The Student Grouping Study
Evaluation Study Plan
Evaluator (institution): IOE
Principal investigator(s): Professor Jeremy Hodgen and Dr Becky Taylor

Education Endowment Foundation

PROJECT TITLE	The Student Grouping Study: investigating the effects of setting and mixed attainment grouping
EVALUATOR (INSTITUTION)	IOE, UCL's Faculty of Education and Society
PRINCIPAL INVESTIGATOR(S)	Professor Jeremy Hodgen Dr Becky Taylor
STUDY PLAN AUTHOR(S)	Professor Jeremy Hodgen Dr Becky Taylor Dr Jake Anders Dr Antonina Tereshchenko Dr Laurie Jacques
STUDY DESIGN	Naturalistic study with matched design
PUPIL AGE RANGE AND KEY STAGE	11-13, KS3
NUMBER OF SCHOOLS	120
NUMBER OF PUPILS	20,000
PRIMARY OUTCOME MEASURE AND SOURCE	Maths attainment (GL Assessment PTM13)
SECONDARY OUTCOME MEASURE AND SOURCE	Self-confidence (self-report survey)

Study Plan version history

VERSION	DATE	REASON FOR REVISION
1.1 [*latest*]	24/07/2024	Standalone IPE plan integrated into main study plan. Updated implementation research questions; amended eligibility criteria; updated analysis of effects to align with statistical analysis plan (ver 1.0); further detail provided on the implementation analysis. Removed references to PCK in relation to impact evaluation.
1.0 [original]	1/11/2019	N/A

Table of Contents

Study Plan version history ... 1
The impact of the pandemic on the study ... 4
Grouping by ability and mixed attainment grouping ... 4
Study rationale and background ... 6
Theory of change ... 8
Research questions ... 9
Analysis of the effects of setting and mixed attainment grouping on student attainment and attitude ... 10
Plain English summary ... 11
Design overview ... 11
Participants ... 11
Outcomes and other data ... 13
Sample size calculations ... 15
Selection of the comparison group and identification assumptions ... 16
Recruitment strategy ... **Error! Bookmark not defined.**
Imbalance between groups ... 16
Primary analysis of the effects of setting and mixed attainment grouping on student attainment ... 17
Robustness checks ... 18
Further analyses ... 18
Analysis of the implementation of setting and mixed attainment ... 19
Theoretical framework ... 20
Design ... 22
Sampling ... 23
    Random selection of case study schools ... 23
Data collection ... 24
    Year 7 (May-July 2023) ... 25
    Year 8 (November 2023 – March 2024) ... 26
Data analysis ... 31
    Heads of Mathematics interviews ... 32
    Mathematics schemes of work ... 32
Cost evaluation ... 35
Ethics ... 35
Data protection ... 35
    Data security ... 36
Personnel ... 37
Steering Board ... 38
Risks ... 38
Timeline ... 39

References 41
Acknowledgments.. 46
APPENDIX A: Review process 47
APPENDIX B: Grouping Study Research Design Options... 48
APPENDIX C: Logic models 51
APPENDIX D: Matching Factors: what factors influence (or are correlated with) school decisions to set/mix?.. 53
APPENDIX E: Additional power calculations.. 58
APPENDIX F: Grouping Study Matching and Simulated Response Exercise...................... 59
APPENDIX G: Steering Board Terms of Reference 82
APPENDIX H: Head of Mathematics Survey... 83
APPENDIX I: TRU Scoring rubric for case study field visits (an extract)............................. 86
APPENDIX J: Student paired interview after the observed lesson...................................... 88
APPENDIX K: Teacher post lesson observation interview (mixed attainment) 91
APPENDIX L: Focus group schedule... 94
Appendix M: Head of Mathematics interview.. 99

The impact of the pandemic on the study

The study began with Year 7 pupils in the academic year of 2019/2020, but was paused in March 2020 because of the COVID-19 pandemic. During the period of lockdowns and school closures, grouping arrangements were disrupted in many schools. On reopening, some schools had changed their practices, or no longer had capacity to participate in the study.

The study restarted in 2022 with a different group of schools and with students in Year 7 in 2022/23.

Grouping by ability and mixed attainment grouping

The study uses a matched design in a natural context, to explore the difference in student outcomes of two approaches to grouping students: grouping by subject ability (or setting), and mixed attainment grouping. As such, the research team will not be delivering an 'intervention', but will be measuring the outcomes of grouping practices already in use in recruited schools. Our description of the practices being compared follows the TIDieR[1] framework.

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: Our review of the literature (Francis et al., 2017) indicated that when students are taught in attainment sets, students in lower-attaining groups make less progress compared with their higher-attaining peers (see also the EEF Teaching and Learning Toolkit, Higgins et al., 2018). However, this effect is very small and is not consistently demonstrated across meta-analyses of experimental research evidence (see, e.g., Steenbergen-Hu et al., 2017, for an overview). In addition, much of the research is not directly relevant to the current context of England because it is either very dated or was conducted in the US.

From the literature and from our own research (Taylor et al., 2018a, 2018b), we have established that the grouping strategy '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 literature review showed that students in lower-attaining sets tend to be allocated teachers of lower quality (and see Francis et al., 2018) 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). Thus, they have less opportunity to learn (see, e.g., Suter, 2017, for a discussion of opportunity to learn). However, Dunne et al.'s (2007, 2011) study of the teaching and learning of students low-attaining sets in English secondary schools suggests that some schools take active steps to

---

[1] http://www.bmj.com/content/348/bmj.g1687

avoid and additionally to mitigate against these effects by, for example, reducing class sizes in lower sets.

We conjecture that opportunity to learn, 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, 2018). We hypothesise that these outcomes are also likely to impact on students' orientation to future participation (cf. Archer et al., 2012).

We conjecture that moderating factors are likely to include those relating to the student (socioeconomic status, ethnicity and sex), the school (characteristics of the whole intake including prior attainment, leadership ethos, and resources) and the teacher (beliefs and attitudes).

This is summarised in a Logic Model (Appendix C: Figure C1).

Who/Where: As with our prior study, Best Practice in Grouping Students, we intend to focus on Year 7 and Year 8 pupils in English state secondary and middle schools. For this study we will focus on mathematics grouping and teaching. Year 7 and Year 8 have been chosen as the first two years of secondary education, and so for the majority of pupils a new phase in their education. Mathematics has been 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.

What/How: Schools will be recruited according to their usual grouping practices in mathematics and will continue to teach students using their usual resources and strategies. Since we propose to compare schools with established practices, random allocation will not be possible. Hence, the schools in this study will be recruited so that, as far as possible, the two groups of schools will be matched on a range of characteristics aside from their grouping practices (grouping by attainment, or mixed-attainment grouping).

We conjecture schools' decisions to set or mix will 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 (Appendix C: Figure C.2) and analysis of which factors influence schools' decisions around grouping practices (Appendix D).

There is likely to be variation between schools in the implementation of grouping, for example mixed attainment schools may choose to have a nurture group, and setting schools will vary in how they allocate students to groups, the fluidity of student movement between groups and the number and distribution of set 'levels'.

