.

Table 5: Twelve evidence-based strategies for improving teaching and learning for low-attaining students (adapted from Hodgen *et al.,* 2020; see also Hodgen *et al.,* 2019; EEF, 2018)

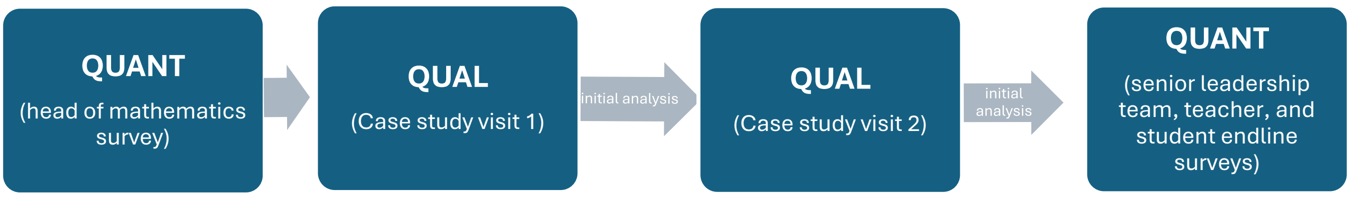
Explicit teaching	A wide variety of teacher-led approaches consisting of crafted instruction, often associated with well-structured practice
Computer-aided instruction	Computer-based systems designed to supplement learning (e.g. Sparx Maths, Heggarty Maths)
Peer tutoring	Tutoring by same-age peers
Heuristics	Explicit strategies for approaching and solving a range of different problems
Manipulatives	Concrete materials that can be manipulated by students to aid understanding and to support the development of mathematical understanding, guided by explicit teaching
Tutoring by adults	Group or one to one additional support provided by teachers or teaching assistants
Feedback to students	Information provided to a student regarding performance or understanding
Representations	Diagrams, graphs, and tools such as number lines, including concrete-pictorial-abstract approaches. Evidence suggests that students need to experience, and compare, multiple representations, but care needs to be taken not to overload students
Feedback to teachers	Information or data on student performance or understanding provided to a teacher
Self-instruction	A set of prompts for students providing structure, or an aide-memoire, for a particular method or set of methods
Cooperative learning	Students collaborate on a shared task in structured programmes, often in groups of mixed attainment
Student-centred learning	A range of student-led or student-mediated approaches to learning, including guided learning approaches

As well as teaching quality and OTL, we anticipated that classroom mathematical experiences would be influenced both by class size and by the ability-labelling of students (Francis *et al.,* 2017; Tereshchenko *et al.,* 2018). Class size was recorded as part of case studies, and was also collected with student data provided by schools. Labelling was explored through teacher interviews and student interviews and focus groups as part of case studies.

Design

The IPE had a sequential mixed-methods design with quantitative surveys and in-depth qualitative case studies. The overall design was informed by the logic model and the methodological framework and is summarised in Figure 7.

Figure 7: Summary of mixed-methods IPE design



Sampling

Surveys

IPE surveys were sent to all those in participating schools who had not been withdrawn from data processing for the purposes of this study:

- Student survey items for the IPE were administered in the same endline survey as the secondary outcome measure items, to all Year 8 students.
- Head of mathematics, mathematics teacher, and senior leader surveys were sent to all relevant staff in each participating school.

Case studies

The case study sample included 12 schools, including six schools using setting and six using mixed attainment grouping.

There was an additional £500 payment available for case study schools in addition to £1,000 for completing the requirements of the impact evaluation.

Random selection of case study schools

Following Maxwell *et al.'s* (2021) assessment of high-quality sampling for qualitative work in IPEs, we used a random sampling approach. Our random case study school sample was stratified by prior attainment in line with evidence that grouping practices are associated with prior attainment of students in secondary schools in England (see Taylor *et al.,* 2020). Full details of the random sampling can be found in Appendix C.23.

The first three schools in each stratum were approached to participate as a case study school. Where a school declined, the next school in the respective stratum was approached until three case study schools in each stratum had agreed to participate. All lower-attaining mixed attainment schools accepted the invitation. Two schools in the higher-attaining mixed attainment school group rejected the invitation and were replaced with the next schools on the list. Three schools in the lower-attaining setting group and four schools in the higher-attaining setting group rejected the invitation and were replaced with the next schools on the list.

The characteristics of the case study schools are summarised in Table 6. Nearly all schools were graded 'Outstanding' or 'Good' by Ofsted, with the exception of school S2 which was graded 'Requires improvement'. Nine schools were mixed sex, while three schools were girls only (there were no boys' schools recruited to the study).

Table 6: Recruited case study schools. %FSM is rounded to the nearest 5%

School	Prior attainment	FSM %	Gender of entry	Ofsted rating	Nurture group
MA1	Lower	20	Mixed	Good	Yes
MA2	Lower	20	Mixed	Good	No
MA3	Lower	20	Mixed	Good	No
MA4	Higher	25	Girls	Outstanding	Yes
MA5	Higher	25	Mixed	Good	No
MA6	Higher	15	Mixed	Good	No
S1	Lower	20	Mixed	Good	–
S2	Lower	25	Mixed	Requires improvement	–
S3	Higher	10	Girls	Outstanding	–
S4	Higher	25	Girls	Good	–
S5	Higher	15	Mixed	Good	–
S6	Lower	20	Mixed	Good	–

Data collection

A complete overview of the data collection methods in relation to the IPE dimensions and research questions is presented in Table 7.

Surveys

Four surveys were administered as part of the IPE[12]:

1. Head of mathematics online survey in the first year of the study to gather data on schools' specific current grouping practices in Year 7 and Year 8, as well as the reasons for these practices. The survey instrument is

[12] In addition to these four surveys there is a student survey administered to Year 7 students, collecting baseline self-confidence data for the effects analysis.

included in Appendix C.7 and included a mixture of closed and open questions and took approximately 15 minutes to complete.

2. Mathematics teacher online survey in the second year of the study to collect data on their training, qualifications and teaching experience, beliefs and attitudes about attainment grouping and mixed attainment, as well as effects on students; and pedagogic practices such as differentiation. This survey contained only closed questions and took approximately 15 minutes to complete.[13]
3. Senior leader online survey in the second year of the study, to collect data on school ethos and strategic decisions about pedagogy and grouping practices. This survey included a mixture of closed and open questions and took approximately 15 minutes to complete.
4. Student endline online survey to collect data on perceptions of teaching; OTL; mathematics aspirations; and attitudes towards grouping practices. This survey included only closed questions and took approximately 30 minutes to complete. This survey addressed research questions for both the IPE and the outcome evaluation.

Surveys completed by adults were emailed directly to participants. Surveys completed by students were completed online in school, under the supervision of a teacher.

Case studies

The focus in the case study work was both on within-case and on between-case comparison (Miles *et al.,* 2019). Within each case study school, we compared the experiences of students with higher and lower mathematics attainment. In the between-case analysis we compared the experiences of students in mixed attainment schools with students in setting schools.

The focus on configurations and associations within a specific case (case-oriented approach) helped us to understand the experiences of students at different attainment levels in relation to grouping practices and teaching approaches within the case study schools. The within-case approach also provided detailed evidence on how attainment grouping in mathematics is implemented across different contexts.

Our research questions demanded comparison within and between two arms of the study. The focus for comparison included experiences of students with low and high-prior attainment, the quality of a mathematics learning environment (including teaching quality and OTL), and contextual factors influencing grouping practices. To identify broader patterns, we collected comparable data across case studies.

Year 7 (May 2023 to July 2023)

Case study visit 1 included:

- An interview with the head of mathematics in each school to gain an overview of the school's approach to grouping students in mathematics, with a focus on equity in their provision.
- Observation of two mathematics lessons: two mixed attainment lessons in each of the six mixed attainment schools; or a low set lesson and a high set lesson in each of the setting schools.
- Post-observation interviews with the class teacher and four focal students from each observed lesson (two paired interviews, see below for further details).

The case study schools were invited to identify two classes to be observed. The schools were asked to select these classes to illustrate their grouping practice 'done well'. We adopted this selection criteria to provide a proxy for the schools' perceptions of effective classrooms for their respective grouping practices.

[13] The original plan was to conduct two teacher surveys, one at the end of the first year of the study and another at the end of the second year of the study. The reason for this was that we had intended to measure teacher quality through a teacher survey of pedagogical content knowledge. However, our attempts to develop this measure were unsuccessful and so we decided to use an alternative measure of teacher quality. This meant that the teacher survey in year one of the study was unnecessary.

As part of the lesson observations:

- Two researchers[14] made 'real-time' notes of events/activity in the lessons. Photos of the classroom board/screen were taken to augment the lesson notes. After the lesson, the researchers used their notes to collaboratively rate the lessons on a scale of 1–3 (using a 5-point scale: 1, 1.5, 2, 2.5, 3) for each of the TRU dimensions using a validated rubric (see Appendix C.8). Where 3 points represents the highest quality and 1.5 and 2.5 points represent elements of the respective higher level.

- Field notes were taken on focal students and then utilised as data sources and interview prompts. Schools were asked to identify four focal in each class. In mixed attainment schools, they were asked to identify two students working at a prior attainment level *above* and two students working at a prior attainment level *below* the majority of the class as determined by the schools' usual assessment data. In sets, schools were asked to identify students working at mid-range prior attainment compared with other students in that same set.[15]

- Lessons were audio-recorded using IRIS Connect software[16] to allow for the transcription of small episodes of interest as the need arose. Lesson narratives were written to summarise each lesson drawing on the two researchers' observation notes and the recordings.

- Paired interviews with the four focal students were conducted after each observed lesson to gain their perspectives on the lesson. In mixed attainment schools the students were paired according to prior attainment with one pair of students with low-prior attainment and one pair with high-prior attainment.

- Class teachers were interviewed after the observed lessons to gain their perspectives and reflections on the lesson in particular with respect to their students' experiences during the lesson.

- Artefacts such as schemes of work for Year 7 and 8 mathematics and lesson plans for each of the observed lessons were collected to use as prompts for teacher interviews and to examine differences in the curriculum offered, and OTL afforded, to different students.

Year 8 (November 2023 to March 2024)

The same schools were visited again in the second year of the study. As well as observations of mathematics lessons and post-lesson teacher and focal student interviews, Case study visit 2 included:

- Focus groups with Year 8 students focusing on typical experiences of mathematical content and cognitive demand.[17] A card-sorting task was used as a stimulus for discussion. The prompts can be found in Appendix C.11. Students selected for the focus groups were, as far as possible, different from the focal students observed in the lessons. For the purposes of focus groups, schools were asked to identify:
    - **Setting schools.** Four students in a low set and four students in a high set.
    - **Mixed attainment schools.** Four students with lower prior attainment from the same class and four students with higher prior attainment from the same class.

Each case study therefore, included:

- Head of mathematics interview;

---

[14] One researcher in each pair was a mathematics specialist (Hodgen, Jacques, Saunders). The mathematics specialist guided the decision as to whether mathematics content was at grade level.
[15] Mid-range students were identified because of the restrictions of setting. Focusing on mid-range students enabled us to exclude students at the extremes of the attainment range, with mid-range students being relatively high attaining in low sets and relatively low attaining in top sets.
[16] See: **www.irisconnect.com/uk/**
[17] Definitions according to the TRU framework referred to above.

- Head of mathematics survey;
- senior leader survey;
- two observations of Year 7 mathematics lessons;
- two observations of Year 8 mathematics lessons;
- four interviews with the observed mathematics teachers;
- two paired interviews with Year 7 focal students with low-prior attainment;
- two paired interviews with Year 7 focal students with high-prior attainment;
- two paired interviews with Year 8 focal students with low-prior attainment;
- two paired interviews with Year 8 focal students with high-prior attainment;
- one focus group with four Year 8 students with low-prior attainment; and
- one focus group with four Year 8 students with high-prior attainment.

An overview of all IPE methods with the total n for the achieved sample is presented in Table 7. Sample targets were achieved or close to achieved for all IPE data collection.

Table 7: IPE methods overview. Where multiple methods addressed the same research question, sources were combined as described in the data analysis section. Achieved samples are given in brackets

IPE dimension	Research question addressed	Research methods	Data collection methods	Achieved sample size	Data analysis methods
Programme differentiation (including compliance)	6	Survey	Online questionnaire	Head of mathematics (n=98) (30 MA + 68 set)	Descriptive statistics
		Interview	Semi-structured interviews	Head of mathematics (n=12)	Deductive coding Thematic analysis – inductive coding
		Interview	Semi-structured interviews	Year 7 teachers (low set) (n=6) Year 7 teachers (high set) (n=5) Year 7 mixed attainment teachers (n=12) Year 8 teachers (low set) (n=6) Year 8 teachers (high set) (n=6) Year 8 mixed attainment teachers (n=12)	Deductive coding Thematic analysis – inductive coding
Students' mathematical experiences	5, 7	Survey	Online questionnaire	Year 8 students (n=14,898) (4,725 MA + 10,223 set)	Descriptive statistics
		Focus groups	Semi-structured interview / sorting task	Lower prior attaining Year 8 students in low sets (n=23) Higher prior attaining Year 8 students in high sets (n=23) Lower prior attaining Year 8 students in mixed attainment (n=23) Higher prior attaining Year 8 students in mixed attainment (n=22)	Deductive coding Thematic analysis – inductive coding
		Paired interview	Semi-structured interviews	Lower prior attaining students in low sets (n=40) Higher prior attaining students in high sets (n=44) Lower prior attaining students in mixed attainment (n=48) Higher prior attaining students in mixed attainment (n=46)	Deductive coding Thematic analysis – inductive coding
		Observation	Structured lesson observations	Year 7 Mixed attainment lessons (n=12) Year 8 Mixed attainment lessons (n=12) Year 7 Low set lessons (n=6) Year 7 High set lessons (n=5) Year 8 Low set lessons (n=6) Year 8 High set lessons (n=6)	TRU analysis Narrative account
Responsiveness of pedagogy	5, 8	Focus groups	Semi-structured interview / sorting task	Lower prior attaining Year 8 students in low sets (n=23) Higher prior attaining Year 8 students in high sets (n=23) Lower prior attaining Year 8 students in mixed attainment (n=23)	Deductive coding Thematic analysis – inductive coding

IPE dimension	Research question addressed	Research methods	Data collection methods	Achieved sample size	Data analysis methods
				Higher prior attaining Year 8 students in mixed attainment (n=22)	
		Paired interview	Semi-structured interviews	Lower prior attaining students in low sets (n=40) Higher prior attaining students in high sets (n=44) Lower prior attaining students in mixed attainment (n=48) Higher prior attaining students in mixed attainment (n=46)	Deductive coding Thematic analysis – inductive coding
		Interview	Semi-structured interviews	Year 7 teachers (low set) (n=6) Year 7 teachers (high set) (n=5) Year 7 mixed attainment teachers (n=12) Year 8 teachers (low set) (n=6) Year 8 teachers (high set) (n=6) Year 8 mixed attainment teachers (n=12)	Deductive coding Thematic analysis – inductive coding
		Survey	Online questionnaire	Mathematics teachers (n=611) (196 MA + 415 set)	Descriptive statistics
Context	9	Survey	Online questionnaire	Senior leaders (n=79) (23 MA + 56 set)	Descriptive statistics
		Survey	Online questionnaire	Head of mathematics (n=98)	Descriptive statistics
		Interview	Semi-structured interviews	Head of mathematics (n=12)	Deductive coding Thematic analysis

Data analysis

The IPE data sources were analysed separately and then synthesised in the final interpretation stage, as appropriate for answering our IPE research questions. However, some aspects of the ongoing qualitative data analysis contributed to the quantitative side of the study during the design of the endline surveys with teachers and students. The process of data analysis is illustrated in Figure 8.

Analysis of survey data
Data from IPE surveys[18] were analysed using descriptive statistics including frequencies and cross-tabulation. Cross-tabulation included analysis based on prior attainment, grouping type, FSM status, gender, and ethnicity. Free-text answers were coded using thematic analysis with themes based on the research questions.