When and how much: Variation in the experiences of pupils with differing levels of prior attainment will be explored thoroughly through the implementation analysis, with particular attention to pupils with low prior attainment and those from disadvantaged backgrounds.

Tailoring: Variations in grouping practices and pedagogy within and between schools will be explored thoroughly in the implementation analysis. In particular we will explore variation in teacher quality and in opportunity to learn (OTL).

Study rationale and background

The question of whether setting or mixed attainment grouping is a more effective strategy for grouping students for teaching in secondary schools is currently a hot and contested topic in England. 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). Whilst 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, Wiliam & Brown, 2000) and attainment (Rui, 2009) and research from the US 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 (Delisile, 2015). Yet, many practitioners are steadfastly and vociferously committed to one approach or the other (e.g. Old & Reddy, 2015), while others are actively seeking advice about attainment grouping, wanting to use the best, evidence-informed practice.

Additionally, much of the existing research was conducted in the US 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 is urgent need for such research. This study will examine 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 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. This study 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 Best Practice in Setting trial found no effect for the intervention compared to a 'business-as-usual' control group (Roy et al., 2018). Our mixed methods process evaluation (largely conducted by the project team) has 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 have found that practical issues such as timetabling impede 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 have 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 learnt 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 cannot 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[2] (Hodgen et al., forthcoming; Taylor et al., forthcoming).

Hence, while these two prior trials have resolved several outstanding questions in the literature, has provided new findings about why setting negatively impacts low attainers, and has identified why this situation is unlikely to improve in spite of good intentions/evidence-informed interventions designed to do so, it was not designed to compare mixed attainment practice with 'ability grouping'. This leaves 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 will:

Provide direct and robust evidence relevant to schools in England comparing the effects of setting and mixed attainment teaching. As previously stated, much of the existing research was conducted in the US 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 is urgent need for such research. The Student Grouping Study will examine 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.

Provide evidence of the effects of setting and mixed attainment teaching relative to other approaches for addressing low attainment or disadvantage. A 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). It is vital that schools have reliable evidence to inform decisions about supporting low attainers and other disadvantaged students. This will be particularly important if the study finds little or no differences between the two approaches on student attainment. Our primary focus in this study is on the experiences of low attaining students. However, in doing this, we will compare their experiences to those of high attaining students.

Provide detailed evidence of how grouping practices are implemented. The Student Grouping Study will provide detailed evidence of contextual influences that shape grouping practices and how attainment grouping in mathematics is implemented in different contexts. As a result, the study will characterise effective and equitable practices 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 will include developing a measure of 'opportunity to learn' mathematics (OTL). In addition, findings regarding implementation of grouping are 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. There is particularly vigorous debate about the impact of attainment grouping in mathematics, where setting is the most prevalent practice. However, the context of mathematics will ensure that the project has the greatest reach and influence, because the findings will be considered by schools as more directly relevant to a wider range of subjects, such as modern foreign languages and science.

Theory of change

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.

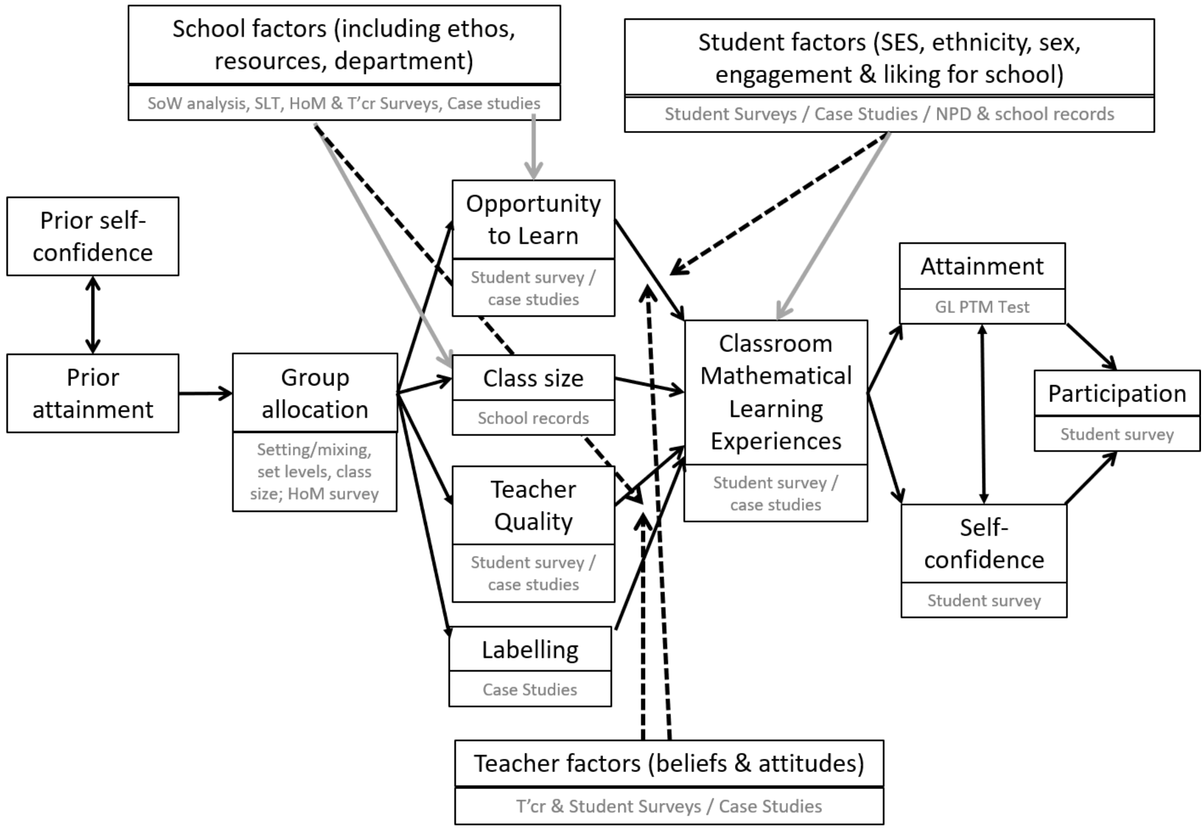
By contrast, setting is the practice of grouping students for teaching by their prior attainment in the subject being taught, with the intention that within each teaching group, students have a similar, narrow, range of prior attainment.

We anticipate that mixed attainment grouping and setting will result in different outputs, in terms of opportunity to learn mathematics (OTL), teacher quality, class size and labelling.

These outputs 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 expect the long-term outcome of participation in mathematics to be affected. We anticipate that students in schools that use mixed attainment grouping will 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. Students in schools that use setting are likely to have different experiences of learning mathematics, based on their prior attainment. As outlined above, research suggests that students in high-attaining sets will be taught a richer curriculum, have access to better teaching and will benefit from positive labels. They are also likely to be taught in larger classes. Students in low-attaining sets are likely to be taught a more restricted curriculum, have access to restricted pedagogies and less good teaching and experience negative labelling. They will be likely to be taught in smaller classes but with access to fewer mathematical role models.

The outputs and short-term outcomes will be moderated by school, student and teacher factors. School factors include the school's ethos or vision, mathematics department goals, and resources. For example, 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 are more likely to be placed in lower-attaining sets. 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



Research questions

Analysis of the effects of setting and mixed attainment grouping on student attainment and attitude:

1. 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.

2. 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 receiving FSM.
3. 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.
4. 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.
5. To what extent do (a) opportunity to learn and (b) teacher quality explain any differential outcomes for different grouping practices, for different set allocations or for students at different attainment levels?

Analysis of the implementation of setting and mixed attainment:

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 Year 7 and Year 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 will provide further information to supplement the findings of research question 5 by examining the factors that shape opportunity to learn 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 will be 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 KS2 (as per the DfE NPD). High attaining students will be defined as the highest attaining tertile.

Analysis of the effects of setting and mixed attainment grouping on student attainment and attitude

In this section, we outline the study design. A number of alternative design options were considered for this study. The advantages and disadvantages of the different designs are summarised in Appendix B.

Plain English summary

Setting by attainment is practised in England to a varying degree and varies according to subject and year group. However, there is limited up-to-date and UK-based evidence about how this practice is carried out by schools, or how choices about student grouping alters outcomes for pupils.

Schools that currently practice mixed attainment grouping in Years 7 and 8 for mathematics will be recruited to the study. These schools will then be matched with schools that practice setting by attainment to provide a comparison group of schools. Differences in outcomes in mathematics for school across these two groups will be looked at, including attainment differences and pupil self-confidence. In total 120 schools will be recruited to the study: 40 schools using mixed attainment schools, and 80 schools using grouping by attainment, or setting.

The study will be a 'naturalistic experiment' comparing existing grouping practices in schools rather than implementing a specific programme. The project will use a matched design in a naturalistic context to investigate the differences in attainment and self-confidence in mathematics of students taught in mixed attainment groups compared to pupils who are taught in sets.

The team at IOE, with oversight from an independent steering board, will be carrying out this research. The project will look at attainment using a standardised test in maths at the end of Year 8. The research will also look at outcomes for FSM students and low-prior attainers. The IOE team will also carry out an implementation analysis to understand the diversity of grouping practices within schools. See acknowledgements for the membership of the steering board.

Design overview

Table 1. Research design.

Design type		
Unit of analysis (school, pupils)		Students clustered in schools
Number of Units to be included in analysis		120 schools (40 mixed attainment schools, 80 matched setting schools)
Outcomes	primary	Attainment in mathematics
	secondary	Self-confidence in mathematics
Outcome sources (instruments, datasets)	primary	GL Assessment Progress Test in mathematics
	secondary	Survey developed by the research team

Participants

As noted earlier, the study initially began with Year 7 pupils in the academic year of 2019/20. Schools were recruited between February and September 2019 and two matched groups of schools were established: Mixed Attainment (33 schools) and Setted (82 schools). This iteration of the study was paused in March 2020 because of the COVID-19 pandemic. During the pandemic, many schools changed their grouping practices[3] and some schools indicated that they no longer had the capacity to take part in the study. Hence, a further recruitment and matching process was carried out between January and October 2022.

Participants will be all students starting Year 7 in September 2022 in recruited schools. They will participate in the study until they finish Year 8, in July 2024. Students in the mixed attainment group will then have received two full years of the 'treatment', i.e. mixed attainment grouping in mathematics.

Two groups of schools have been recruited: one group that is already teaching mathematics to Year 7 and Year 8 students in attainment sets, and one group that is already teaching mathematics to Year 7 and Year 8 students in mixed-attainment groups.

Schools were eligible to participate in the mixed attainment 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 may additionally have a 'nurture group', in which the very lowest attaining students are taught separately.[4]

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 that use 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.

Werecruited matched schools, with a ratio of 1:2 Mixed Attainment to Setting schools (see sample size calculations below). The 'quasi-randomisation date' was 16 June 2022, the date at which recruitment to the mixed attainment group was complete and the matching procedure to identify matched setting schools was carried out.

Our initial eligibility criteria (see Hodgen et al., 2019) included free schools and middle schools. However, during the matching process a number of mainly free schools did not have sufficient historic attainment data available for the matching process and were excluded from the study. Additionally, in the initial matching simulation exercise, it was not possible to identify any sufficiently close matches for most of the middle schools, largely due to the small number of these schools. Hence, middle schools were also excluded from the study.

---

[3] Taylor, B., Hodgen, J., Jacques, L., Tereshchenko, A., Cockerill, M., & Kwok, R. K. W. Access to mathematics learning for lower secondary students in England during school closures: implications for equity and quality. *Teachers and Teaching*, 1-15. https://doi.org/10.1080/13540602.2022.2062717

[4] Schools may operate a 'nurture group' and still be included in the mixed attainment group, provided that this is 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 & Whitebread, 2007; Mazenod et al., 2018).

Outcomes and other data

Primary outcome measures

The primary outcome measure will be attainment in mathematics, measured using the paper version of the GL Assessment Progress Test Mathematics (PTM). This will allow comparison with the two trials in our previous study. In addition, PTM is very well matched to the current mathematics curriculum and it may be possible to create sub-scales for different mathematical topics. However, we note that in some previous trials, there has been evidence of floor and ceiling effects associated with the PTM, particularly for the standardised scores.

Testing would be carried out in schools at the end of the second year of the study (students at the end of Year 8).

In order to avoid attrition, and consequent bias, collection of the attainment outcome data will be supported by support staff, who will ensure that schools arrange to participate in outcome testing and that the data is collected, visiting schools where necessary. A financial incentive of £1000 will also be offered to both study groups, payable at the end of the second year of the study when data collection is complete.

In our analysis we will control for prior attainment (Key Stage 2 test outcomes in mathematics, obtained from the NPD).

Secondary outcome measures

We propose to use mathematics self-confidence survey items[5] developed for the Best Practice in Grouping Students project as the secondary outcome measure. Additionally, we will use general self-confidence survey items and adapt items measuring students' orientation towards further participation as additional secondary outcomes. We will administer the survey at two points: as part of a short machine-read survey close to the start of the study (students in Year 7) and again at the end of Year 8 as part of a longer online survey. In our analysis we will control for baseline self-confidence.

Other data Mediator variables

Teacher quality: There is strong evidence that teacher knowledge is a key indicator of teacher quality (e.g., Baumert et al., 2010: Coe et al, 2015; Hill et al., 2005) However, indicators such as highest qualification achieved have proved too coarse-grained as measures of teacher quality (Shulman, 1986). Hence, 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, which demonstrated that pedagogical content knowledge (or knowledge of mathematics tasks, students' mathematical learning and instructional strategies in mathematics) was a key factor in understanding teacher effectiveness (Baumert et al., 2010). A key challenge was to reduce the length of the instrument in order to ensure that any surveys do not introduce additional burdens on teachers and, in addition, reduce response rates. A measure was developed using items from the CEMENT survey (developed by Merrilyn Goos and Kim Beswick and validated with Australian secondary teachers, Goos, 2013) and the Teacher Education and Development Study in Mathematics (TEDS-M) survey (from the international comparative study of mathematics teachers).[6] However, the results of the validation process were not promising. There were strong indications of multi-dimensionality that did not reflect the Baumert et al. (2010) theoretical model of teacher knowledge that we are using. Specifically, the factors did not align clearly with mathematics content knowledge (MCK) and pedagogical content knowledge (PCK). As an alternative, we will use two

[5] See Francis et al. (2017) for further information about the self-confidence survey items.
[6] See https://www.iea.nl/studies/iea/teds-m

dimensions of the teacher quality student survey developed and validated by Evidence-Based Education[7]: *understanding the content*, and *activating hard thinking.* The Evidence-Based Education student survey is based on dimensions from the Great Teaching Toolkit evidence review. Understanding the content addresses teacher knowledge and understanding of curriculum content, sequencing, pedagogies and common student misconceptions. Activating hard thinking addresses structuring learning, explanation, questioning, interactions with students and embedding and activating learning.