Analysis of case study data
It is usually desirable in cross-case analysis to combine within-case and between-case analytic strategies (Miles *et al.,* 2019). We started by writing up each case using a standard set of TRU variables, which underpin all data sources. Once each case study narrative was completed, we moved to the analysis of series of cases within each arm of the study (i.e. setting and mixed attainment cases) using pre-defined themes developed from the research questions and emerging cross-cutting subthemes derived from the data. The final systematic comparison between the cases in setting and mixed attainment groups was informed by our research questions. All data management, analysis, and write-ups were supported by computer-assisted qualitative data analysis software NVivo 12 (Lumivero LLC, Denver CO, USA).

Lessons and lesson narratives

<u>Coding based on the TRU framework</u>

The lesson narratives produced by mathematics specialists were firstly analysed using the deductive coding framework (column A) and for evidence of the 12 evidence-based strategies (Table 5, page 31). Inductive qualitative content analysis (Mayring, 2019) was then used to identify examples of classroom activity (content) that characterise qualities of the five domains building on the TRU instrument developer's analysis (Schoenfeld, 2018). Since classrooms are complex environments, some segments were simultaneously coded to more than one TRU dimension. Differential patterns in teaching practices and student experiences between mixed attainment classes, low sets, and high sets were explored using cross-tabulation analyses to create matrices: e.g. TRU dimension x grouping practice (MA + low set + high set) and TRU dimension x rating scored (1, 1.5, 2, 2.5, 3).

Structural coding of lessons was also used to indicate the occurrences of classroom activity involving whole class, group work, and individual work. This enabled further interrogation of differential classroom activity between mixed attainment groups, low sets, and high sets.

Post-lesson focal student and teacher interviews

<u>Deductive coding based on the TRU framework</u>

Segments of interview text were deductively coded according to the conceptual framework set out in (columns B and C) together with codes relating to the 12 evidence-based strategies set out in Table 5 (page 31). Since the interview schedules (see Appendix C.9 to C.12) were designed using the TRU framework, it was possible to code segments according to the interview schedule items. However, since the interviews were semi-structured, some responses were coded to more than one domain code. For instance, responses about mathematical content may also involve cognitive demand and were therefore, coded to both.

[18] The baseline student survey was not part of the IPE.

Pattern coding

Thematic analysis using Braun and Clarke's (2019) six-stage approach to reflexive thematic analysis involved further inductive coding to identify themes, which described student or teacher experiences and perspectives within and across the five TRU dimensions in the context of different grouping practices.

The post-lesson interview data was triangulated with the lesson narratives as part of the school case study analysis. The lesson plans collected during field visits were utilised to support this analysis.

Focus groups with students

As above, the analysis started with deductive coding according to the TRU framework for mathematical content and cognitive demand (see Appendix C.8). Thematic analysis using Braun and Clark's (2019) approach involved inductive coding to identify themes, which described students' collective experiences and perspectives on these two TRU dimensions to pattern experiences of the different grouping practices from the perspective of students with differing levels of prior attainment.

Heads of mathematics interviews

The deductive coding following the TRU conceptual constructs (see Appendix C.8) was followed by inductive coding to support both identification of patterns across cases and important local factors in case study schools. The analysis of this dataset therefore, integrated case-oriented (i.e. as contributing to understanding case dynamics) and variable-oriented analytic strategies (i.e. themes that cut across cases).

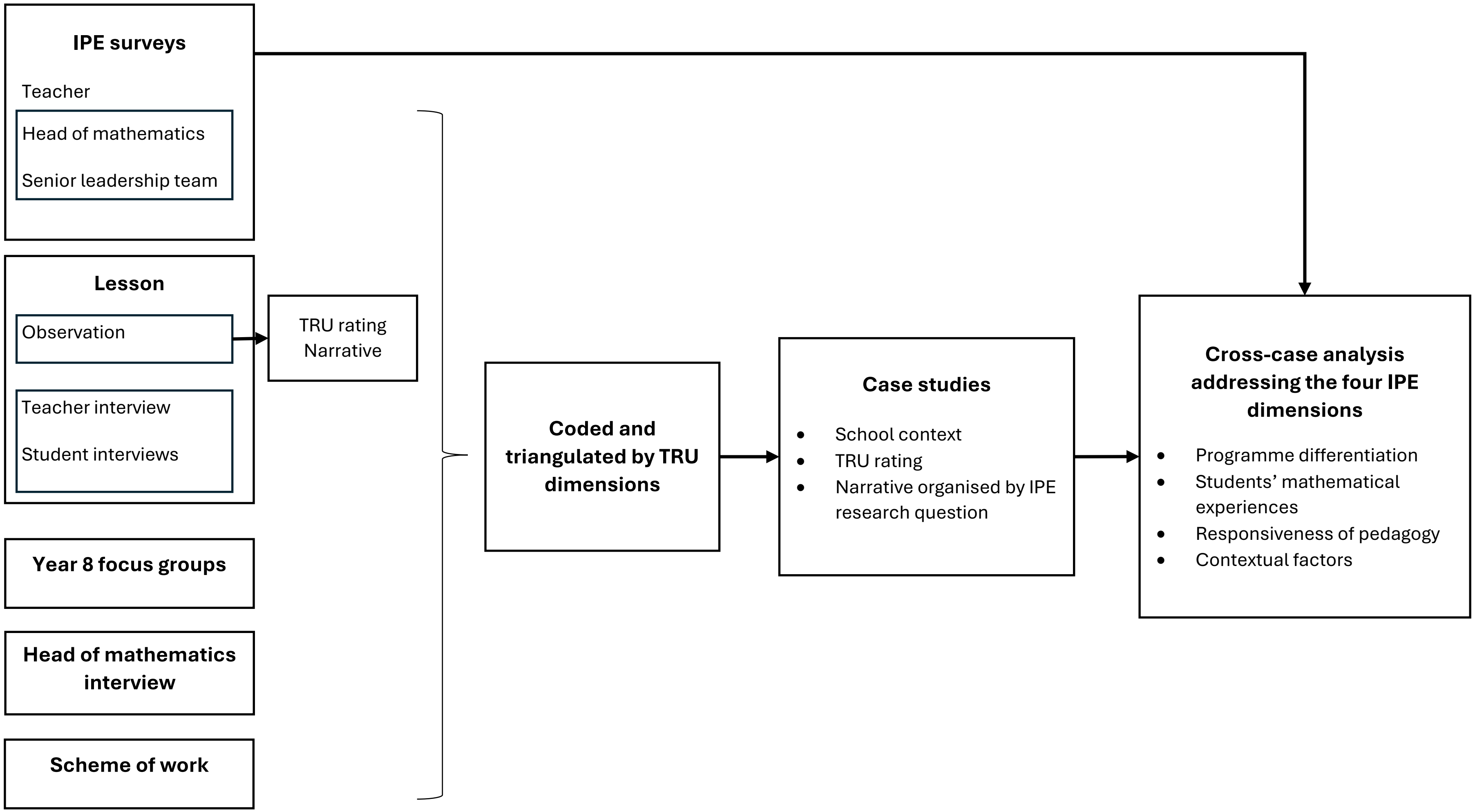
Mathematics schemes of work

The deductive analysis of these documents will focus on the dimensions of implementation and theoretical constructs such as OTL. The triangulation of this data with heads of mathematics interviews in each case study school will produce a picture of intended provision for different students in each school, while the subsequent analysis will systematically compare it within and across the two arms of the study.

Triangulation of data sources

First, we combined the multiple data sources by triangulating using our theoretical framework of the TRU (see Appendix C.24). Second, we compared and contrasted findings against each research question, looking for agreement and dissonance among sources (see Table 7). Full case studies are included in Appendix D.

Figure 8: Illustration of the process of within-case and cross-case analysis in the IPE



Costs

We were not evaluating an intervention in this study, so there were no specific costs associated with implementation. Nevertheless, we investigated the costs associated with maintaining different grouping practices. We anticipated that some schools in our sample may have changed their grouping practices within the preceding five years, however, we did not find examples of this.

We planned to estimate any costs using implementation evidence from interviews and the head of mathematics survey, but in the event no schools changed their practices so we could not estimate the costs of this.

Timeline

Table 8: Timeline. For team initials see Project team, page 13

Dates	Activity	Staff responsible / leading
November 2018 – December 2018	Ethical and data protection approval sought Data security Protocol written Recruitment manager and administrator appointed	BT/AT/JH/JA
January 2019	MoUs agreed	BT/AT
February 2019 – July 2019	Recruitment and matching Prepare student baseline survey Prepare school practices survey	MC/BT BT/AT BT/AT
August 2019 – September 2019	Statistical analysis plan written	JH/JA
September 2019 – October 2019	Parent/carer ethical consent sought via schools (withdrawal) UPNs collected from schools for NPD matching request Head of mathematics survey Baseline student self-confidence survey	Admin Admin BT/AT BT/Admin
September 2019 – August 2020	Survey instrument design and validation Qualitative instrument design and piloting (September 2019 – February 2020) Qualitative data collection (March 2020 – July 2020)	RK AT AT/LJ
Study interrupted due to COVID-19 pandemic and restarted in 2022	–	–
January 2022	Ethics and data protection revised	BT
February 2022 – July 2022	Recruitment and matching Update and prepare student baseline survey Update and prepare school practices survey	Admin/MC/JA BT/AT BT/AT
September 2022 – October 2022	Contacting schools about collecting data, processing opt-out forms Receive Year 7 student information from schools (student names, dates of birth, school identification, and UPN) Baseline student self-confidence survey	Admin Admin BT/Admin
September 2022 – July 2023	Qualitative instrument design and piloting (September 2022 – February 2023) Wave 1 qualitative data collection (March 2023 – July 2023)	AT AT/LJ/BT/JH/MC/PS

Dates	Activity	Staff responsible / leading
March 2023 – April 2023	Heads of mathematics online survey	BT
March 2023	Case study schools selected and recruited	AT/LJ
May 2023 – July 2023	*Case study school phase 1 visits Year 7*: Classroom observation and face-to-face interviews with students, teachers, and heads of mathematics, and documents (lesson materials, schemes of learning).	AT/LJ
November 2023 – March 2024	*Case study school phase 2 visits Year 8*: Classroom observation and face-to-face interviews with students and teachers; focus groups with students; and documents (lesson materials, schemes of learning).	AT/LJ
May 2024 – June 2024	Senior leader survey	BT
May 2024 – July 2024	Mathematics teacher survey	BT
May 2024 – July 2024	*Outcome measure data collection:* Student endline survey (outcome self-confidence, OTL, perceptions of mathematics teaching) PTM Teacher survey Senior leader survey	BT/Admin
July 2024 – January 2025	Outcome analysis Implementation analysis Report writing	JH/JA/NB/DS/HB/BT AT/BT/JH JH/BT

Impact evaluation results

Participant flow including losses and exclusions

The flow of participants is detailed in Figure 9. The recruitment and matching process is described in the 'Methods' section above. Initially 148 schools were approached for participation as mixed attainment schools. A total of 32 schools agreed to participate and were included in the matching process. However, ultimately two of these did not provide student data and so were excluded from the study and are not included in the participant flow diagram. A total of 800 setting schools were approached for participation in the study, including 50 potential matches for the two schools that did not supply student data. The participant flow diagram therefore, shows 750 matched schools approached for participation. A total of 80 matched setting schools initially agreed to participate. Of these, seven setting schools did not provide student data. This resulted in the exclusion of two further mixed attainment schools, which were matched only to setting schools that did not provide student data.

The final recruited numbers of schools that provided student data were 28 mixed attainment schools and 69 matched setting schools.

A further seven setting schools withdrew post-matching. This resulted in one more mixed attainment school being excluded from the final analysis as all its matched setting schools had withdrawn.

The final sample for analysis therefore, comprised 27 mixed attainment schools and 62 matched setting schools.

The results of the MDES calculations at planning, recruitment, and analysis stages are presented in Table 9. The MDES at analysis stage was lower than that at planning or recruitment stages, thus increasing the power of the trial. This was due to two factors: a much higher than anticipated correlation between Key Stage 2 mathematics scores and the outcome measure; and a much lower school-level ICC than anticipated.

Figure 9: Participant flow diagram. Counts of numbers of students lost to analysis include whole schools as well as students within retained schools who did not complete endline testing

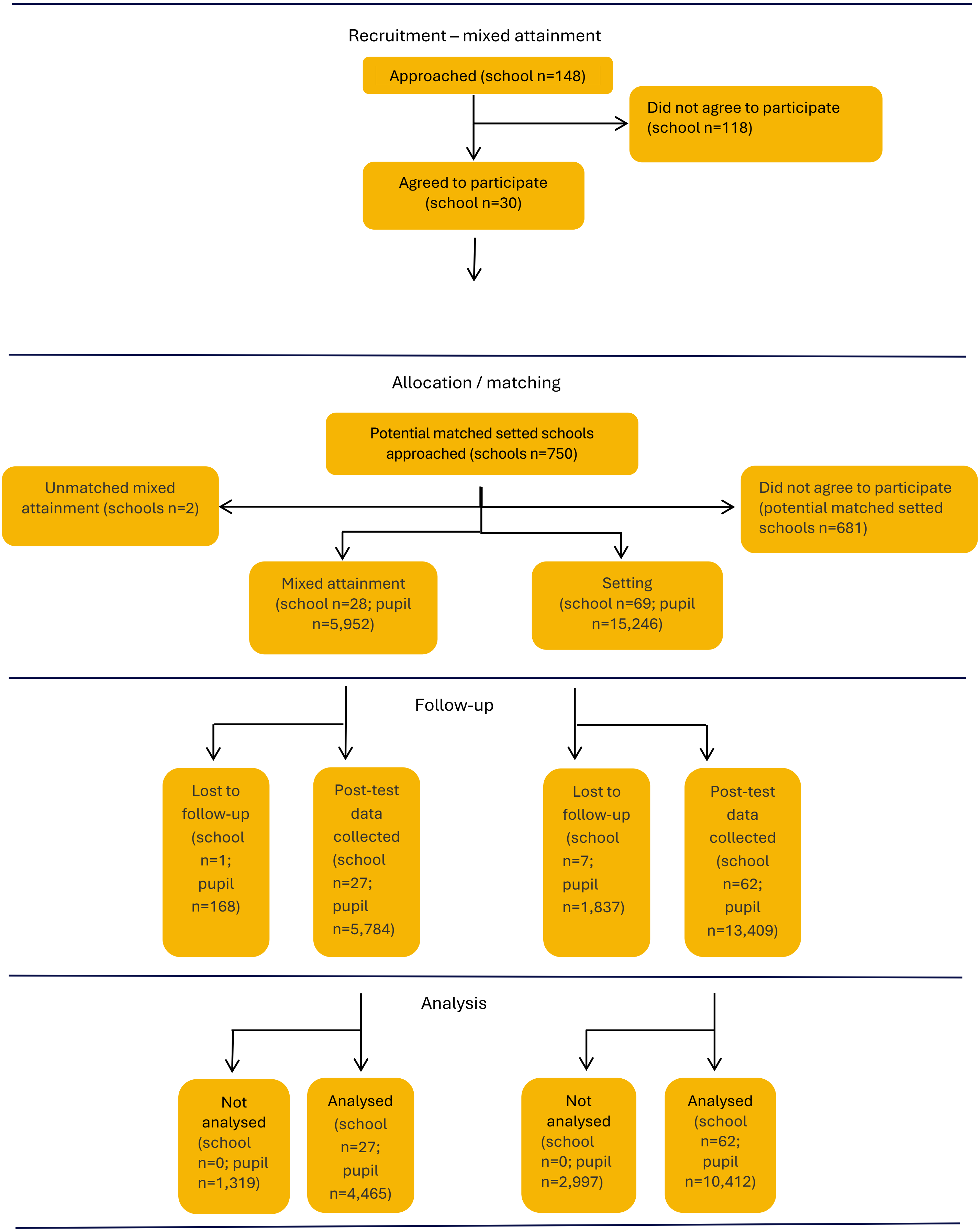


Table 9: MDES at different stages. LPA means low-prior attainment

		Study plan		Recruited matched sample		Analysis	
		Overall	FSM / LPA	Overall	FSM	Overall	FSM / LPA
MDES[a]		0.201	0.209	0.231	0.238	0.125	0.170
Pre-test/post-test correlations	Level 1 (student)	0.75	0.75	0.75	0.75	0.73	0.71
	Level 2 (school)	0.38	0.38	0.38	0.38	0.75	0.76
ICCs	Level 2 (school)	0.15	0.15	0.15	0.15	0.07[b]	0.06
Alpha		0.05	0.05	0.05	0.05	0.05	0.05
Power		0.8	0.8	0.8	0.8	0.8	0.8
One-sided or two-sided?		Two-sided	Two-sided	Two-sided	Two-sided	Two-sided	Two-sided
Average cluster size (students in schools)		100	25	190	40	167	32
Average cluster size (matched groups of schools)		6	6	3.5	3.5	3.3	3.3
No. of schools	Mixed attainment	40	40	28	28	27	27
	Setting	80	80	69	69	62	62
	Total:	120	120	97	97	89	89
No. of students	Mixed attainment	4,000	1,000	5,320	1,120	4,465	921
	Setting	8,000	2,000	13,110	2,760	10,412	1,962
	Total:	12,000	3,000	18,430	3,880	14,877	2,883

[a] *The MDES estimates at study plan and recruitment stages differ slightly to those in the Statistical Analysis Plan (Anders et al., 2024) due to a correction to the original code. The proportions of mixed to setted schools were corrected from 0.25 to 0.33 and 0.29 for the study plan and recruited sample calculations respectively. These corrections had a negligible effect on the MDES for the study plan and resulted in a very small increase in power for the recruited sample.*
[b] *The MDES estimates at analysis stage are based on the actual ICCs, which are smaller than our initial assumption. Hence, despite a smaller sample size, the power has improved. See text for further details.*

Attrition

Attrition rates are shown in Table 10. Overall attrition is 29.8%. Differential attrition between the two groups is 6.7 percentage points, which is within the United States What Work Clearinghouse's (2022) optimistic threshold for differential attrition.