Opportunity to Learn (OTL): We will use OTL to measure the extent to which students are offered the full curriculum. OTL is defined simply as the time allocated for learning different topics (Carrol, 1963). This has been used to compare the effects of curriculum exposure in research from the US (e.g., Schwartz, 1995) and international surveys, such as TIMSS and PISA (e.g., Suter, 2017; Schmidt & Burroughs, 2016, Burstein, 1992). However, developing and validating a robust measure of OTL is a challenge, particularly a measure that is relatively short and easy to complete (Suter, 2017). There is evidence that classroom activity can be measured using a relatively short rating scale-based survey (Nitz et al, 2014). We will measure OTL through the student and teacher surveys, which will each be validated through at least two, and possibly three, pilot administrations in addition to cognitive interviewing. The instrument will be used to measure within-class, as well as between-class, variation in OTL in order to compare OTL at pupil-level across the two groups of schools.

Class size: We will collect data from schools on students' class, teacher and class size from schools. These data will be collected close to the start of the study and after all students have been allocated to classes for mathematics.[8]

Student attitudes: We will explore the possibility of collecting additional data through a student survey administered to a sub-sample of students within schools to measure student liking for school and level of engagement. We anticipate that student attitudes to setting and mixed attainment grouping may also influence their engagement and measures of these would also be included in the student survey.

Other data: Potential moderator variables

As noted above, we conjecture that moderating factors are likely to include those relating to the student (SES (FSM), ethnicity and sex), the school (characteristics of the whole intake including prior attainment, leadership ethos, and resources) and the teacher (beliefs and attitudes). We note, however, that the clustered structure of our sample will limit the factors that we can investigate using a quantitative moderation analysis and some of these factors will be investigated either using sub-group analysis and/or a combination of descriptive and case study data (school ethos and resources, teacher beliefs and attitudes).

We will adapt and validate research instruments developed for our prior study Best Practice in Grouping Students to create scales to measure student liking for school and teacher beliefs (including attitudes to setting and mixed attainment grouping, beliefs about students and expectations).

Other data: Additional survey data for the implementation analysis

Our existing survey instruments include scales for measuring:

- Student perceptions of how grouping is practised in their school

[7] Evidence-Based Education. (2022). *Great Teaching Toolkit: Student Surveys*. Evidence-Based Education. https://evidencebased.education/great-teaching-toolkit-cpd/

[8] Some schools use the first few weeks of Year 7 to assess students before class allocation.

- Teachers' pedagogical practices

We will review and revalidate these scales, in order to achieve a survey that is not overly burdensome to complete. We also intend to develop and validate new scales for student attitudes to mixed attainment grouping, and students' opportunity to learn in mathematics.

Our student focus group and teacher interview schedules explore the above issues in greater depth, probing the explanations behind identified patterns. Our classroom observational work will record the pedagogic practices and behaviours to which students are subject, and the classroom dynamics precipitated, as well as verifying or otherwise practices articulated by interview respondents. We will use the qualitative data collection in the implementation analysis to develop further our logic model in order to improve our explanation of the mechanisms underlying the different grouping practices, and to provide a better understanding of how to overcome the challenges of implementing more equitable practices.

Sample size calculations

MDES calculations were carried out using the R package PowerUpR. Initial calculations were based on a 2-level model with students clustered within schools. We note that the power calculations for the planned and actual recruited samples reported in the Statistical Analysis Plan, also using PowerUpR, were based on a 3-level model (students clustered within schools and schools clustered within matched groups) because this was judged to better reflect the actual design and better estimates of the MDES.

We show the MDES calculation assuming a pre-post test correlation of 0.75 (at the student-level) in Table 2. We expect the pre-post test correlation between the KS2 and PTM to be at least 0.5 and, hence, believe a correlation of 0.75 is achievable if we include additional covariates in the model alongside the KS2 score. We follow the practice currently favoured by the EEF of assuming the school-level correlation to be 50% of the student-level, although we consider this assumption to be conservative. MDES calculations are also shown for FSM and low prior attainment sub-groups. See the statistical analysis plan for further information.

We calculated a range of estimates based on different samples and group sizes and two levels of pre-test to post-test correlations (see Appendix E).

Table 2. MDES

		Study Plan	
		OVERALL	FSM / Low prior attainment
MDES		0.199	0.207
Pre-test/ post-test correlations	level 1 (pupil)	0.75	0.75
	level 2 (school)	0.38	0.38
Intracluster correlations (ICCs)	level 2 (school)	0.15	0.15
Alpha		0.05	0.05
Power		0.8	0.8
One-sided or two-sided?		Two-sided	Two-sided
Average cluster size		100	25
Number of schools	Intervention	40	40
	Comparison	80	80
	Total	120	120

Number of pupils	Intervention	4000	1000
	Comparison	8000	2000
	total	12000	3000

Selection of the comparison group and identification assumptions

We recruited schools to the mixed attainment group first. We estimated that there are between 120 and 190 eligible schools in England teaching mathematics in mixed attainment groups to Year 7 and Year 8. 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 identify schools, and cold-calling schools to ask about grouping practices. We anticipated that recruitment might be difficult. After mixed attainment schools were recruited, we undertook to identify a group of setted schools matched to each mixed attainment school. Details of the matching are described in detail in the statistical analysis plan.

The study initially recruited schools between February and September 2019. Two matched groups of schools were recruited: Mixed Attainment (33 schools) and Setted (82 schools). The study then began with Year 7 pupils in the academic year of 2019/2020, but was paused in March 2020 because of the COVID-19 pandemic. During the pandemic, many schools changed their grouping practices and some schools indicated that they no longer had the capacity to take part in the study. Hence, a further recruitment and matching process was carried out during 2022. This second, and final, recruitment exercise is described in detail in the statistical analysis plan).

Imbalance between groups

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 will test balances between groups on a fixed set of variables for means, medians, standard deviations, and skewness at both school- and student-level. We will report the comparison of means in a similar way to that required for a standard EEF RCT trial with standardized differences using Glass's delta (arithmetically the same, but conceptually different to effect sizes in this setting). Unstandardised differences in means, medians, standard deviations and skewness will also be reported. We will also plot overlapping kernel density plots of these characteristics between the treatment and matched comparison groups to give an overall impression of the different distributions. This will be done for the following variables:

- Prior attainment (KS2), current and historical;
- % FSM;
- % of low and high attainers;
- School size;
- % FSM;
- % EAL;
- Academy status;
- % at IDACI quintiles;
- OFSTED rating;
- % urban.