Table 10: Student-level attrition from the evaluation (primary outcome)

		Mixed attainment	Setting	Total
No. of students	At matching	5,952	15,246	21,431
	Analysed	4,465	10,412	14,877
Student attrition (from allocation to analysis)	Number	1,487	4,834	6,554
	Percentage	25.0%	31.7%	29.8%
No. of schools	At matching	28	69	97
	Analysed	27	62	89
School attrition (from allocation to analysis)	Number	1	7	8
	Percentage[a]	3%	10%	8%

[a] Percentages are provided for school attrition to enable a comparison across the two groups and between student and school attrition. We note that the number of schools is less than 100 and, hence, these percentage figures for school attrition should be interpreted with caution.

Student and school characteristics

Table 11 presents the baseline characteristics of the mixed attainment and setted groups of schools when weighted for the number of matched comparator schools and the number of students in each school at analysis stage.

The balance at student level for Key Stage 2 mathematics is good and within acceptable levels.[19] Similarly, the balances at student level for the proportions of low and high-attaining students, FSM, and EAL are within acceptable levels. There are some indications of slight imbalances for general self-confidence levels and IDACI score, indicating that, on average, the students in the mixed attainment group had slightly lower self-confidence and were slightly less deprived than the students in the setted group. Balances between groups on a fixed set of variables for means, SDs, and skewness at both school and student levels, and for the subgroup of students eligible for FSM were tested and found to be acceptable. See Appendix C.19 for further details.

The aim of the matching exercise was to primarily achieve good balance at student level. There was some evidence of imbalance at school level. In short, the schools in the mixed attainment group were larger and fewer of them were academies or rated as 'Outstanding' by Ofsted, which may be related to teacher quality. The mixed attainment group also had slightly higher proportions of FSM and EAL students. These differences were broadly similar to those anticipated by the matching process. See Appendix C.19 for a comparison to the matching process.

[19] The threshold for low risk of bias due to imbalance are a standardised difference of ≤0.05 (EEF, 2019). A standardised difference of between 0.05 and 0.1 is classified as constituting a moderate risk of bias.

Table 11: Baseline characteristics of groups weighted for number of matched comparator schools and the number of students in each school

School-level characteristics	Mixed attainment group		Setted group		
	n/N (missing)	Mean / proportion	n/N (missing)	Mean / proportion	
2019 School average Key Stage 2 attainment of intake	27 (0)	29.1	62 (0)	29.2	
2018 School average Key Stage 2 attainment of intake	27 (0)	28.7	62 (0)	28.8	
2017 School average Key Stage 2 attainment of intake	27 (0)	28.6	62 (0)	28.9	
School low-prior attainment proportion	27 (0)	8.2%	62 (0)	8.3%	
School high-prior attainment proportion	27 (0)	45.0%	62 (0)	47.1%	
Number of students in Year 7 reported by school	27 (0)	238	62 (0)	229	
School FSM proportion	27 (0)	22.2%	62 (0)	19.4%	
School EAL proportion	27 (0)	20.2%	62 (0)	14.2%	
School academy status	27 (0)	68%	62 (0)	77%	
IDACI Quintile 1	27 (0)	27%	62 (0)	27%	
IDACI Quintile 2		17%		22%	
IDACI Quintile 3		30%		25%	
IDACI Quintile 4		18%		16%	
IDACI Quintile 5		7%		11%	
Ofsted Outstanding	27 (0)	30%	62 (0)	39%	
Ofsted Good	27 (0)	56%	62 (0)	59%	
Ofsted Requires Improvement	27 (0)	15%	62 (0)	3%	
Urban	27 (0)	81%	62 (0)	90%	
Student-level characteristics	**n/N (missing)**	**Mean / proportion**	**n/N (missing)**	**Mean / proportion**	**Standardised difference**
Proportion of students in lower tertile of Key Stage 2 mathematics scores	4,465 (0)	29%	10,412 (0)	29%	–
Proportion of students in upper tertile of Key Stage 2 mathematics scores	4,465 (0)	35%	10,412 (0)	36%	–
FSM6 status	4,465 (0)	21%	10,412 (0)	19%	–
EAL status	4,465 (0)	17%	10,412 (0)	15%	–

Key Stage 2 mathematics score	4,465 (0)	77.0	10,412 (0)	77.2	-0.01
General self-confidence[a20]	3,414 (1,051)	15.49	7,906 (2,506)	15.16	0.06
Mathematics self-confidence[b21]	2,669 (1,796)	17.16	7,392 (3,020)	17.17	0.00
FSM6 status	4,465 (0)	21%	10,412 (0)	19%	0.05
EAL status	4,465 (0)	17%	10,412 (0)	15%	0.05
Individual IDACI score	4,465 (0)	0.14	10,412 (0)	0.15	-0.07

[a] *Note: A higher general self-confidence score means a lower level of self-confidence.*
[b] *Note: A higher mathematics self-confidence score means a lower level of self-confidence.*

Outcomes and analysis

Primary analysis

Table 12 presents the results for all students for the primary outcome measure (mathematics attainment as measured by the GL PTM13 test). It shows the unadjusted means for the mixed attainment and setted groups after weighting was applied to account for the number of matched comparator setted schools for each mixed attainment school and for the number of students in each school. Using the primary analysis model outlined in the 'Methods' section, including various student- and school-level covariates alongside the baseline measure (Key Stage 2 mathematics scores), we calculated a Hedges' g effect size of -0.05 (95% CI: -0.14, 0.03) for the impact on the mathematics scores for all students when taught in mixed attainment groups as compared to being taught in setted groups. This small negative effect is equivalent to one month's progress and is not significantly different to zero. This provides little evidence of a difference in the primary outcome between the two groups on the mathematics scores of all students in the cohort (research question 2a).

Table 12 also presents the results for low-, middle-, and high-prior attaining students. For low-prior attaining students, using the distributional change analysis model outlined in the methods section, we calculated a Hedges' g effect size of 0.04 (95% CI: -0.05, 0.13) for the impact on their mathematics scores when taught in mixed attainment groups as compared to being taught in setted groups. This small positive effect is equivalent to zero month's progress and is not significantly different to zero. This addresses research question 1 and provides little evidence of a difference in the primary outcome between the two groups for low-prior attaining students. In contrast, the impact of being taught in a mixed attainment group on the mathematics scores of the high-prior attaining groups was negative and significantly different to zero. We calculated a Hedges' g effect size of -0.14 (95% CI: -0.25, -0.03), which is equivalent to two months' progress. The effect on the distributional gap between the low- and high-prior attaining students was a Hedges' g effect size of -0.18 (95% CI: -0.26, -0.10) for the mixed attainment schools in comparison to the setted schools. In other words, the 'gap' between the mathematics scores of the low- and high-prior attainers was smaller in the mixed attainment schools. However, this was largely due to significantly different lower scores for the high-prior attaining students rather than higher scores for the low-prior attaining students.

Finally, Table 12 presents the results for the FSM subgroup of students and we calculated a Hedges' g effect size of 0.01 (95% CI: -0.08, 0.10) for the impact on their mathematics scores when taught in mixed attainment groups as compared to being taught in setted groups. This negligible effect is equivalent to zero month's progress and is not significantly different to zero. This provides no evidence of a difference in the primary outcome between the two groups for the FSM subgroup of students (research question 2b).

Table 12: Primary analysis

Outcome	Unadjusted means (weighted sample)				Effect size		
	Mixed attainment group		Setted group				
	n (not matched)	Mean (95% CI)	n (not matched)	Mean (95% CI)	Total n (intervention; control)	Hedges' g (95% CI)	P-value
PTM13 (all students)	4,465 (2,168)	102.3 (102.0, 102.70)	10,412 (5,563)	102.5 (102.2, 102.8)	14,877	-0.051 (-0.135, 0.033)	.232
PTM13 (FSM)	921	97.4 (96.6, 98.2)	1,962	95.8 (95.2, 96.3)	2,883	0.008 (-0.083, 0.099)	.860
Distributional change model							
PTM13 (low attainers)	1,302	91.2 (90.7, 91.7)	3,076	90.0 (89.7, 90.3)	4,378	0.044 (-0.045, 0.133)	.330
PTM13 (high attainers)	1,565	112.6 (112.1, 113.1)	3,574	113.9 (113.5, 114.2)	5,139	-0.137 (-0.248, -0.025)	.017
PTM (Middle attainers): Reference group	1,598	101.4 (101.0, 101.8)	3,762	101.3 (101.1, 101.6)	5,360	-0.043 (-0.132, 0.045)	.334
PTM13 (low- / high-attainment gap)	–	–	–	–	–	-0.181 (-0.264, -0.097)	.001

Secondary analysis

The results of the secondary outcome (self-confidence) analyses are presented in Table 13. These use models analogous to those used in the primary analyses. We hypothesised that mixed attainment teaching would have a positive effect on self-confidence for the low-attaining students. However, these results show a moderate, statistically significant, negative effect on general self-confidence for low-prior attaining students (*g* = -0.13, 95% CI: -0.23, -0.04) (research question 3) with a smaller, and non-statistically significant, negative effect for all students (*g* = -0.07, 95% CI: -0.13, 0.00) (research question 4a). The effect on the general self-confidence of the FSM subgroup of students was similar to the effect for all students, but was not significantly different to zero (*g* = -0.08, 95% CI: -0.19, 0.04) (research question 4b). The results for mathematics self-confidence are very similar, with a moderate, statistically significant, negative effect on mathematics self-confidence for low-prior attaining students (*g* = -0.19, 95% CI: -0.27, -0.10) with a smaller, statistically insignificant, negative effect for all students (*g* = -0.06, 95% CI: -0.13, 0.00). The effect on the mathematics self-confidence of the FSM subgroup of students was similar to the effect for all students, but was not significantly different to zero (*g* = -0.11, 95% CI: -0.24, 0.03).

Table 13: Secondary analysis

Outcome	Unadjusted means (weighted sample)				Effect size		
	Mixed attainment group		Setted group				
	n (not matched)	Mean (95% CI)	n (not matched)	Mean (95% CI)	Total n (intervention; control)	Hedges' g (95% CI)	P-value

General self-confidence (all students)[a]	3,236	16.3 (16.1, 16.5)	7033	16.0 (15.8, 16.1)	10,029	-0.066 (-0.134, 0.003)	.059
General self-confidence (low attainers): Distributional change model with middle attainers as the reference group	880	19.3 (19.0, 19.7)	1815	18.7 (18.5, 19.0)	2,695	-0.130 (-0.226, -0.035)	.007
General self-confidence (FSM)	643	17.4 (16.9, 17.9)	1221	17.4 (17.1, 17.8)	1,864	-0.075 (-0.191, 0.041)	.205
Mathematics self-confidence (all students) Distributional change model with middle attainers as the reference group[b22]	2508	18.6 (18.3, 18.9)	6568	18.5 (18.4, 18.7)	9,076	-0.064 (-0.129, 0.001)	.055
Mathematics self-confidence (low attainers)	629	23.5 (23.0, 24.0)	1671	22.1 (21.8, 22.4)	2,300	-0.185 (-0.272, -0.098)	.000
Mathematics self-confidence (FSM)	469	20.2 (19.6, 20.9)	1132	19.9 (19.5, 20.3)	1,601	-0.106 (-0.242, 0.029)	.122

[a] General self-confidence measure is negatively coded; a higher score meant lower self-confidence. This is corrected (reversed) in the effect sizes for ease of interpretation.
[b] Mathematics self-confidence measure is negatively coded; a higher score meant lower self-confidence. This is corrected (reversed) in the effect sizes for ease of interpretation.

Missing data analysis

As outlined in the methods and pre-specified in the Statistical Analysis Plan (Anders *et al.,* 2024), we proposed to implement our missing data strategy, if more than 10% of data in the primary modelling is missing. This threshold was met (see section on 'Attrition'). In order to assess whether it is plausible to consider these data as MAR, we ran a logistic regression model to predict missingness in outcome data using variables in the imbalance testing. The results of this analysis indicated that missingness was related to observed variables, supporting the plausibility of the MAR assumption. We then re-estimated the treatment effect for cases for which only the outcome variable was missing using the pre-specified model. This analysis indicated no statistically significant effect for the intervention as in the main analysis (see Appendix C.18).

Analysis in the presence of non-compliance

The head of mathematics survey indicated that no school changed their grouping practices during the trial. Since there was no non-compliance, this analysis was not conducted.

Additional analyses

As outlined in the 'Methods' section, in order to address research question 5, exploratory analyses were planned to investigate the role of OTL and teacher quality in mediating the overall changes in the primary outcome. Since the impact on the primary outcome was not statistically significant either for all students or for low-prior attaining students, we did not expect to find statistically significant results from these mediation analyses. Nevertheless, these exploratory analyses were conducted and, as expected, the results were not statistically significant. However, somewhat counterintuitively, these analyses suggested that OTL and teacher quality were associated with low test scores after controlling for mixed/setting treatment (Table 14). See Appendix C.17 for a more detailed summary of these results.

Table 14: Mediation analyses effect sizes

	Hedges' g (95% CI)	P-value
Full sample analysis: All students (N=11,248)		
OTL model: Total effect	-0.019 (-0.106, 0.068)	0.669
OTL model: Indirect effect	-0.006 (-0.013, 0.001)	0.097
Teacher quality model: Total effect	-0.021 (-0.112, 0.069)	0.648
Teacher quality model: Indirect effect	-0.008 (-0.020, 0.004)	0.197
Analysis for low-attaining students (N=3,127)		
OTL model: Total effect	0.124 (-0.004, 0.253)	0.057
OTL model: Indirect effect	-0.002 (-0.008, 0.005)	0.655
Teacher quality model: Total effect	0.117 (-0.010, 0.245)	0.071
Teacher quality model: Indirect effect	-0.009 (-0.021, 0.003)	0.155

As a robustness check on the analysis of distributional change, we conducted a quantile regression analysis. This was specified at the decile points of the outcome distribution in addition to the 25th, 50th, and 75th percentiles as pre-specified in the Statistical Analysis Plan (Anders *et al.,* 2024), due to some anomalous results at the 25th percentile.[23] This indicated a similar pattern to the main distributional analysis (Table 15). A summary of the results of this analysis is presented in Appendix C.16.

Table 15: Quantile analysis effect sizes

Percentile	Hedges' g (95% CI)
10th	-0.051 (-0.174, 0.073)
20th	-0.028 (-0.126, 0.071)
25th	-0.039 (-0.042, -0.035)
30th	-0.029 (-0.123, 0.066)
40th	-0.030 (-0.112, 0.051)
50th	-0.031 (-0.108, 0.046)
60th	-0.036 (-0.108, 0.036)
70th	-0.039 (-0.116, 0.038)
75th	-0.039 (-0.118, 0.040)
80th	-0.042 (-0.129, 0.045)

[23] The result at the 25th percentile was unexpectedly large. However, the decile results, particularly those at the 20th and 30th percentiles, indicated this as an anomaly. See Table C.16.9 in Appendix C.16.

90th	-0.049 (-0.127, 0.029)

As a sensitivity check, as described in the 'Methods' section, we conducted a replication of the primary and distributional change analyses using a sample matched at student level. These analyses indicated similar results to the main analyses. A summary of these results is presented in Appendix C.16.

A sensitivity analysis found no evidence to indicate that school self-administration of final testing had any influence on the outcomes. See Appendix C.16 for a summary of these results.

As outlined in the methods, we had planned to conduct an additional sensitivity analysis using the approach described by Rosenbaum (2015). However, this method is designed to assess the sensitivity of statistically significant results and, hence, was judged not to be informative in this case.

Estimation of ICC

The ICC at school level for the primary and secondary outcomes are presented in Table 16. These are lower than the estimates at planning and analysis stage, resulting in increased power to detect effects.

Table 16: ICC estimates

Outcome	ICC
Primary analysis (PTM13)	0.074
Secondary analysis (General self-confidence)	0.033
Secondary analysis (Mathematics self-confidence)	0.029

IPE results

Compliance

Schools were asked to confirm their grouping prior to the endline survey. No schools notified us of a change in attainment grouping practices between recruitment and the end of the trial, so compliance is taken to be 100%. Variation in implementation of setting and mixed attainment grouping is explored further in programme differentiation, below.

Implementation dimensions

As outlined in the 'Methods' section above, this study was a natural experiment and sought to understand the different ways in which schools implement either 'mixed attainment grouping' or 'setting'. Hence, a focus on fidelity of implementation was not appropriate. We therefore draw on the surveys and case studies (see Appendix D for the full 12 case study narratives) to report on the following four dimensions of implementation:

5. Programme differentiation.
6. Students' mathematical experiences.
7. Responsiveness of pedagogy.
8. Contextual factors.

Programme differentiation

Organisation of mathematics classes
The number of teaching groups in schools is summarised in Table 17, and depended on the size of the student cohort. The range of numbers of teaching groups was similar in the mixed attainment and setting schools. The number of teaching groups ranged from 3 to 15 in mixed attainment schools and 4 to 13 in setting schools.

Table 17: Head of mathematics survey: How many mathematics classes are there in Year 7?

	Mixed (n=30)	Setting (n=68)
3	1	0
4	2	1
5	2	2
6	3	9
7	3	10
8	4	18
9	5	7
10	3	10
11	4	4
12	1	6
13	0	1
14	0	0
15	2	0

In the head of mathematics survey (98 responses: mixed attainment n=30; setting n=68), the most frequently mentioned reasons for choosing a mixed attainment grouping approach, given in response to open questions, were a focus on equity (n=9), and the belief that 'everyone can be a mathematician' (n=9). Slightly fewer heads of mathematics reported that they felt it was important for students to work with peers of different levels of ability (n=6), and that mixed attainment grouping helped avoid negative labelling of low-attaining students (n=5) and led to better learning (n=5).

The heads of mathematics from setting schools reported most frequently in the survey that they believed setting led to better learning (n=23). Other important reasons for choosing setting were that it helped support low-attaining students (n=18), stretch the most able (n=13), and that they felt that it was important for students to work with peers of a similar 'ability' level (n=10).

A number of interesting contrasts between setting and mixed attainment schools emerged. A few heads of mathematics in setting schools also claimed that setting helped remove negative labelling of low-attaining students (n=4). A few heads of mathematics in setting schools reported that setting led to ease of teaching (n=6), while no head of mathematics in mixed attainment schools gave this reason. Conversely only one head of mathematics in a setting school gave equity (n=1) or the belief that everyone can be a mathematician (n=0) as a reason to choose setting; and no head of mathematics in mixed attainment schools responded that their approach facilitated stretching the most able or supporting low-attaining students. A few heads of mathematics in setting schools said that setting was appropriate given the training, experience, or skills of their teachers (n=6), for example, one reported that: *'The majority of teachers have also taught in this way throughout their career and so a shift to mixed ability would be particularly challenging to manage'* (Head of mathematics).

Twelve heads of mathematics from the 30 mixed attainment schools reported in the head of mathematics survey that they had a 'nurture group' alongside mixed attainment groups. All but one gave the reason for this as being to improve curriculum access for low attainers, for example, saying that nurture group students are *'unable to access the curriculum due to their inability to read, write or use numbers at the appropriate level'* (Head of mathematics). Further reasons given included that it was a benefit for some students to be taught in smaller groups (n=5); or to support students with additional socio-emotional or learning needs (n=4).

In setting schools, the number of set levels ranged from two to nine, with the majority of schools having three or four set levels. Numbers of set levels can be found in Table 18.

Table 18: Head of mathematics survey: How many different ability or attainment level sets are there?[a][24]

Set levels	N
2	2
3	19
4	22
5	11
6	6
7	2
8	5
9	1

[a] Heads of mathematics were given the following supporting information to help them with answering the question:
This may not be the same as the total number of sets. The following examples may help you choose your answer:
If there is 1 top set, 2 middle sets and 1 bottom set you would answer 3.
If there are 2 top sets, 2 upper middle sets, 2 lower middle sets and 2 bottom sets, you would answer 4.
If there are 8 sets all at different ability levels, you would answer 8.

When asked to provide additional information about their school's approach to grouping, a small number of heads of mathematics in mixed attainment schools reported that their school started setting either in Year 9 (n=1) or Year 10 (n=2). One head of mathematics mentioned that they continue with mixed attainment throughout the school, to the end of Year 11. A small number indicated that they were either new to mixed attainment grouping (n=1) or that it was an established practice of at least six years (n=2). GCSE tier of entry was mentioned by three heads of mathematics, with two of these saying that Key Stage 4 grouping was by GCSE tier. Two heads of mathematics spontaneously expressed negative views about mixed attainment grouping with one saying that the difference between higher and lower attainers was 'vast' by Year 9, when setting was initiated in their school. The other reported negative effects of mixed attainment grouping: *'the best students are held back and the confidence of the less able students is shattered as they witness others repeatedly having more success than them'* (Head of mathematics).

Heads of mathematics in setting schools generally provided additional information about the structure of their groupings. However, a small number mentioned that the curriculum was differentiated for different groups (n=5); commented on positive behaviour (n=1) or negative behaviour (n=2) in sets; or on the provision for low-attaining students; or those with additional needs (n=3). Other issues raised by one head of mathematics in each case, were how students are allocated to sets, parental preferences for a top set, regular review of grouping practices; and movement of students between sets.

In the head of mathematics interviews, several heads of mathematics in mixed attainment schools referred to having a nurture group or one (smaller) classroom having more students with Education Health and Care Plans (EHCPs) to facilitate interventions and support. Special Educational Needs and Disabilities (SEND) teams were involved in constructing groups and designing support within mixed attainment structures. SEND students were discussed by one head of mathematics in a setting school (S4). These children were generally working at the lower end of setting and were provided additional support by the SEND team. The head of mathematics from setting school S6 described sitting SEND students in the aisle so that teachers could provide support.

Allocation of students to classes

Heads of mathematics were asked in the survey to report on how they allocate students to classes in Year 7 (Table 19). Key Stage 2 SATs test scores were used in the great majority of setting schools (n=60) and in about half of mixed attainment schools (n=14). Just under two-thirds of setting schools reported using their own test of student attainment (n=43), while this was used by very few mixed attainment schools (n=6). Similarly, commercially available tests were used more frequently in setting schools (n=24) than in mixed attainment schools (n=6). Much more common in mixed attainment schools was to use information from students' primary schools (n=15, compared with n=10 in setting schools). Random allocation was used in two-fifths of mixed attainment schools (n=12) and in no setting schools. Teacher judgements of student abilities were more commonly used in setting schools (n=20) than in mixed attainment schools (n=4). Other sources of information were used less frequently in both types of schools. Where additional tests were used, they were typically conducted in the first half-term of the school year (n=24 for commercial tests, n=44 for the school's own tests), however, a very small number of schools conducted tests before students had even started in the school, for example, on a 'moving up' day (n=6 for commercial tests, and n=2 for the school's own tests).

Table 19: Head of mathematics survey: Which sources of information were used to allocate students to Year 7 mathematics groups this year? Select all that apply.

	Mixed (n=30)	Setting (n=68)
National Curriculum Key Stage 2 test results	14	60
National Curriculum Key Stage 2 teacher assessments	2	9
A commercially available test (e.g. CATs, MidYIS)	6	24
The school's own test of student attainment	6	43
Teacher judgements of students' abilities	4	20
Teacher observations of student behaviour	1	3
A parent's judgement of their child's ability	0	1
Information from students' feeder schools	15	10
The results of random allocation of students	12	0
Taught in form groups	4	0

Heads of mathematics were invited to provide more information about why they used their specific approaches to allocating students to groups. The most popular response among heads of mathematics in mixed attainment schools was that they used their chosen sources to achieve balanced groups in terms of student prior attainment and other characteristics (e.g. gender, ethnicity, FSM) (n=12). A small number of heads of mathematics in mixed attainment schools mentioned either identification of students with additional needs (n=3); managing friendship groups within classes (n=2); or that it was school policy to use the available data (n=2). By contrast, the majority of heads of mathematics in setting schools reported that they used the data, which they felt gave the most accurate or reliable measure of student attainment in mathematics (n=39). Many of these specifically called into question the reliability of Key Stage 2 SATs tests or suggested that their own tests provided more useful data because of the types of questions, control over test conditions, or timing. Where heads of mathematics were happy to rely on Key Stage 2 SATs, they favoured these because they were readily available (n=14). Much smaller numbers of heads of mathematics mentioned achieving a balance of students in groups (n=2) or linking grouping to GCSE predictions or flight paths (n=2).

Allocation of teachers to classes

Heads of mathematics were also asked about how they allocated teachers to mathematics groups (Table 20, Table 21). This question was asked separately for mixed attainment and setting schools, with different options. In mixed attainment schools, heads of mathematics were most likely to report that teachers were allocated to classes randomly (n=17, approximately three-fifths), or that allocation to exam classes took precedence and Year 7 allocation happened later (n=11, approximately one-third). Smaller numbers reported that they had little choice in teacher allocation (n=2), teachers could express their own preferences (n=2), or teachers were allocated based on their strengths (n=4).

Within setting schools the most frequently reported mechanisms were to achieve balance for teachers, either by rotating them between ability classes and year groups from year to year (n=43) or to provide each teacher with a balance of set levels and ages (n=36). Allocation of teachers to Year 7 classes was also a lower priority in setting schools, with a substantial number of heads of mathematics saying that exam classes were staffed first (n=28). Other responses indicated that lower-attaining sets were prioritised for subject expert (n=13) or experienced (n=9) teachers; or that teacher strengths were taken into account (n=18). Subject expert teachers were prioritised for high-attaining classes in only a small number of schools (n=6).

Table 20: Head of mathematics survey: How did you allocate Year 7 teachers to mathematics classes for this school year 2022/2023? (Mixed attainment schools; n=30)

	N
Teachers are randomly allocated to different classes	17
I have little choice over how teachers are allocated to Year 7 classes	2
Teachers are allocated to classes based on their personal preferences	2
Teachers are allocated to classes based on their particular strengths	4
Teachers are allocated first to GCSE/A level classes and later to Year 7	11

Table 21: Head of mathematics survey: How did you allocate Year 7 teachers to mathematics classes for this school year 2022/2023? (Setting schools; n=68)

	N
Subject expert teachers tend to be placed with the higher ability classes	6
Subject expert teachers tend to be placed with the lower ability classes	13
Teachers are rotated between ability classes and year groups from year to year	43
Teachers are randomly allocated to different classes	4
I have little choice over how teachers are allocated to Year 7 classes	2
Teachers are allocated to classes based on their personal preferences	0
Teachers are allocated to classes based on their particular strengths	18
Experienced teachers tend to be placed with the higher ability classes	2
Experienced teachers tend to be placed with the lower ability classes	9
Experienced teachers tend to be placed with the middle classes	3
Inexperienced teachers tend to be placed with the higher ability classes	2
Inexperienced teachers tend to be placed with the lower ability classes	1
Inexperienced teachers tend to be placed with the middle classes	1
Teachers are allocated first to GCSE/A level classes and later to Year 7	28
Teachers are allocated so that they have a balanced set of classes (and ages)	36

As noted above, teacher strengths were taken into account in some schools when deciding which class to allocate them to. Heads of mathematics who indicated this was a factor were asked which teacher strengths were considered. Responses are summarised in Table 22 and find that in the mixed attainment schools strengths were considered in relation to which teachers might be suited to teaching a nurture group. In setting schools, the majority reported taking into account strengths in managing challenging behaviour (n=16), teaching students with SEND (n=15), subject knowledge (n=14), or teaching students with high attainment (n=13). Fewer heads of mathematics mentioned building relationships with vulnerable students (n=8 in setting schools).

Table 22: Head of mathematics survey: Which teacher strengths are taken into consideration when allocating teachers to groups? Answered by four heads of mathematics in mixed attainment schools and 18 heads of mathematics in setting schools

	Mixed (n=4)	Setting (n=18)
Teaching students with SEND	3	15
Teaching students with high attainment	1	13
Managing challenging behaviour	2	16
Building relationships with vulnerable students	2	8

Subject knowledge	2	14
Other strengths	1 – knowledge of our nurture curriculum is the biggest factor	0

Heads of mathematics were also invited to give additional details of why they allocated teachers in particular ways. A few heads of mathematics in setting schools noted that ensuring that teachers were allocated a range of different classes was beneficial for teacher development as it enabled teachers to develop their skills in teaching different ages and levels, and beneficial for teacher well-being because it ensured fairness in workload because high sets were larger than low sets and so amounts of marking varied. A few heads of mathematics in setting schools also commented on the presence of specialist and non-specialist teachers in their departments. Where heads of mathematics identified that all their teachers were skilled specialists there was more focus on balance than where there were some non-specialists, with two heads of mathematics suggesting that non-specialists were more comfortable with middle sets, and a larger number reporting that non-specialists were never allocated to A level, and preferably not to GCSE classes. A few heads of mathematics in mixed attainment schools mentioned that they took familiarity with the nurture curriculum or working with students with SEND into consideration when allocating teachers to groups.

Timetable constraints were a concern for heads of mathematics in both mixed attainment and setting schools. Typically teachers were allocated to A level and GCSE classes first, leaving limited options for allocation of teachers to Key Stage 3 classes. This could result in Key Stage 3 being more likely to have split classes (classes taught by more than one teacher) or being taught by teachers with a poor track record.

Students' mathematical experiences

Evidence about students' mathematical experiences was collected in various ways. In the student endline survey (n=14,898), we asked about student experiences of learning mathematics, focusing on the elements summarised in Table 23.[25] In case study visits, we used the TRU framework to structure lesson observations and interviews with students (see Appendix D for the full case study narratives).