Comparing our treated sample with the following samples:

- English schools;
- Pool of potential comparators identified by matching;
- Recruited comparison sample.

Primary analysis of the effects of setting and mixed attainment grouping on student attainment

Given the exploratory nature of this naturalistic study we do not consider it appropriate to adopt the EEF's "standard" approach, a 2-level multilevel model, because it relies on strong assumptions that may not hold within the PSM framework. Rather, using a weighted pupil-level sample, we will estimate the following linear regression model to address Research Question 2:

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

where y is the outcome variable for pupil 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 pupil- 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. Standard errors will be 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 vector of pupil- and school-level control variables ($\boldsymbol{X_{ij}}$) will consist of the following covariates:

- Individual KS2 prior attainment
- School average KS2 attainment of intake
- School low prior attainment proportion
- School high prior attainment proportion
- School cohort number of pupils
- Individual FSM
- School FSM proportion
- Individual EAL status
- School EAL proportion
- School academy status
- Individual IDACI
- School composition IDACI quintile
- Ofsted grade
- Urban/rural classification

The analysis will be on an intention to treat basis analogous to that used for a standard RCT. We will specify a date ("quasi-randomisation date") by which schools and students will be included in the sample and pre-trial data is to be collected. This will be done once we have recruited treated schools, finalised our pool of 10 matched comparators for each school, and recruited the two comparators within this. Schools, and students, will then be included in the analysis whether or not they subsequently change their grouping practices and whether or not pupils move between schools.

Weights will be applied to reflect the number of matched comparator schools recruited corresponding to each mixed attainment school. In order to explore potential distributional changes, we will adapt the primary analysis model, using the same weighted pupil-level sample.

Anders et al.'s (2017) *Complex Whole-School Interventions* report suggests a 'difference-in-differences' approach in certain cases. Because this is a naturalistic study, we do not consider this appropriate, because outcomes in previous years, such as GCSE, Progress 8 or other valued added measures, will have been directly influenced by schools' grouping practices.

Please see the statistical analysis plan for a detailed discussion.

Robustness checks

We propose to carry out a range of robustness checks as part of our matching exercise as follows: We will ensure that our robustness check specifications are themselves robust prior to conducting the checks on the results of matching. These will include (1) checks on the approach to matching that might lead us to adjust our approach to this, and (2) checks on how well our recruited sample reflects our matched sample, and in order to understand, and potentially adjust for, any problems induced by the recruitment process.

We will conduct these checks at a school-level after matching has been carried out, and later, for (2), also at a student-level once we have accessed the NPD (but prior to matching in the outcomes data).

In addition, we will conduct checks to assess the robustness of our primary model, including multiple imputation (if appropriate, see missing data below).

Further analyses

Secondary outcome analyses

The secondary outcome analyses will use the same approach to modelling as in the main analysis.

Quantile regression analyses

As a robustness check on the analysis of distributional change, we will use quantile regression to explore distributional changes in performance associated with being in a school with setting. We will specify models analogous to our primary analysis for the following points of the outcome distribution: the 25th, 50th (median), and 75th percentile.

Subgroup analyses

We will conduct sub-group analyses for FSM students and for low attaining students. In line with EEF's guidance for sub-group analysis for RCTs, we will first estimate a model on the full sample adding a covariate for our sub-group of interest and an interaction between this covariate and the treatment indicator. We will then run a separate sub-group analysis identical to the primary analysis model on the sample defined as falling in the sub-group. This will be done separately for each sub-group (FSM and low attainment). We will additionally examine the model dependence of these effect size estimates by presenting the range of estimates across different models.

Treatment effects in the presence of non-compliance

Our approach to compliance will be informed by the implementation analysis using further sub-group analyses to examine how differences in compliance influence outcomes. Whatever the school-level indicator of compliance chosen, we expect to explore compliance using a simple sub-group analysis model, rather than a CACE/instrumental variables analysis. We do this because we do not believe the assumptions of these approaches necessarily generalise to a matched sample analysis. Our approach to compliance is set out in detail in the statistical analysis plan.

Mediation analyses

We intend to conduct mediation analysis using two mediators identified in the logic model: opportunity to learn (OTL), and teacher quality.

This will follow the principles of mediation analysis set out by Baron and Kenny (1986) with appropriate significance tests, as described by Sobel (1986), which are well recognised in the literature. That said, we are aware of limitations of this approach and will explore promising alternatives, such as that proposed by Imai et al. (2010; see also Hayes, 2013; Preacher, 2015). We propose to pre-specify one ‘primary’ mediation analysis to explore the importance of differences in opportunity to learn and teacher quality (measurement of which is discussed further elsewhere). The model is specified in detail in the statistical analysis plan.

Missing data

We will adopt an approach that we have used previously (see Anders and Shure, 2018), which are fully specified in the statistical analysis plan. This will describe and summarise the extent of missing data in the primary and secondary outcomes, and in the model associated with the analysis. Reasons for missing data will also be described. For all models we will implement a missing data strategy if more than 10% of data in the model is missing.

Effect size calculation

Effect sizes will be calculated using the Hedges’ G ES for cluster randomised trials as per the current EEF (2022) statistical analysis guidance for evaluations (and based on the primary ITT analysis and conditional on the covariates in this model). This, and the calculation of confidence intervals, are fully specified in the statistical analysis plan.

Analysis of the implementation of setting and mixed attainment

As outlined above, the study takes place in a natural context and compares two matched groups of schools that implement contrasting approaches to grouping students for mathematics lessons: mixed attainment and setting. In contrast to a randomised controlled trial, the study does not aim to evaluate the impact and implementation of a well-described intervention. Rather, the aim of our IPE is to investigate how these broad approaches are understood and implemented across schools and classes. Our purpose is 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 current EEF 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 is a natural experiment and seeks to understand the different ways in which schools implement either ‘mixed attainment grouping’ or ‘setting’. Hence, a focus on fidelity of implementation is not appropriate. We have therefore adapted the EEF IPE dimensions as described below.

In this study, we will examine 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, (RQ6), 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 opportunity to learn (RQ7, RQ5);

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 will include consideration of set movement (RQ8, RQ5);

Contextual factors: factors that explain any differences in grouping practice and how they are implemented (RQ9). This will include whether there are any factors that especially enable 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 will, together with RQ5 of the effects analysis, examine the extent to which the hypothesised logic model provides a plausible explanation of the processes by which the different grouping approaches impact on student mathematical outcomes and whether, and how, this logic model should be adapted.

Theoretical framework

The purpose of this research is to document and compare teachers’ pedagogy and students’ mathematical learning experiences when they are grouped in mixed attainment or setting classrooms, as well as to explore the factors that influence these processes. We will focus on students with lower prior attainment by contrasting their experiences with high prior attaining students (see Hodgen et al., 2023a; Hodgen et al., 2023b).