The survey responses from students in mixed attainment schools and in setting schools were mostly reasonably similar. Slightly more students in mixed attainment schools agreed or strongly agreed that their teacher asked students to come up to the board to show their ideas (57% in mixed attainment schools vs 51% in setting schools). This accords with our lesson observations where we observed positive, and less positive, student experiences in both mixed attainment and setted classrooms. In some lessons, students at all attainment levels were encouraged, and supported, to explain their reasoning and to engage in productive struggle (e.g. see Appendix D: MA2, Y8 L1; MA4 Y7 L1). This was not limited to mixed attainment. For example, in one low set lesson (Appendix D: S4, Y7), the teacher provided dialogic support to enable students to reason and explain about the primes and factors, and discuss and correct errors, even though they found the content difficult. We also observed lessons with a mix of stronger and weaker features. For example, in a Year 8 high set lesson that was largely procedurally focused and examination-oriented, the students were nonetheless encouraged to query the teachers' methods (Appendix D: S3, Y8). However, across both types of schools, we observed many lessons dominated by teacher talk with few opportunities for student input or agency.

More differences emerged when comparing the responses of students with high-, medium-, and low-prior attainment within mixed attainment schools and within setting schools (Table 24). Overall, in both mixed attainment and in setting schools, students with high-prior attainment were more likely than students with medium prior attainment, who were in turn more likely than those with low-prior attainment, to always or frequently feel comfortable talking about their mistakes in class. Similar proportions of students in mixed attainment and in setting schools with high and with medium prior attainment said they always or frequently felt comfortable talking about their mistakes in class. However, students with low-prior attainment were slightly more likely to report always or frequently feeling comfortable talking about mistakes in class if they were in a setting school (49%) than a mixed attainment school (44%).

[25] Tables summarising findings from the teacher and student surveys included in the main report contain percentages only. Full tables including n can be found in Appendices C.13 and C.14.

Similar proportions of students with medium and low-prior attainment reported that they always/frequently understood ideas in mathematics better when they could use a diagram, picture number line, or graph. However, a slightly smaller proportion of students with high-prior attainment reported this in mixed attainment schools (65%) compared with setting schools (71%). Similarly, a slightly smaller proportion of students with high-prior attainment in mixed attainment schools (63%) than in setting schools (68%) reported that their teacher always/frequently linked learning to real-life situations, while responses between study groups were very similar for students with medium- and low-prior attainment.
We observed the use of real-life contexts in both low and high sets, so this (slight) difference for high-attaining students may relate to differences in how higher and lower-attaining students perceive school mathematics contexts as meaningful and authentically 'real-life'. In one low-attaining class (S5, Y7), for example, we observed a class demonstration task which was to share a time period of 40 minutes in proportion to a ratio of 3:5. It is not immediately obvious that this task is one that would be meaningful and realistic to many students. Cooper and Dunne's (2018) research suggests that students from higher SES backgrounds tend to have less difficulty with such tasks in comparison to students from low SES backgrounds.

The overwhelming majority of the lessons observed in both mixed attainment and setting schools were dominated by teacher-led whole-class teaching with all students tackling the same sets of tasks and problems. Often this was interspersed with brief opportunities for paired discussion of less than one minute duration, or the use of responses on individual whiteboards, which the teacher used to check understanding. This is in line with the approach used in schemes such as White Rose Mathematics (WRM), which a number of teachers described using or adapting. We observed very few instances of groupwork. Typically, the focus was on procedures and methods with little focus on developing understandings of the underlying concepts. In some schools (e.g. MA4), students reported that they were encouraged to use their own methods, although in the majority of lessons observed students were presented with just one method. Teachers appeared to place considerable emphasis on ensuring almost all students followed the lesson content and, supporting this, in post-lesson interviews teachers mentioned the goal of mathematics mastery for all learners as shaping their approach.[26] This often resulted in a slow pace in low sets and many mixed attainment classes. In mixed attainment classes, we observed teachers giving some high-attaining students challenges when they had completed the class activity, but we did not observe any discussion of these challenges. Top set lessons were dominated by teacher-led activity, but the pace of these lessons was faster.

Table 23: Student endline survey: How much do you agree with the following statements about mathematics lessons?

	Mixed attainment		Setting	
	% A/SA	%SD/D	% A/SA	% SD/D
I feel comfortable talking about my mistakes in class	50	50	52	48
I understand ideas in maths better when I can use a diagram, picture, number line or graph	68	32	71	29
I understand ideas in maths better when I can use objects to help me, e.g. cubes	58	42	60	40
My teacher uses diagrams to show how to correct a mistake	64	36	63	37
My teacher links what we are learning to real-life situations	61	39	63	37
My teacher asks students to come up to the board to show their ideas	57	43	51	49
My teacher often helps me individually with my work	61	39	63	37

Categories collapsed: SA/A=strongly agree and agree; D/SD=disagree and strongly disagree.

Table 24: Student endline survey: How much do you agree with the following statements about maths lessons?

		High		Middle		Low	
		% A/SA	% SD/D	% A/SA	% SD/D	% A/SA	% SD/D
I feel comfortable talking about my mistakes in class	Mixed	55	45	51	49	44	56
	Setting	54	46	51	49	49	52
	Mixed	65	35	71	29	70	31

[26] The Teaching for Mastery approach is a national initiative that was introduced for mathematics teaching in England in 2014. It aims to achieve mastery of the curriculum content for all, or most, students. It is supported by a national network of Maths Hubs coordinated by the National Centre for Excellence in the Teaching of Mathematics.

I understand ideas in maths better when I can use a diagram, picture, number line or graph	Setting	71	30	72	28	70	31
I understand ideas in maths better when I can use objects to help me, e.g. cubes	Mixed	50	50	61	39	65	35
	Setting	54	46	62	38	65	35
My teacher uses diagrams to show how to correct a mistake	Mixed	64	36	64	36	62	38
	Setting	62	38	64	36	63	37
My teacher links what we are learning to real-life situations	Mixed	63	37	60	40	59	42
	Setting	68	32	62	38	60	40
My teacher asks students to come up to the board to show their ideas	Mixed	53	47	55	45	53	48
	Setting	54	46	51	49	53	48
My teacher often helps me individually with my work	Mixed	59	40	60	40	63	37
	Setting	62	38	62	38	66	34

Comparing responses for students in mixed attainment and in setting schools, with high-, middle-, and low-prior attainment. Categories collapsed: SA/A=strongly agree and agree; D/SD=disagree and strongly disagree.

Student participation in mathematics

We hypothesised that students' mathematical experiences would influence their later participation in mathematics. To investigate this, we included four items from the PISA student questionnaire (OECD, 2013b) in the student endline survey. Student responses were very similar, regardless of whether they attended a school in the mixed attainment group or in the setting group (Table 25). All items showed a trend of students with high-prior attainment being most likely to agree or strongly agree that mathematics would be useful for future study, would help with getting a job and that parents believed mathematics was important for the student's future career, and that parents like mathematics. Those with low-prior attainment were least likely to agree or strongly agree, regardless of whether they were in a mixed attainment school or a setting school (Table 26). The differences between mixed attainment and setting schools were small, suggesting that grouping had no relationship with participation in mathematics, while prior attainment was strongly correlated.

Table 25: Student endline survey: How much do you agree with the following statements about maths? Items taken from PISA student questionnaire (OECD, 2013b)

	Mixed attainment			Setting		
	% SA/A	% N	% D/SD	% SA/A	% N	% D/SD
Mathematics is an important subject for me because I need it for what I want to study later on	62	24	14	64	23	13
I will learn many things in mathematics that will help me get a job	63	24	13	65	23	13
My parents believe that mathematics is important for my career	70	23	1	69	22	8
My parents like mathematics	46	33	21	47	32	21

Categories collapsed: SA/A=strongly agree and agree; N=neither agree nor disagree; D/SD=disagree and strongly disagree.

Table 26: Student endline survey: How much do you agree with the following statements about maths? Items taken from PISA student questionnaire (OECD, 2013b). Comparing responses from students in mixed attainment and setting schools, high-, middle-, and low-prior attainment

Item	Group	High			Middle			Low		
		% SA/A	% N	% D/SD	% SA/A	% N	% D/SD	% SA/A	% N	% D/SD
Mathematics is an important subject for me because I need it for what I want to study later on	Mixed	70	19	11	60	26	14	55	28	17
	Setting	70	19	11	63	24	13	58	27	15
I will learn many things in mathematics that will help me get a job	Mixed	71	19	10	62	25	13	56	28	16
	Setting	70	19	11	63	24	13	60	26	15
My parents believe that mathematics is important for my career	Mixed	79	17	4	69	24	7	61	27	12
	Setting	76	18	6	68	23	9	62	26	11
My parents like mathematics	Mixed	55	31	15	46	34	20	36	35	29
	Setting	57	27	16	45	33	22	36	37	27

Categories collapsed: SA/A=strongly agree and agree; N=neither agree nor disagree; D/SD=disagree and strongly disagree.

Responsiveness of pedagogy

Teacher beliefs about students learning mathematics

In the endline survey, teachers were asked a number of questions about their beliefs about students learning mathematics, and several differences emerged between teachers in mixed attainment and in setting schools (Table 27).

Teachers in mixed attainment schools were much more likely than those in setting schools to agree or strongly agree that all students should have the OTL the whole mathematics curriculum, regardless of prior attainment (81% in mixed attainment schools vs 66% in setting schools). It is striking that this belief is held by a large majority of teachers in mixed attainment schools. Teachers in mixed attainment schools were also more likely than those in setting schools to agree or strongly agree that all students can do well in mathematics (83% in mixed attainment schools vs 74% in setting schools). It is notable that so many teachers in mixed attainment schools believe in universal opportunity and potential in mathematics. Many fewer teachers in mixed attainment schools, but still more than in setting schools, agreed or strongly agreed that all students benefit from the same types of pedagogy when learning mathematics, regardless of prior attainment (52% in mixed attainment schools vs 43% in setting schools).

Unsurprisingly, there was strong divergence in teachers' beliefs about the most beneficial approaches to grouping students for learning mathematics. Teachers in setting schools were far more likely to agree or strongly agree that students should be grouped with others of a similar prior attainment level (53% in mixed attainment schools vs 92% in setting schools), while teachers in mixed attainment schools were much more likely to support groups with a wide range of prior attainment (60% in mixed attainment schools vs 22% in setting schools). It is notable here that support for groups with a wide range of prior attainment was much weaker overall, and that just over half of teachers in mixed attainment schools believed that students benefited from being in classrooms with a similar prior attainment level, i.e. setting or similar.

Teachers were much more similar in their beliefs about students' need for challenge and support. A large majority of teachers in mixed attainment and in setting schools agreed or strongly agreed that all students benefited from being challenged (93% in mixed attainment schools and 91% in setting schools) and supported (90% in mixed attainment schools and 95% in setting schools).

Table 27: Teacher endline survey: About students learning maths

	Mixed attainment			Setting		
	% SA/A	% N	% D/SD	% SA/A	% N	% D/SD
All students should have the OTL the whole maths curriculum, regardless of prior attainment.	81	3	17	66	5	29
All students benefit from the same types of pedagogy when learning maths, regardless of prior attainment.	52	11	37	43	11	46
Students benefit from learning in maths classrooms with students of a similar prior attainment level	53	26	21	92	5	3
Students benefit from learning in maths classrooms with a wide range of prior attainment in maths	60	19	22	22	18	60
All students can do well in maths	83	7	10	74	10	17
All students benefit from being challenged mathematically, regardless of their prior attainment	93	3	5	91	4	5
All students benefit from support when learning mathematics, regardless of their prior attainment	90	5	5	95	3	2

Categories collapsed: SD/D=strongly disagree and disagree; N=neither agree nor disagree; A/SA=agree and strongly agree.

Teachers' pedagogical practices

In the endline survey, teachers were asked a number of questions about their pedagogical practices, and revealed a number of similarities and differences emerged between teachers in mixed attainment and in setting schools (Table 28).

Classroom management and assessment practices, as reported by teachers, were similar across the two groups of schools with, for example, more than 90% of teachers reporting that they frequently or always told students to follow classroom rules, that they frequently or always sat students who need most help to enable easy access and that they frequently or always observed students and provided immediate feedback. Similar proportions of teachers reported that they frequently or always referred to problems from everyday life and explained the utility of new mathematics knowledge (71% mixed attainment compared with 66% setting). There were some slight differences between the two groups on review and goal setting: more teachers in in setting schools reported that they frequently or always presented a summary of recently learned content (68% compared with 63% in mixed attainment schools); and set goals at the beginning of instruction (73% compared with 66% in mixed attainment schools).

There were larger differences in teachers' reports of the way they approached the learning of mathematics and the extent to which they offered students opportunities for problem-solving and critical thinking. For example, teachers in mixed attainment schools were more likely to report that they frequently or always presented tasks with no obvious solution (46% vs 25%), that they used problem-solving tasks as a starting point (39% vs 22%), they gave tasks that require critical thinking (75% vs 65%) and allowed students to decide on their own procedures for solving complex tasks (53% vs 42%). Mixed attainment teachers were also more likely to report that they frequently or always used group work (33% vs 18%). Further, echoing the student survey findings, mixed attainment teachers were more likely to report that they frequently or always used representations (82% vs 69%), manipulatives (29% vs 18%), and brought students up to the board to show their ideas (41% vs 29%). In contrast, teachers from setting schools were slightly more likely to report that they were frequently or always explicit about explaining expectations for learning (91% vs 77%) and made connections between new and old topics (88% vs 80%). These reported differences were not strongly supported by our observations of lessons: for example, we saw very little, small group or pair work in any schools, few open-ended tasks or problem-solving opportunities, and limited use of manipulatives. While we did observe one lesson with an open-ended task in a mixed attainment school, for the most part lessons in mixed attainment and setting schools were broadly similar.

Table 28: Teacher endline survey: Thinking about your teaching of mathematics to Year 7 and Year 8, how often do you do the following?

	Mixed		Setting	
	% N/O	% F/A	% N/O	% F/A
I present a summary of recently learned content	37	63	32	68
I set goals at the beginning of instruction	34	66	27	73
I explain what I expect the students to learn	23	77	9	91
I explain how new and old topics are related	20	80	12	88
I present tasks for which there is no obvious solution	54	46	75	25
I give tasks that require students to think critically	25	75	36	65
I have students work in small groups to come up with a joint solution to a problem or task	67	33	82	18
I ask students to decide on their own procedures for solving complex tasks	47	53	58	42
I tell students to follow classroom rules	5	95	6	94
I tell students to listen to what I say	7	93	5	96
I calm students who are disruptive	5	95	7	93
When the lesson begins, I tell students to quieten down quickly	6	94	8	92
I refer to a problem from everyday life or work to demonstrate why new knowledge is useful	29	71	34	66
I let students practise similar tasks until I know that every student has understood the subject matter	25	75	19	81
I have students come up to the board to show their ideas	59	41	71	29
I give students manipulatives to use in lessons to help them learn	71	29	82	18
I use representations in lessons to help students learn	18	82	31	69
I observe students when working on particular tasks and provide immediate feedback	9	91	5	95

I seat students who need the most help where I can easily get to them	5	95	4	96
I give students problem-solving tasks as an extension	22	78	27	73
I give students problem-solving tasks as a starting point for learning	61	39	78	22

Categories collapsed: N/O=never and occasionally; F/A=frequently and always.

Schemes of work

Schemes of work were collected from five schools: two from mixed attainment schools, and three from setted schools. The mathematical content was similar across all five, reflecting the nature of the national curriculum and the national focus on mastery. For example, two of these schools had adopted WRM,[27] a scheme currently adopted by around a fifth of schools (Observatory for Mathematical Education, 2025). WRM encourages the implementation of mastery through a 'small steps' approach. In these schools, the small steps approach was used to describe and structure the sequence of lessons, although in some cases (e.g. 'understand the meaning of equivalence') these 'small steps' describe relatively large and cognitively demanding conceptual ideas. There were some noteworthy differences between the two groups. In particular, the two mixed attainment schools described cognitive challenge for students at all attainment levels in terms of regular investigative and thinking lessons (MA5) and depth and making connections (MA4). In contrast, the two non-WRM setted schemes of work described the curriculum in terms of differentiated content divided into learning outcomes for support, core and extension (S1) and for each of four set levels (S5) with significant overlap between the different levels.