The theory of change that underpins this study is outlined above and the logic model is found in Figure 1 and Appendix C. Central to this model and to this IPE is the concept of students’ mathematical learning experiences and how teaching quality and opportunity to learn affect this. Our understanding of classroom mathematical learning experiences is underpinned by the ‘Teaching for Robust Understanding framework (TRU Math) (Schoenfeld and the Teaching for Robust Understanding Project, 2016), which allows us to evaluate the quality of a mathematics learning environment from the students’ perspective. The framework enables 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 is clear, correct, and well justified (tied to conceptual underpinnings).
- Cognitive Demand: The extent to which students' sense-making and 'productive struggle' is promoted and they are engaged in classroom interactions that create and maintain an environment of intellectual challenge.
- Access: The extent to which classroom activity structures invite and support active engagement for a diverse range of students so that they experience meaningful and equitable access to concepts and mathematical practices.
- Agency, Ownership and Identity: The extent to which students' experiences support the construction of positive disciplinary identities e.g. 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 is elicited, challenged, and refined and how the classroom environment is responsive to student thinking.

Opportunity to learn (OTL) generally refers to educational inputs and instructional processes within a school setting necessary for helping students to achieve the intended outcomes. Kurz (2011) provided a framework for OTL based on time on instruction, instructional content, and quality of instruction. The OTL instrument developed for the effects analysis focused on the first two aspects: time and content. In the IPE we extend this 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 are 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 will draw on the 12 evidence-based strategies and approaches identified by Hodgen et al. (2020) as having the potential to improve teaching and learning of mathematics for low-attaining pupils when used in a balanced manner. These strategies are summarised in Table 3. We note that, for each strategy, the manner in which the strategy is implemented is crucial to its likely effectiveness. Hence, we will use the TRU in order to differentiate the quality of how these strategies are implemented.

Table 3. Twelve evidence-based strategies for improving teaching and learning for low-attaining pupils (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 (CAI)	Computer-based systems designed to supplement learning (e.g., Sparx Maths, Heggarty Maths).

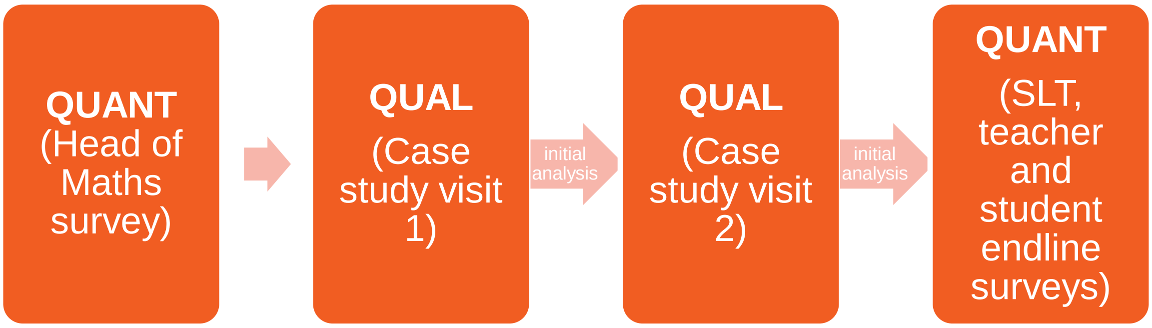
Peer tutoring	Tutoring by same-age peers.
Heuristics	Explicit strategies for approaching, and solving, a range of different problems.
Manipulatives	Concrete materials that can be manipulated by pupils to aid understanding and to support the development of mathematical understanding, guided by explicit teaching.
Tutoring by adults	Group or one-to-one additional support provided by teachers or teaching assistants (TAs).
Feedback to pupils	Information provided to a student regarding performance or understanding.
Representations	Diagrams, graphs and tools such as number lines, including concrete-pictorial-abstract [CPA] 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 anticipate that classroom mathematical experiences will be influenced both by class size and by the ability-labelling of pupils (Francis et al., 2017; Tereshchenko et al., 2018). Class size will be recorded as part of case studies, and will also be collected with pupil data provided by schools. Labelling will be explored through teacher interviews and student interviews and focus groups as part of case studies.

Design

The IPE has a sequential mixed-methods design with quantitative surveys and in-depth qualitative case studies. The overall design is informed by the logic model and the theoretical framework and is summarised in Figure 2.

Figure 2. Summary of mixed-methods IPE design.



Sampling

Surveys

Surveys will be sent to all those in participating schools who have not been withdrawn from data processing for the purposes of this study:

- Student surveys will be administered to all Year 8 students who have not been withdrawn from data processing by their parent/carer.
- Head of mathematics, mathematics teacher and senior leader surveys will be sent to all relevant staff in each participating school.

Case studies

The case study sample will include twelve schools, including six schools using setting and six using mixed attainment grouping.

As outlined in the study plan (Hodgen et al., 2019) schools completing all aspects of data collection will receive £1000, with an additional £500 for case study schools.

Random selection of case study schools

Following Maxwell et al.'s (2021) assessment of high-quality sampling for qualitative work in IPEs, we propose to use a random sampling approach. Our random case study school sample will be stratified by prior attainment in line with evidence that grouping practices are associated with prior attainment of pupils in secondary schools in England (see Taylor et al., 2020). Further, purposive sampling is not appropriate as there is very limited research indicating how high quality grouping practices are associated with particular school characteristics. A school level 'prior attainment variable' will first be produced, based on the average Key Stage 2 point score for students from each school who sat GCSE examinations in summer 2019 and for whom data was therefore available. Schools will then be ranked by prior attainment and divided into two equally-sized strata. The characteristics of the strata for the achieved sample are described in Table 4.

Table 4. Number of schools by grouping practices, stratified by high and lower prior attainment

	Total schools	Prior attainment variable range **Higher** stratum	No of schools	Prior attainment variable range **Lower** stratum	No of schools

Mixed	30	30.4-29.0	15	28.8-27.0	15
Setting	80	31.5 - 29.3	40	29.2 - 25.7	40

Randomisation will be conducted using the Data Analysis tool pack in MS Excel. To enable replication, random seed values have been generated using the RAND function (see Table 5). Using these seed values, four sets of random numbers will be generated from a uniform distribution. These sets of randomly-generated numbers will then be assigned to the schools in each of the four strata. Schools will then be re-ranked within each stratum by random number, from lowest to highest.

The first three schools in each stratum will be approached to participate as a case study school. Where a school declines, the next school in the respective stratum will be approached until three case-study schools in each stratum have agreed to participate, resulting in a final sample of twelve case study schools.

Table 5. Random seed values used to generate random number sequences

	Seed (set)
Mixed Attainment Higher	52
Mixed Attainment Lower	66
Setting Higher	22
Setting Lower	18

Data collection

A complete overview of the data collection methods in relation to the IPE dimensions and research questions is presented in Table 6.

Surveys

Four surveys will be administered as part of the IPE[9]:

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 included in Appendix H and includes a mixture of closed and open questions.
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 contains only closed questions and will take approximately 15 minutes to complete.
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

[9] 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.

includes a mixture of closed and open questions and will take approximately 15 minutes to complete.