Contextual factors

Heads of mathematics reported in the survey on who had been responsible for decisions about attainment grouping in their school (Table 29). In the majority of cases in setting schools and in about half of mixed attainment schools, the decision was made by the head of mathematics. Mixed attainment grouping was more likely than setting to be determined by the school's senior leadership team or by the Multi-Academy Trust. In two cases, survey responses implied that the head of mathematics themselves may not have fully supported the decision to use mixed attainment grouping, for example, *'The previous head of department and current head teacher believes there is an advantage to teaching mixed attainment over sets in mathematics'*.

Table 29: Head of mathematics survey: Whose decision is it to group in this way (mixed attainment/setting)?

	Mixed (n=30)	Setting (n=68)
Head of mathematics/middle leader	14	43
Senior leadership team	8	12
Multi-Academy Trust	3	0
By mutual agreement within the mathematics department	5	13

Findings from case study schools were very similar. For example, in school MA1,[28] mixed attainment grouping had been introduced to promote equity in curriculum and expectations, and to prevent the creation of 'sink sets' with poor behaviour. In school MA2, mixed attainment was used as much as possible in Years 7 and 8 for social justice motivations, so that all students would have the same opportunities regardless of prior experiences. At school MA5, it was felt that there was a successful track record of mixed attainment grouping in mathematics, and a belief that students should have equal access to learning because they develop at different rates and could excel at different areas of mathematics.

In setting schools, some heads of mathematics mentioned the challenge of teaching students with a large range of prior attainment in the same classroom, or said that it was easier to plan and deliver lessons to students grouped in sets. For example, the head of mathematics at school S5 said that mixed attainment was impractical for the 'massive range in ability' of students. Some heads of mathematics mentioned that setting was a longstanding tradition in their school. Two heads of mathematics said that they felt setting reduced anxiety or stress for low-attaining students. Heads of mathematics at setting schools also mentioned that they felt that students were making good progress in sets.

[27] See: https://whiteroseeducation.com/
[28] All case study schools have been given pseudonyms.

What factors associated with the specific school context influence grouping practices?
As schools in the mixed attainment and setting groups were matched for the purposes of the study, overall we would expect their school contexts to be similar. It is notable that heads of mathematics in the mixed attainment and setting groups reported different responses to their similar contexts.

For example, as detailed in Appendix D, the head of mathematics at school MA5 believed that mixed attainment grouping was the most effective way to group their intake, which included a high proportion of students from disadvantaged backgrounds. By contrast the head of mathematics at school S2 felt that their significant proportion of students from disadvantaged backgrounds required a supportive environment in sets.

The head of mathematics at school MA2 felt that mixed attainment enabled them to spread students from a small number of feeder primary schools that achieved high outcomes in mathematics across a large number of classes, whereas the head of mathematics at school S6 regarded setting as most appropriate to meet the needs of their students, who came from a large number of primary schools and had diverse backgrounds.

The influence of parents was also mentioned, with the head of mathematics at school S5 saying that parents of their aspirational intake would frequently challenge grouping decisions, particularly in relation to GCSE tier of entry. Their response was to communicate consistently to students and parents that group placement was designed to benefit the student by aligning their work to their 'ability' level and that places in higher groups needed to be earned. At school MA4, the head of mathematics gave a contrasting perspective. Here, mixed attainment grouping was felt to reduce the anxiety that girls and their parents experienced over set placement and movement between sets. Because all students learned together, everyone had the chance to encounter different approaches, higher-level skills and problem-solving opportunities.

Mathematics teacher characteristics
The teacher endline survey received 611 complete responses from teachers in 94 schools. There were 196 responses from teachers in 30 mixed attainment schools, and 415 responses from teachers in 64 setting schools.

Teachers were asked a number of questions about their experience, training, and development. Table 30 provides data about teachers' highest qualifications in mathematics. The majority of teachers had at least an undergraduate degree in mathematics or a related subject (e.g. economics, business, sciences), although the proportion was slightly higher in mixed attainment schools (81%) than in setting schools (77%).

Table 31 shows the routes into mathematics teaching taken by respondents. The majority of respondents had become mathematics teachers through the PGCE mathematics route (61% in mixed attainment schools and 56% in setting schools). Similar proportions of teachers had trained through the PGCE route to teach subjects other than mathematics (6% in mixed attainment schools and 5% in setting schools) or had taken other initial teacher training routes (33% in mixed attainment schools and 40% in setting schools). Teachers who responded 'Another route' had predominantly qualified via an undergraduate pathway (BA or BEd with QTS, n=36) or Graduate Teacher Programme (n=10). Smaller numbers (fewer than ten) had followed other pathways including training overseas or via an apprenticeship, assessment-only ITT, PGDE, further education, primary, teaching out of specialism, converting from another subject or were working as an unqualified teacher. The same proportion of teachers regarded themselves as specialist mathematics teachers in both types of school (Table 32: 90% in both types of school).

Overall, teachers in mixed attainment schools tended to be slightly less experienced than teachers in setting schools (Table 33), with 36% of teachers in mixed attainment schools having taught for less than five years, compared with 29% of teachers in setting schools. Conversely, setting schools had more experienced teachers with 45% of teachers having at least 11 years of experience, compared with 39% in mixed attainment schools.

Table 30: Teacher endline survey: What is your highest qualification in mathematics?

	Mixed		Setting	
	N	%	N	%
Postgraduate degree (MA, PhD, equivalent)	32	16	51	12
Undergraduate degree – mathematics	77	39	173	42
Undergraduate degree – related subject	49	25	97	23
A level/equivalent	28	14	69	17
GCSE/equivalent	6	3	21	5
Other	4	2	4	1

Table 31: Teacher endline survey: What was your route to teaching mathematics?

	Mixed		Setting	
	N	%	N	%
PGCE mathematics	119	61	232	56
PGCE other	12	6	21	5
School-based ITT mathematics	30	15	81	20
School-based ITT other	0	0	7	2
Teach First mathematics	4	2	11	3
Another route	31	16	63	15

Table 32: Teacher endline survey: Do you consider yourself to be a specialist mathematics teacher?

	Mixed		Setting	
	N	%	N	%
Specialist	176	90	375	90
Non-specialist	20	10	40	10

Table 33: Teacher endline survey: How many years have you been teaching?

	Mixed		Setting	
	N	%	N	%
0–2 years	37	19	54	13
3–5 years	33	17	65	16
6–10 years	50	26	108	26
11–20 years	51	26	127	31
21+ years	25	13	61	15

Continuing professional development (CPD)

Teachers were asked to estimate how many days they had spent on CPD with a mathematics focus over the last school year (Table 34). The median number of days spent on CPD with a mathematics focus led by a member of school staff was higher in mixed attainment schools (two days) than in setting schools (one day). The mean number of days spent on CPD with a mathematics focus led by an external facilitator was overall lower than for school-led CPD, and was also higher in mixed attainment schools (half-days) than in setting schools (zero days). We note that teachers appeared to find these questions difficult to answer, as the range of numbers of days of CPD reported was very large. Some teachers recorded surprisingly large numbers of days, amounting to up to 15% of their annual contracted days. Significant proportions of teachers reported spending no days on CPD with a mathematics focus, particularly with external facilitation.

Table 34: Teacher endline survey: Please estimate how many days this school year (2023/2024) you spent on CPD with a mathematics focus

	Mixed	Setting
	...led by someone from your school	
N	190	400
Mean (SD)	4.1 (6.0)	2.4 (3.8)
Median	2	1
% receiving zero days	14.2	26.3
	...led by someone external to your school	
N	193	401

Mean (SD)	2.0 (2.9)	1.0 (2.4)
Median	0.5	0
% receiving zero days	45.6	64.3

Experience of different grouping practices

Unsurprisingly there were differences between teachers in mixed attainment and setting schools in their prior experiences of attainment grouping practices (Table 35). While nearly all teachers in mixed attainment schools said that they had experience of mixed attainment teaching, this was the case for only around half of teachers in setting schools. The majority (8%) of teachers in mixed attainment schools had experienced setting. Streaming was less common, experienced by around two-fifths of teachers in mixed attainment schools and a quarter of teachers in setting schools; and while over half of teachers in mixed attainment schools had experience of nurture groups, this was the case for only a quarter of teachers in setting schools.

Table 35: Teacher endline survey: Which of the following types of grouping for maths have you experienced as a teacher? Please select all that apply

	Mixed		Setting	
	N	%	N	%
Mixed attainment	194	99	217	52
Setting	159	81	397	96
Streaming	78	40	107	26
Nurture group	106	54	106	26
Total in group	196	100	415	100

Summary

- What are schools' specific current grouping practices in Year 7 and Year 8 for mathematics? What are the reasons for these practices?

Heads of mathematics in mixed attainment schools were more likely to report that grouping practices were motivated by a desire for equity or a belief that everyone could be a mathematician, while heads of mathematics in setting schools were more likely to say that their chosen approach was motivated by a belief that it would lead to better teaching. Some mixed attainment schools had nurture groups, established to meet the need of the lowest-attaining students. There was significant variation in the numbers of classes, and in setting schools the number of set levels. This was largely influenced by cohort size.

- What factors associated with the specific school context influence mixed attainment grouping or setting practices?

Due to the matching approach, schools in the mixed attainment and setting groups were contextually similar. In setting schools, heads of mathematics identified the experience of their teachers as meaning that setting was the more appropriate practice, however, setting had also been experienced by the majority of teachers in mixed attainment schools. Teachers in both groups had entered teaching through similar training routes, and a similar proportion of teachers in both groups considered themselves to be mathematics specialists. Teachers in mixed attainment schools were on average slightly less experienced than those in setting schools, but there was no evidence that this influenced decisions about grouping.

- What pedagogic practices do teachers use in Year 7 and Year 8 mathematics lessons, for mixed attainment grouping/setting practices? To what extent, if any, is this influenced by the attainment of students?

Teachers reported using some similar pedagogic approaches, but there were also some points of difference between teachers in mixed attainment and setting schools. However, our observations indicated that there were few differences in practice. Lessons in mixed attainment and in setting schools were, in general, quite similar. Lessons in mixed attainment schools were often more similar to those for low sets than for high sets.

- How do students with low-prior attainment experience mixed attainment grouping or setting? What are the beneficial and detrimental effects of these experiences associated with mixed attainment grouping or setting?

Students in mixed attainment and in setting schools made similar reports about their experiences of mathematics lessons. This was borne out by our observations, where we saw lessons in both types of school and for both high and low sets predominantly featuring teacher-led whole-class instruction followed by individual practice. There were very few opportunities for pair or group work in any of the lessons observed. Students with low-prior attainment had similar experiences in mixed attainment and in setting schools.

Cost

In both mixed attainment and setting schools, grouping practices were schools' usual practices. No new intervention was implemented in either group of schools for the purposes of this study. As reported above, we asked teachers about the number of days' CPD with a mathematics focus they had received in the last year. The median number of days' CPD delivered by a school colleague reported by teachers in mixed attainment schools was two days, compared with one day in setting schools. For CPD delivered by an external provider, the mean number of days reported by teachers in mixed attainment schools was a half-day, compared with zero days in setting schools. This means that on average teachers in mixed attainment schools were receiving one day more internal training and half-day more external training than colleagues in setting schools.

None of the case study schools reported having recently changed to mixed attainment grouping from setting, so it was not possible to estimate from their experiences the cost of moving from one practice to the other.

It seems reasonable however, to conclude that there may be small additional continuous training required in implementing mixed attainment grouping in mathematics when compared with setting. Survey data suggests that this would include no more than one day of internal training per year and, thus, estimated to be of minimal cost.

Table36: Additional cost of delivering mixed attainment grouping in mathematics in Years 7 and 8, compared with setting

Item	Type of cost	Cost	Total cost over three years	Total cost per student per year over three years
Teacher CPD - school-led	Annual training and development	Nil	Nil	Nil
Teacher CPD – external provider	Annual training and development	Nil	Nil	Nil
Total	–	Nil	Nil	Nil

Conclusion

Table 37: Key conclusions

Key conclusions
Students in schools using mixed attainment grouping made one month's less progress in mathematics, on average, compared to students in schools using setting. This is our best estimate of impact, which has a low/moderate security rating.
Low-prior attaining students in schools using mixed attainment grouping made similar progress in mathematics, compared to students in schools using setting.
High-prior attaining students in schools using mixed attainment grouping made, on average, two fewer months' progress in mathematics compared to students in schools using setting.
Students in receipt of FSM made similar progress in mathematics, on average, in schools using mixed attainment and schools using setting.
The IPE lesson observations indicated that schools' teaching practices are broadly similar across mixed attainment and setted schools and that overall practices in most case study schools appear to be driven by national policy around mastery. In general, despite well-intentioned policies in mixed attainment schools around equity and challenge for high attainers, only setting schools appear to be challenging high-prior attaining students.

Impact evaluation and IPE integration

Interpretation

The impact analysis findings show that, for all students, there was a small negative, non-significant effect on the progress of students taught in mixed attainment classes compared to those taught in setted classes. In terms of our primary objective, we found a zero effect of mixed attainment teaching on the mathematics attainment of low-prior attaining students. We also found, a small negative, non-significant effect on the attainment of middle prior attaining students, and a moderate negative, statistically significant effect on the attainment of high-prior attaining students. Hence, there is some evidence for a reduction in the mathematics attainment gap between low- and high-prior attaining students when taught in mixed attainment classes. This appears to be largely due to less progress being made by the high-prior attaining students, rather than greater progress by the low-prior attaining students. For the subgroup of students in receipt of FSM, there was effectively no difference in attainment, indicating that, contrary to some qualitive evidence, students from disadvantaged backgrounds are not further disadvantaged by the practice of setting. In addition, we found statistically significant, negative effects on general self-confidence for students in mixed attainment groups, which were small for the student group as a whole and for students in receipt of FSM, and moderate for the low-prior attaining students.

Students taught in mixed attainment classes make one month's less progress in mathematics compared to students in schools using setting. Although there was a small numerical difference in test scores, it is possible that this reflects normal variation among students and schools, rather than a real impact of the grouping method itself. This finding is broadly consistent with the EEF Toolkit (EEF, 2021) and the wider literature (Steenbergen-Hu *et al.,* 2016). Similarly, the findings of a small gain for low-prior attaining students and a reduction in the attainment gap between low- and high-prior attaining students are consistent with existing research (e.g. Terrin and Triventi, 2023). However, the finding of a moderate, negative for high-prior attaining students is only partially consistent with the EEF Toolkit (EEF, 2021), which indicates a smaller, but still negative effect, for high-prior attainers, broadly equivalent in magnitude to a small gain for low-prior attainers.

In a study of setted schools based on the EEF-funded earlier projects, 'BPGS', we found that students placed in higher sets made greater gains than those placed in lower sets (Hodgen *et al.,* 2023b). On the basis of this earlier work and the wider literature on setting in England (e.g. Ireson *et al.,* 2005; Boaler *et al.,* 2000), we hypothesised that, in England, the negative impact on low-prior attaining students would have a greater negative impact than previously thought due to being placed in lower sets and that low-prior attaining students would have higher levels of self-confidence when taught in mixed attainment classes. Neither of these hypotheses was supported by our findings. We also hypothesised that students from disadvantaged backgrounds, as indicated by FSM status, would have better outcomes when taught in mixed attainment classes, because the proportion of these students who have lower than average attainment is disproportionately high

(Noyes *et al.,* 2023) and because the qualitative literature on attainment grouping suggests this (Dunne *et al.,* 2007). This hypothesis was not supported by our findings.