4. **Student endline online survey** to collect data on perceptions of teaching; opportunity to learn; mathematics aspirations; and attitudes towards grouping practices. This survey includes only closed questions and will take approximately 30 minutes to complete. This survey addresses research questions for both the IPE and the outcome evaluation.

Case studies

The focus in the case study work is both on within-case and on between-case comparison (Miles et al., 2019). Within each case study school, we will compare the experiences of students with higher and lower mathematics attainment. In the between-case analysis we will compare 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) helps 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 will also provide detailed evidence for practitioners on how attainment grouping in mathematics is implemented across different contexts.

Our research questions demand comparison within and between two arms of the study. The focus for comparison includes experiences of students with low and high prior attainment, the quality of a mathematics learning environment (including teaching quality and opportunity to learn), and contextual factors influencing grouping practices, as well as other variables to be developed in analyses. To identify broader patterns, we will collect comparable data across case studies.

Year 7 (May-July 2023)

Case study visit one will include:

- 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 students from each observed lesson (two paired interviews).

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

As part of the lesson observations:

- Two researchers will make 'real time' notes of events/ activity in the lessons. Photos of the classroom board/screen will be taken to augment the lesson notes. After the lesson, the researchers will use 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 (Appendix I). Where 3 represents the highest quality and 1.5 and 2.5 represent elements of the respective higher level.

- Field notes will be taken on focal students (four per lesson; balanced by prior attainment in mixed attainment classrooms) and then utilised as data sources and interview prompts. Schools will be asked to identify four focal students with parental consent in advance of each lesson observed. In mixed attainment schools, they will be asked to identify two students working at a prior attainment level *above* the majority of the class 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 will be asked to identify students working at mid-range prior attainment compared with other students in that same set.
- Lessons will be audio-recorded using IRIS Connect software to allow for the transcription of small episodes of interest as the need arises. Lesson narratives will be written to summarise each lesson drawing on the two researchers' observation notes and the recordings.
- Paired interviews with the four focal students will be conducted after each observed lesson to gain their perspectives on the lesson. In mixed attainment schools the students will be paired according to prior attainment with one pair of students with low prior attainment and one pair with high prior attainment.
- Class teachers will be interviewed after the observed lessons to gain their perspectives and reflections on the lesson in particularly with respect to their students' experiences during the lesson.
- Artefacts such as schemes of work for Year 7 and 8 maths and lesson plans for each of the observed lessons will be collected to use as prompts for teacher interviews and to examine differences in the curriculum offered, and opportunity to learn afforded, to different students.

Year 8 (November 2023 – March 2024)

The same schools will be visited again in the second year of the study. Case study visit two will include:

- Observations of two mathematics lessons (as above).
- Post-observation interviews with Year 8 mathematics teachers and focal students after the observed lessons (as above).
- Focus groups with Year 8 students focusing on typical experiences of mathematical content and cognitive demand[10]. These students will ideally be different from the focal students observed in the lessons. For the purposes of focus groups, schools will be 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.

[10] Definitions according to the TRU framework referred to above.

Each case study will therefore include:

- Head of Mathematics interview
- 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
- One focus group with four Year 8 students with high prior attainment

An overview of all IPE methods with the total numbers of participants is presented in Table 6.

Table 6. IPE methods overview. Where multiple methods address the same research question, sources will be combined as described in the data analysis section.

IPE Dimension	RQ addressed	Research Methods	Data collection methods	Sample size and sampling criteria	Data analysis methods
Programme differentiation (including compliance)	RQ6	Survey	Online questionnaire	HOM (n=110)	Descriptive statistics
		Interview	Semi-structured interviews	HOM (n=12)	Deductive coding Thematic analysis – inductive coding
		Interview	Semi-structured interviews	Y7 teachers (low set) (n=6) Y7 teachers (high set) (n=6) Y7 mixed attainment teachers (n=12) Y8 teachers (low set) (n=6) Y8 teachers (high set) (n=6) Y8 mixed attainment teachers (n=12)	Deductive coding Thematic analysis-inductive coding
Students' mathematical experiences	RQ5, RQ7	Survey	Online questionnaire	Year 8 students (n=9000)	Descriptive statistics
		Focus groups	Semi-structured interview/ sorting task	Lower prior attaining Y8 students in low sets (n=24) Higher prior attaining Y8 students in high sets (n=24) Lower prior attaining Y8 students in mixed attainment (n=24) Higher prior attaining Y8 students in mixed attainment (n=24)	Deductive coding Thematic analysis – inductive coding
		Paired Interview	Semi-structured interviews	Lower prior attaining students in low sets (n=48)	Deductive coding Thematic analysis-inductive coding

IPE Dimension	RQ addressed	Research Methods	Data collection methods	Sample size and sampling criteria	Data analysis methods
				Higher prior attaining students in high sets (n=48) Lower prior attaining students in mixed attainment (n=48) Higher prior attaining students in mixed attainment (n=48)	
		Observation	Structured lesson observations	Y7 Mixed attainment lessons (n=12) Y8 Mixed attainment lessons (n=12) Y7 Low set lessons (n=6) Y7 High set lessons (n=6) Y8 Low set lessons (n=6) Y8 High set lessons (n=6)	TRU analysis Narrative account
Responsiveness of pedagogy	RQ5, RQ8	Focus groups	Semi-structured interview/ sorting task	Lower prior attaining Y8 students in low sets (n=24) Higher prior attaining Y8 students in high sets (n=24) Lower prior attaining Y8 students in mixed attainment (n=24) Higher prior attaining Y8 students in mixed attainment (n=24)	Deductive coding Thematic analysis – inductive coding
		Paired Interview	Semi-structured interviews	Lower prior attaining students in low sets (n=48)	Deductive coding Thematic analysis – inductive coding

IPE Dimension	RQ addressed	Research Methods	Data collection methods	Sample size and sampling criteria	Data analysis methods
				Higher prior attaining students in high sets (n=48) Lower prior attaining students in mixed attainment (n=48) Higher prior attaining students in mixed attainment (n=48)	
		Interview	Semi-structured interviews	Y7 teachers (low set) (n=6) Y7 teachers (high set) (n=6) Y7 mixed attainment teachers (n=12) Y8 teachers (low set) (n=6) Y8 teachers (high set) (n=6) Y8 mixed attainment teachers (n=12)	Deductive coding Thematic analysis-inductive coding
		Survey	Online questionnaire	Maths teachers (n=500)	Descriptive statistics
Context	RQ9	Survey	Online questionnaire	Senior leaders (n=110)	Descriptive statistics
		Survey	Online questionnaire	HOM (n=110)	Descriptive statistics
		Interview	Semi-structured interviews	HOM (n=12)	Deductive coding Thematic analysis

Data analysis

The IPE data sources will be 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 will contribute to the quantitative side of the study during the design of the endline surveys with teachers and students.