The EEF Toolkit (EEF, 2021) judges the current evidence on the effects of attainment grouping on student progress to be very limited. Indeed, despite numerous meta-analyses (e.g. Steenbergen-Hu *et al.,* 2016), much of the quantitative and experimental literature is dated and situated in the United States context of tracking (Rui, 2009; e.g. see also, Hodgen *et al.,* 2023b, for a discussion). The English educational context is very different to the United States and includes an entitlement the whole-mathematics curriculum. We believe our study to be the first large-scale robust study of the effects of attainment grouping and the first to robustly examine the effects on disadvantaged students. Our study, thus, adds to the evidence base about attainment grouping in general and provides robust evidence on the impact of attainment grouping in the context of the English educational system, including those from disadvantaged backgrounds.

It may be that the current context of mathematics education in England has had a particularly strong influence on our finding of a moderate, negative effect on the attainment of high-attaining students. In 2017, England introduced Teaching for Mastery, an approach focused on achieving mastery of the basic content for all and greater depth for some. The intention was that some Teaching for Mastery was influenced by the mathematics teaching in East Asian systems with a specific focus on teaching in Shanghai. A key takeaway from the Shanghai Teacher Exchange was to slow the pace of lessons (Boylan *et al.,* 2019). The lessons that were observed in the current project were very different to the mathematics lessons observed in our previous project, BPGS, prior to the introduction of mastery. Compared to the lessons observed for BPGS, the lessons observed for the current project in general had less differentiation, involved a greater degree of teacher-led explicit teaching, almost no opportunities for pair or group work, or for exploratory/open-ended activities and, aside from the current top set lessons, lessons were generally slower paced.

For the current project, we observed top and bottom set lessons from setted schools as well as lessons from mixed attainment schools. Classroom mathematical learning experiences were highly variable and not consistently associated with high/low sets or mixed attainment groups. We observed small numbers of lessons for all types of class that were rated highly on the TRU, but many more lessons that were rated at the lower end of the TRU scale. Nevertheless, we found lessons from mixed attainment schools tended to be more similar to the bottom set than top set in terms of pace and, although there were opportunities for challenging extension activities for high-attaining students, these opportunities frequently involved mathematics that was not directly related to the lesson content and were rarely discussed in class. We did however, see some examples of practice that might limit opportunities for high-prior attaining students, for example, there being a lack of resources to stretch high-prior attainers in the classroom, or a focus on using high-prior attainers to support low-prior attainers rather than providing carefully structured pair or group work that might allow high-, middle-, and low-prior attainers to mutually support one another. We also saw examples of classwork being too easy for high-prior attainers, or high-prior attainers reporting that they were going over familiar material rather than learning something new. Thus, contrary to the aims of Teaching for Mastery, in mixed attainment classes, the opportunities for developing more depth in understanding for the high-attaining students were limited. In the absence of teaching strategies that focus on depth of understanding, this may disadvantage the highest attaining students. In contrast, top sets provided more challenge through a faster pace and a greater emphasis on GCSE (even though these students were in Year 7 and Year 8, three or four years before the GCSE examination).

In our logic model, we hypothesised that schools' attainment grouping practices would indirectly influence student outcomes through the mediation of factors including OTL and teacher quality. We found no statistically significant differences for students as a whole, for low-prior attaining students or for students in receipt of FSM, thus limiting the power of the mediation analysis to detect statistically or practically significant effects. Hence, unsurprisingly, we found little evidence to suggest that the effects that we found were mediated by either OTL or teacher quality. As noted above, evidence from lesson observations largely corroborated this finding, although top sets tended to be higher paced providing greater OTL. Although we found few differences in overall school ethos or resourcing based on teacher and school leader surveys, evidence from the IPE case studies did suggest some differences between the two groups of schools. For example, heads of mathematics in mixed attainment schools appeared more frequently to believe that all students could be successful in mathematics regardless of prior attainment and mixed attainment schools tended to take more account of information received from primary schools. As we expected, nurture groups and low sets tended to be smaller and have more access to teaching assistants. However, the largely inconclusive findings from the impact evaluation render the question of the logic model somewhat moot.

Limitations and lessons learned

In conclusion, for students overall, those in mixed attainment classes made one month less progress in mathematics, compared with those taught in setted classes. This result was not statistically significant, meaning that there was uncertainty around the result, with the possible difference for students in schools using mixed attainment grouping ranges from two months less progress to no difference in progress. However, we found that for high prior attaining students in schools using mixed attainment grouping made, on average, two fewer months' progress in mathematics compared to students in schools using setting. This finding was statistically significant, meaning that it was unlikely to be due to chance. The IPE indicated that schools practices are broadly similar across mixed attainment and setted schools and that overall practices appear to be driven by national policy around mastery rather than grouping practices. In general, despite well-intentioned policies in mixed attainment schools, only setting schools appear to be challenging high-prior attaining students.

There were a number of strengths to this study. The impact evaluation was based on a comparison of matched samples of mixed attainment and setted schools. While matching is more prone to bias than a random sample of schools, we took steps to ensure that the threat of bias was low (see Appendix C.22). The IPE took a mixed-methods approach combining quantitative evidence with robust case studies based on a random selection of schools that were requested to choose their strongest practice for observations, although the number of case study schools was relatively small. Unlike the EEF-funded RCTs, this was a natural experiment based on schools' existing grouping practices. As a result, compliance was high; there were no schools that changed their practices. In general, schools and teachers were highly committed to the grouping approach that their school had adopted, although some heads of mathematics in mixed attainment schools expressed some doubts about the approach.

One limitation of our study relates to the selection of the two matched groups of schools and to the relatively high levels of attrition. At the point of recruitment, there were relatively few schools in England that met our mixed attainment criteria (Taylor *et al.,* 2020). The matching process yielded smaller numbers of matched setted schools than were anticipated and only some of these schools agreed to take part in the trial. In addition, attrition was relatively high. A small number of schools dropped out due to changes in school leadership. A larger number of schools dropped out as a consequence of this, because they were no longer matched. Hence, although the two groups of schools were well-matched, the schools may not fully represent the national population of schools. In addition, there were slight imbalances in the baseline characteristics of the two groups. The schools in the mixed attainment group were larger and fewer of them were Academies or rated as 'Outstanding' by Ofsted. The Ofsted rating imbalance suggests that the quality of teaching in the mixed attainment group overall was lower than that for the setting group, although we note that the numbers, and imbalances, were small.

A second limitation is that our focus was on attainment tertiles rather than on allocation to low, middle, or high sets. This may be problematic, due to the misallocation of students to sets (Connolly *et al.,* 2020). For example, some high-prior attainers may not have been placed in top sets. We intend to carry out additional analysis beyond that planned in the Statistical Analysis Plan (Anders *et al.,* 2024) to explore this.

A third limitation is that some schools supervised outcome tests themselves. This is unlikely to be a significant source of bias. There is no reason to believe any malpractice and a sensitivity analysis found no evidence of bias.

A fourth limitation is that we studied schools with established grouping practices. Hence, we cannot comment on the short-term impacts nor on the costs associated with a school changing their practices.

Future research and publications

Currently, the EEF Toolkit (EEF, 2021) recommends that schools choosing to use setting and streaming should consider how the approach will enable more effective teaching for all students, including lower-attaining students. One implication of our study is that schools choosing to use mixed attainment should also consider how the approach will enable more effective teaching for all students, including higher-attaining students. Our findings indicated that, in general, the top sets in the setted schools challenge high-attaining students in ways that the mixed attainment schools do not. The IPE indicated that, in general, mixed attainment schools do consider and put in place strategies for stretching and challenging high attainers. More research is needed into both these issues: what strategies do setted schools use to challenge high attainers and which of these are most effective, and what strategies could mixed attainment schools use to effectively challenge high attainers within mixed attainment classes.

Our study shows only a small and non-significant effect for mixed attainment teaching on the progress of low-prior attaining students. The problem of low attainment in mathematics remains a serious problem and it is therefore, important that research considers other approaches to raising the attainment of this group of students. More research is needed into methods of increase challenge and depth of understanding for all.

Finally, although we found little evidence to suggest that the effects of mixed attainment were mediated by either OTL or teacher quality, we did find some weak evidence to suggest that our measures of OTL and teacher quality may be negatively associated with student progress. If replicated elsewhere, this would suggest that these measures are not capturing the most relevant aspects of either factor. For example, OTL can be conceptualised as comprising the amount, or breadth, of content taught together with the academic learning time devoted to the topics (e.g. Kurz, 2011). It may be that our OTL measure captures breadth of content covered, but not the time in which students are engaged. Hence, it may not capture the extent to which students have the time to develop depth of understanding. We note also that our teacher quality measure was based wholly on student report and it may be that students are not fully aware of some aspects of teacher quality. Further research is needed to develop and validate the measure of these factors.

References

Allen, R., Jerrim, J., Parameshwaran, M. and Thompson, D. (2018) *'Properties of Commercial Tests in the EEF Database'*. London: Education Endowment Foundation.

Anders, J., Brown, C., Ehren, M., Greany, T., Nelson, R., Heal, J., Groot, B., Sanders, M. and Allen, R. (2017) *'Evaluation of Complex Whole-School Interventions: Methodological and Practical Considerations'*. London: Education Endowment Foundation.

Anders, J., Hodgen, J., Taylor, B. and Bretscher, N. (2024) *'The Student Grouping Study: Investigating the Effects of Setting and Mixed Attainment Grouping. Statistical Analysis Plan'*. London: Education Endowment Foundation. Available at: https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/student-grouping-study

Archer, L., DeWitt, J., Osborne, J., Dillon, J., Willis, B. and Wong, B. (2012) 'Science Aspirations, Capital, and Family Habitus: How Families Shape Children's Engagement and Identification With Science'. *American Educational Research Journal,* 49: 5, 881–908. https://doi.org/10.3102/0002831211433290

Archer, L., Francis, B., Miller, S., Taylor, B., Tereshchenko, A., Mazenod, A., Pepper, D. and Travers, M.-C. (2018) 'The Symbolic Violence of Setting: A Bourdieusian Analysis of Mixed Methods Data on Secondary Students' Views About Setting'. *British Educational Research Journal,* 44: 1, 119–140. https://doi.org/10.1002/berj.33

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., Klusmann, U., Krauss, S., Neubrand, M. and Tsai, Y-M. (2010) 'Teachers' Mathematical Knowledge, Cognitive Activation in the Classroom, and Student Progress'. *American Educational Research Journal,* 47: 1, 133–180. http://dx.doi.org/10.3102/0002831209345157

Boaler, J., Wiliam, D. and Brown, M. (2000) 'Students' Experiences of Ability Grouping – Disaffection, Polarisation and the Construction of Failure'. *British Educational Research Journal,* 26: 5, 631–648. https://doi.org/10.1080/713651583

Boylan, M., Wolstenholme, C., Demack, S., Maxwell, B., Jay, T., Adams, G., and Reaney, S. (2019) *Longitudinal evaluation of the Mathematics Teacher Exchange: China-England - Final Report*. London: Department for Education.

Braun, V. and Clarke, V. (2019) 'Reflecting on Reflexive Thematic Analysis'. *Qualitative Research in Sport, Exercise and Health,* 11: 4, 589–597. https://doi.org/10.1080/2159676X.2019.1628806

Burris, C.C., Heubert, J.P. and Levin H.M. (2006) 'Accelerating Mathematics Achievement Using Heterogeneous Grouping'. *American Educational Research Journal,* 43: 1, 137–154. https://doi.org/10.3102/00028312043001105

Caliendo, M. and Kopeinig, S. (2008) 'Some Practical Guidance for the Implementation of Propensity Score Matching'. *Journal of Economic Surveys,* 22: 1, 31–72. https://doi.org/10.1111/j.1467-6419.2007.00527.x

Connolly, P., Taylor, B., Francis, B., Archer, L., Hodgen, J., Mazenod, A., and Tereshchenko, A. (2020) 'The misallocation of students to academic sets in maths: a study of secondary schools in England'. *British Educational Research Journal*, 45 (4), 873–897. https://doi.org/10.1002/berj.3530

Cooper, B. and Dunne, M. (1998) 'Anyone for Tennis? Social Class Differences in Children's Responses to National Curriculum Mathematics Testing'. *The Sociological Review*, 46 (1), 115–148. https://doi.org/10.1111/1467-954x.00092

Cooper, P. and Whitebread, D. (2007) 'The Effectiveness of Nurture Groups on Student Progress: Evidence From a National Research Study'. *Emotional and Behavioural Difficulties,* 12: 3, 171–190. doi:10.1080/13632750701489915

Delisle, J.R. (2015) *'Why Differentiation Doesn't Work'.* EducationWeek. Available at: www.edweek.org/ew/articles/2015/01/07/differentiation-doesnt-work.html (accessed 20 May 2020).

DfE, (2015) '*Supporting the attainment of disadvantaged pupils: articulating success and good practice'*. London: DfE.

Dunne, M., Humphreys, S., Sebba, J., Dyson, A., Gallannaugh, F. and Muijs, D. (2007) *'Effective Teaching and Learning for Pupils in Low Attaining Groups. Research Report DCSF-RR011'.* Nottingham: Department for Children, Schools and Families

Dunne, M., Humphreys, S., Dyson, A., Sebba, J., Gallannaugh, F. and Muijs, D. (2011) 'The Teaching and Learning of Pupils in Low-Attainment Sets'. *The Curriculum Journal,* 22: 4, 485–513. doi:10.1080/09585176.2011.627206

Education Endowment Foundation (EEF). (2017) *'Improving Mathematics at Key Stages Two and Three: Guidance Report'*. London: Education Endowment Foundation. Available at: https://educationendowmentfoundation.org.uk/education-evidence/guidance-reports/maths-ks-2-3 (accessed 22 April 2026).

Education Endowment Foundation (EEF). (2019) 'Classification of the Security of Findings From EEF Evaluations (Version 2.0)'. London: Education Endowment Foundation. Available at: www.education.ox.ac.uk/wp-content/uploads/sites/3/2024/12/classifying_the_security_of_eef_findings_2019.pdf (accessed 22 April 2026).

Education Endowment Foundation (EEF). (2021) *'Teaching and Learning Toolkit'*. London: Education Endowment Foundation. Available at: https://educationendowmentfoundation.org.uk/education-evidence/teaching-learning-toolkit (accessed 22 April 2026).

Education Endowment Foundation (EEF). (2022) 'Statistical analysis guidance for EEF evaluations'. London: Education Endowment Foundation. Available at: https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluation/evaluation-guidance-and-resources/evaluation-design (accessed 22 April 2026).

Education Endowment Foundation (EEF). (2022) *'Implementation and Process Evaluation Guidance for EEF Evaluations'*. London: Education Endowment Foundation. Available at: https://educationendowmentfoundation.org.uk/public/files/Evaluation/Setting_up_an_Evaluation/IPE_guidance.pdf (accessed 20 April 2026).

Evidence-Based Education. (2022) *'Great Teaching Toolkit: Student Surveys'*. Evidence Based Education.

Francis, B., Archer, L., Hodgen, J., Pepper, D., Taylor, B. and Travers, M.-C. (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. https://doi.org/10.1080/0305764X.2015.1093095

Francis, B., Connolly, P., Archer, L., Hodgen, J., Mazenod, A., Pepper, D., Sloan, S., Taylor, B., Tereshchenko, A. and Travers, M.-C. (2017) 'Attainment Grouping as Self-Fulfilling Prophecy? A Mixed Methods Exploration of Self Confidence and Set Level Among Year 7 Students'. *International Journal of Educational Research,* 86, 96–108. https://doi.org/10.1016/j.ijer.2017.09.001

Francis, B., Craig, N., Archer, L., Hodgen, J., Mazenod, A., Taylor, B. and Tereshchenko, A. (2019) *'Teacher 'Quality' and Attainment Grouping: The Role of Within-School Teacher Deployment in Social and Educational Inequality'. Teaching and Teacher Education,* 77, 183–192. https://doi.org/10.1016/j.tate.2018.10.001

Francis, B., Craig, N., Hodgen, J., Taylor, B., Tereshchenko, A., Connolly, P., and Archer, L. (2020) 'The impact of tracking by attainment on pupil self-confidence over time: demonstrating the accumulative impact of self-fulfilling prophecy'. *British Journal of Sociology of Education*, *41*(5), 626–642. https://doi.org/10.1080/01425692.2020.1763162.

Graham, J.W., Olchowski, A.E. and Gilreath, T.D. (2007) 'How Many Imputations are Really Needed? Some Practical Clarifications of Multiple Imputation Theory'. *Prevention Science,* 8, 206–213. https://doi.org/10.1007/s11121-007-0070-9

Ho, D., Imai, K., King, G. and Stuart, E. (2011) 'MatchIt: Nonparametric Preprocessing for Parametric Causal Inference'. *Journal of Statistical Software,* 42: 8, 1–28. https://doi.org/10.18637/jss.v042.i08

Hodgen, J., Foster, C., Marks, R. and Brown, M. (2018) *'Evidence for Review of Mathematics Teaching: Improving Mathematics in Key Stages Two and Three'.* London: Education Endowment Foundation. Available at:

Hodgen, J., Adkins, M., Ainsworth, S. and Evans, S. (2019) *'Catch Up® Numeracy: Evaluation Report and Executive Summary'.* London: Education Endowment Foundation. Available at: https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/catch-up-numeracy (accessed 22 April 2026).

Hodgen, J., Foster, C., Coe, R., Brown, M., Higgins, S. and Küchemann, D. (2021) *'Low Attainment in Mathematics: An Investigation Focusing on Year 9 Students in England'.* London: UCL Institute of Education.

Hodgen, J., Bretscher, N., Hardman, M., Anders, J. and Lawson, H. (2023) *'SMART Spaces: Spaced Learning Revision Programme. Evaluation Report'.* London: Education Endowment Foundation. Available at: https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/smart-spaces (accessed 22 April 2026).

Hodgen, J., Foster, C., Brown, M. et al. (2023a) 'Low-Attaining Secondary School Mathematics Students' Perspectives on Recommended Teaching Strategies'. *International Journal of Science and Mathematics Education,* 22, 1325–1343. https://doi.org/10.1007/s10763-023-10420-8

Hodgen, J.; Taylor, B., Francis, B., Craig, N., Bretscher, N., Tereshchenko, A., Connolly, P. and Mazenod, A. (2023b) 'The Achievement Gap: The Impact of Between-Class Attainment Grouping on Pupil Attainment and Educational Equity Over Time'. *British Educational Research Journal,* 20: 2, 209–230. https://doi.org/10.1002/berj.3838

Hodgen, J., Taylor, B., Anders, J., Tereshchenko, A. and Jacques, L. (2024) *'The Student Grouping Study: Evaluation Study Plan'*. London: Education Endowment Foundation. Available at: https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/student-grouping-study (accessed 22 April 2026).

Hoffmann, T.C., Glasziou, P.P., Boutron, I., Milne, R., Perera, R., Moher, D., et al. (2014) 'Better Reporting of Interventions: Template for Intervention Description and Replication (TIDieR) Checklist and Guide'. *BMJ;* 348, g1687. https://doi.org/10.1136/bmj.g1687

Ireson, J., Hallam, S. and Hurley, C. (2005) 'What are the Effects of Ability Grouping on GCSE Attainment? *British Educational Research Journal,* 31:4, 443–458. https://doi.org/10.1080/01411920500148663

Imai, K., Keele, L. and Tingley, D. (2010) 'A General Approach to Causal Mediation Analysis'. *Psychological Methods,* 15, 4, 309–334. https://doi.org/10.1037/a0020761

Jackson, B. (1964) *'Streaming: An Education System in Miniature'*. London: Routledge and Kegan Paul.

Kurz, A. (2011) 'Access to What Should Be Taught and Will Be Tested: Students' Opportunity to Learn the Intended Curriculum'. In: S.N. Elliott, R.J. Kettler, P.A. Beddow and A. Kurz (Eds.), *'The Handbook of Accessible Achievement Tests for All Students: Bridging the Gaps Between Research, Practice, and Policy (pp. 99–129)'*. New York: Springer.

Mayring, P. (2019) 'Qualitative Content Analysis: Demarcation, Varieties, Developments'. *Forum Qualitative Sozialforschung Forum: Qualitative Social Research,* 20: 3. http://dx.doi.org/10.17169/fqs-20.3.3343

Maxwell, B., Stevens, A., Demack, S., Coldwell, M., Wolstenholme, C., Reaney-Wood, S., Stiell, B. and Lortie-Forgues, H. (2021) *'Review: EEF Implementation and Process Evaluation (IPE) Quality Pilot'*. London: Education Endowment Foundation. Available at: https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluation/eef-evaluation-reports-and-research-papers/syntheses-of-eef-evaluations/eef-implementation-and-process-evaluation-ipe-quality-pilot (accessed 22 April 2026).

Mazenod, A., Francis, B., Archer, L., Hodgen, J., Taylor, B., Tereshchenko, A. and Pepper, D. (2018) 'Nurturing Learning or Encouraging Dependency? Teacher Constructions of Students in Lower Attainment Groups in English Secondary Schools'. *Cambridge Journal of Education*, 49: 1, 53–68. https://doi.org/10.1080/0305764X.2018.1441372

Miles, M.B., Huberman, A.M. and Saldaña, J. (2019) *'Qualitative Data Analysis: A Methods Sourcebook (4th Edn.)'*. London: SAGE Publications Ltd.

Muijs, D. and 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. https://doi.org/10.1080/00131881.2010.524750

Noyes, A., Brignell, C., Jacques, L., Powell, J. and Adkins, M. (2023) *'The Mathematics Pipeline in England: Patterns, Interventions and Excellence'*. Nottingham: Centre for Research in Mathematics Education (CRME). Available at: https://nottingham-repository.worktribe.com/OutputFile/21372569 (accessed 22 April 2026).

Oakes, J. (1985) '*How schools structure inequality'*. New Haven: Yale University Press.

Observatory for Mathematical Education. (2025) *'Review of Mathematical Education 2025'*. Nottingham: University of Nottingham.

Organisation for Economic Co-operation and Development (OECD). (2013a) *'PISA 2012 Results: What Makes Schools Successful? Resources, Policies and Practices (Volume IV)'*. PISA, OECD Publishing. https://doi.org/10.1787/9789264201156-en

Organisation for Economic Co-operation and Development (OECD). (2013b) *'PISA 2012 Results: Ready to Learn (Volume III): Students' Engagement, Drive and Self-Beliefs'*. Paris: OECD Publishing. http://dx.doi.org/10.1787/9789264201170-en

Old, A. and Reddy, B. (2015) *'Mixed Ability Works Debate'*. YouTube. Available at: https://youtu.be/YUa_CI-stRo (accessed 22 April 2026).

R Core Team (2021) 'R: A language and environment for statistical computing. R Foundation for Statistical Computing'. Vienna, Austria. Available from: https://www.R-project.org/ (accessed 22 April 2026).

Reid, M., Clunies-Ross, L., Goacher, B., and Vile, C., (1981) '*Mixed-ability teaching: problems and possibilities'*. Windsor: NFER-Nelson.

Rosenbaum, P.R. (2015) 'Some Counterclaims Undermine Themselves in Observational Studies'. *Journal of the American Statistical Association,* 110: 512, 1389–1398. https://doi.org/10.1080/01621459.2015.1054489

Roy, P., Styles, B., Walker, M., Morrison, J., Nelson, J. and Kettlewell, K. (2018a) *'Best Practice in Grouping Students Intervention A: Best Practice in Setting Evaluation Report and Executive Summary'*. London: Education Endowment Foundation. Available at: https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/best-practice-in-grouping-students (accessed 22 April 2026).

Roy, P., Styles, B., Walker, M., Morrison, J., Nelson, J. and Kettlewell, K. (2018b) *'Best Practice in Grouping Students Intervention B: Best Practice in Mixed Attainment Evaluation Report and Executive Summary'*. London: Education Endowment Foundation. Available at: https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/best-practice-in-mixed-attainment-grouping (accessed 22 April 2026).

Rubin, D. (2004) *'Multiple Imputation for Nonresponse in Surveys'*. New York: John Wiley and Sons.

Rui, N. (2009) 'Four Decades of Research on the Effects of Detracking Reform: Where Do We Stand?—A Systematic Review of the Evidence'. *Journal of Evidence-Based Medicine,* 2, 164–183. https://doi.org/10.1111/j.1756-5391.2009.01032.x

Schoenfeld, A.H. and the Teaching for Robust Understanding Project. (2016) *'An Introduction to the Teaching for Robust Understanding (TRU) Framework'*. Berkeley, CA: Graduate School of Education. Available at: http://map.mathshell.org/trumath.php (accessed 22 April 2026).

Schoenfeld, A.H. (2018) 'Video Analyses for Research and Professional Development: The Teaching for Robust Understanding (TRU) Framework'. *ZDM Mathematics Education,* 50, 491–506. https://doi.org/10.1007/s11858-017-0908-y

Steenbergen-Hu, S., Makel, M.C. and Olszewski-Kubilius, P. (2016) 'What One Hundred Years of Research Says About the Effects of Ability Grouping and Acceleration on K–12 Students' Academic Achievement: Findings of Two Second-Order Meta-Analyses'. *Review of Educational Research,* 86: 4, 849–899. https://doi.org/10.3102/0034654316675417

Suter, L.E. (2017) 'How International Studies Contributed to Educational Theory and Methods Through Measurement of Opportunity to Learn Mathematics'. *Research in Comparative and International Education,* 12: 2, 174–197. https://doi.org/10.1177/1745499917711549

Taylor, B., Francis, B., Archer, L., Hodgen, J., Pepper, D., Tereshchenko, A. and Travers, M.-C. (2017) 'Factors Deterring Schools From Mixed Attainment Teaching Practice'. *Pedagogy, Culture and Society,* 25: 3, 327–345. https://doi.org/10.1080/14681366.2016.1256908

Taylor, B., Francis, B., Craig, N., Archer, L., Hodgen, J., Mazenod, A., Tereshchenko, A. and Pepper, D. (2018). Why Is It Difficult for Schools to Establish Equitable Practices in Allocating Students to Attainment 'Sets'? *British Journal of Educational Studies*, 67: 1, 5–24. https://doi.org/10.1080/00071005.2018.1424317

Taylor, B., Hodgen, J., Tereshchenko, A. and Gutiérrez, G. (2020) 'Attainment Grouping in English Secondary Schools: A National Survey of Current Practices'. *Research Papers in Education,* 37: 2, 199–220. https://doi.org/10.1080/02671522.2020.1836517

Taylor, B., Hodgen, J., Jacques, L., Tereshchenko, A., Cockerill, M. and Kwok, R.K.W. (2024) 'Access to Mathematics Learning for Lower Secondary Students in England During School Closures: Implications for Equity and Quality'. *Teachers and Teaching,* 30: 7–8, 971–985. https://doi.org/10.1080/13540602.2022.2062717

Tereshchenko, A., Francis, B., Archer, L., Hodgen, J., Mazenod, A., Taylor, B., Pepper, D. and Travers, M.-C. (2018) 'Learners' Attitudes to Mixed-Attainment Grouping: Examining the Views of Students of High, Middle and Low Attainment'. *Research Papers in Education,* 34: 4, 425–444. https://doi.org/10.1080/02671522.2018.1452962

Terrin, É. and Triventi, M. (2023) 'The Effect of School Tracking on Student Achievement and Inequality: A Meta-Analysis'. *Review of Educational Research,* 93: 2, 236–274. https://doi.org/10.3102/00346543221100850

Tingley, D., Yamamoto, T., Hirose, K., Keele, L. and Imai, K. (2014) 'Mediation: R Package for Causal Mediation Analysis'. *Journal of Statistical Software*, 59: 5, 1–38. Available at: www.jstatsoft.org/v59/i05/ (accessed 22 April 2026).

Weidmann, B. and Miratrix, L. (2021) 'Lurking Inferential Monsters? Quantifying Selection Bias in Evaluations of School Programs'. *Journal of Policy Analysis and Management,* 40: 3, 964–986. https://doi.org/10.1002/pam.22236

What Works Clearinghouse. (2022). *'What Works Clearinghouse: Procedures and Standards Handbook, Version 5.0'.* U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance (NCEE). Available at: https://ies.ed.gov/ncee/wwc/Handbooks (accessed 22 April 2026).

Appendix A: The EEF cost rating

Table A: Cost rating

Cost rating	Description
£ £ £ £ £	*Very low:* less than £80 per pupil per year.
£ £ £ £ £	*Low:* up to about £200 per pupil per year.
£ £ £ £ £	*Moderate:* up to about £700 per pupil per year.
£ £ £ £ £	*High:* up to £1,200 per pupil per year.
£ £ £ £ £	*Very high:* over £1,200 per pupil per year.

Appendix B: Security classification of evaluation findings

OUTCOME: GL Assessment Progress Test in Mathematics 13

Rating	Criteria for rating			Initial score		Adjust		Final score
	Design	**MDES**	**Attrition**					
5 🔒	Randomised design	<= 0.2	0-10%					
4 🔒	Design for comparison that considers some type of selection on unobservable characteristics (e.g. RDD, Diff-in-Diffs, Matched Diff-in-Diffs)	0.21 - 0.29	11-20%					
3 🔒	Design for comparison that considers selection on all relevant observable confounders (e.g. Matching or Regression Analysis with variables descriptive of the selection mechanism)	0.30 - 0.39	21-30%	3		Adjustment for threats to internal validity **[1]**		
2 🔒	Design for comparison that considers selection only on some relevant confounders	0.40 - 0.49	31-40%					2
1 🔒	Design for comparison that does not consider selection on any relevant confounders	0.50 - 0.59	41-50%					
0 🔒	No comparator	>=0.6	>50%					

Threats to validity	Risk rating	Comments
Threat 1: Confounding	Moderate/high	Matching was carried out using a broad range of observable variables. These variables were included in the model, and sensitivity and robustness analyses were undertaken. Evidence of imbalance on some variables remained after matching (e.g., Ofsted rating), though this was within acceptable bounds. However, there were other unmeasured differences that were not accounted for which may have influenced the outcome (e.g., pupil participation, teacher experience, teaching quality), particularly given the small number and atypical nature of the mixed attainment schools. At the school level, most of the initially identified matched comparator schools declined to participate, introducing a potential threat to validity. However, sensitivity analyses using student-level matching supported the main findings.
Threat 2: Concurrent Interventions	Low	No substantial evidence of risks. Schools did not change their teaching practice.
Threat 3: Experimental effects	Low	No substantial evidence of risks. Schools did not change their teaching practice, though they were aware of being part of this naturalistic experiment.
Threat 4: Implementation fidelity	Low	No substantial evidence of risks. Schools implemented their maths grouping practice as usual so there was no ideal implementation fidelity or compliance.
Threat 5: Missing Data	Moderate	There was a substantial level of missing student-level data (29.8%). Analyses that accounted for missing data produced results consistent with the complete-case analyses. Although some differential attrition was observed, this remained within acceptable thresholds. The impact of missing data on statistical power

		was reduced due to higher-than-expected level 2 correlation values and a lower-than-expected intraclass correlation coefficient (ICC).
Threat 6: Measurement of Outcomes	Low	The GL assessment is a validated and reliable outcome measure and there was no evidence of floor or ceiling effects. The collection of outcome data was supported by administrators in the majority of schools – 29 schools administered the tests themselves, but no evidence of bias was identified in the sensitivity analysis.
Threat 7: Selective reporting	Low	Clearly and comprehensively reported, following the planned analysis in the study plan and statistical analysis plan.

- **Initial padlock score:** 3 Padlocks – Matched design, powered to 0.20, 29.8% attrition.
- **Reason for adjustment for threats to validity**: 1 Padlock – Some moderate concerns about confounding and missing data at a magnitude that weakens the overall causal claim that can be made.
- **Final padlock score:** Initial score adjusted for threats to validity = 2 Padlocks.

Further appendices:

The further appendices can be found published on the EEF website.

Education Endowment Foundation

The Education Endowment Foundation
5th Floor, Millbank Tower,
21–24 Millbank,
London,
SW1P 4QP

https://educationendowmentfoundation.org.uk

@EducEndowFoundn

Facebook.com/EducEndowFoundn