Analysis of survey data

Data from surveys will be analysed using descriptive statistics including frequencies and cross-tabulation. Cross-tabulation will include analysis based on prior attainment, grouping type, FSM status, gender and ethnicity. Free text answers will be 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 will write up each case using a standard set of TRU variables which underpin all data sources. Once each case is well understood and within-case analysis is complete, we will move 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 sub-themes derived from the data. The final systematic comparison between the cases in setting and mixed attainment groups will be informed by our research questions. All data management, analysis and write-ups will be supported by computer assisted qualitative data analysis software (Nvivo 12).

Lessons and lesson narratives

TRU rating

Attributed scores for each domain (whole class (WC) and/or small group work (SG)) according to the TRU framework rubric (Appendix I) will be analysed to explore whether there is any differential quality in the five dimensions experienced by students in mixed attainment classes compared with low sets and high sets.

Coding based on the TRU framework

The lesson narratives produced by maths specialists will firstly be analysed using the deductive coding framework (Table 5, column A) and for evidence of the twelve evidence-based strategies. Inductive qualitative content analysis (Mayring, 2019) will then 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 may be 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 will be explored using cross-tab 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). For example, it might be possible to illustrate different formative assessment practices in mixed attainment classes compared with high sets and indicate whether the practices in mixed attainment are rated higher or lower than in high sets according to the TRU rubric.

Structural coding of lessons will also be used to indicate the occurrences of classroom activity involving whole class, group work and individual work. This will enable further interrogation of differential classroom activity between mixed attainment, low set and high set.

Post-lesson focal student and teacher interviews

Deductive coding based on the TRU framework

Segments of interview text will be deductively coded according to the conceptual framework set out in Table 7 (columns B and C) together with codes relating to the twelve evidence-based strategies set out in Table 1. Since the interview schedules (see Student Interview: Appendix J and Teacher interview: Appendix K) have been designed using the TRU framework, it will be possible to code segments according to the interview schedule items. However, since the interviews are semi-structured, some responses may be coded to more than one domain code. For instance, responses about mathematical content may also involve cognitive demand and will therefore be coded to both.

Pattern coding

Thematic analysis using Braun and Clark's (2019) 6-stage approach to reflexive thematic analysis will involve further inductive coding to identify themes which describe student or teacher experiences and perspectives within and across the five TRU dimensions in the context of different grouping practices. For example, this approach will identify patterns to describe the range of low prior attaining students' experiences and compare these with the experiences of students from different grouping practices.

The post-lesson interview data will be triangulated with the lesson narratives to create lesson cases to facilitate the multi-case study analysis. The lesson plans collected during field visits will be utilised to support this analysis.

Focus groups with students

As above, the analysis will start with deductive coding according to the TRU framework for *mathematical content* and *cognitive demand* (Appendix L). Thematic analysis using Braun and Clark's (2019) approach will involve inductive coding to identify themes which describe 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 (Appendix M) will be 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 integrates case-orientated (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 opportunity to learn. 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

We will bring together the multiple data sources first by triangulating using our theoretical framework of the TRU (see Table 7). Second we will compare and contrast findings against each research question, looking for agreement and dissonance among sources (see Table 6).

Table 7. Deductive coding framework using the teaching for robust understanding framework for different data sources

TRU Dimension code	(A) Lesson narratives	(B) Focal student interviews	(C) Teacher interviews
Mathematical content	Examples from the lesson of mathematical content, discussed by teacher and/or students	Segments of the interview where students reflect on the mathematical content in the lesson observed. (Interview schedule q1-q4)	Segments of the interview where the teacher reflects on the intended and lived mathematical content of the lesson observed.
Cognitive demand	Examples from the lesson where students are engaged in different forms of mathematical thinking (e.g. single word responses, explaining own reasoning, explaining interpretation of others' thinking, asking the teacher/ class mathematical questions)	Segments of the interview where students reflect on the level of mathematical challenge they experienced in the lesson observed. (Interview schedule q3-q10)	Segments of the interview where the teacher reflects on the intended and lived levels of mathematical challenge in the lesson observed.
Access	Examples from the lesson where teaching practices engage different students' participation. (e.g. whiteboard sharing, group work)	Segments of the interview where students reflect on their own and others' participation in the lesson observed. (Interview schedule q11-q14)	Segments of the interview where the teacher describes the practices they used to ensure wide participation and reflects on the different participation patterns of different students in the lesson observed.
Agency, ownership and identity	Examples from the lesson of teaching practices involving students' mathematical contributions i.e. where these are built on or not. (e.g. the use of a student's mathematical error, bringing a student to the board to be 'the teacher')	Segments of the interview where the students reflect on the usefulness of their own and others' contributions in the lesson observed. (Interview schedule q15-q18)	Segments of the interview where the teacher reflects on moments in the lesson where students' contributions were useful or not in the observed lesson.
Formative assessment	Examples from the lesson where the teacher solicits student responses and does or does not make use of them.	Segments of the interview where the students reflect on how well they think they did in the lesson and how they think their teacher knows how well they did. (Interview schedule q19-q22)	Segments of the interview where the teacher reflects on what they learned about their students' during the lesson and moments where the teacher did or did not choose to develop on students' ideas.

Cost evaluation

We are not evaluating an intervention in this study, so there are no specific costs associated with implementation. Nevertheless, we will investigate the different costs associated with maintaining the different grouping practices. In addition, we anticipate that some schools in our sample are likely to have changed their grouping practices within the last 5 years, and we will investigate the costs of these changes in terms of staff training, materials and management support. Where appropriate, we will follow the latest EEF Guidance on Cost Evaluation in estimating these costs. Using implementation evidence from interviews and "light-touch" survey, we will estimate actual costs alongside marginal costs. We note that, given the timing of any changes to practice, these costs are likely to be rough estimates. Costs will be reported as an average cost per student in the cohort being researched. We are minded to calculate the average costs over three years to facilitate comparison with standard EEF trials.

Ethics

The project has full ethical approval from UCL Institute of Education Research Ethics Committee (reference REC1139).

We will provide information about the research to parents/carers of all students, students themselves and teachers prior to the collection of UPNs from participating schools, and allowing participants, or their parents/carers, to withdraw their data from the research (for ethics purposes; see below for a discussion of data protection considerations). This will cover NPD matching, collection of school data, participation in surveys and focus groups, and data archiving. We will seek opt-in consent from all students and teachers prior to participation in surveys and focus groups.

It is important to understand that consent is only one of a range of conditions for using personal data available in the current DPA and in the GDPR. Under the GDPR it will still be possible to process data for public interest purposes (as per condition 5(d) from Schedule 2 of the current Data Protection Act, which will have a direct parallel in the GDPR). The current advice that we have received indicates that, in the studies such as the current proposal, and where we do seek access to the most sensitive personal data, consent will not be required, and that offering an opt-out is a good way of demonstrating that the project does not impinge on anyone's rights.

Data protection

The project has been fully approved for compliance with data protection regulations including GDPR by UCL's data protection team (registration number: Z6364106/2018/11/03 social research).

Students and their parents or carers, and teachers, will be informed of the proposed data processing and given an opportunity to object to this, and withdraw their, or their child's, data. The information which will be provided to parents/carers, pupils and teachers explains in clear and plain non-technical language the purpose to which we will put the data, that they can object to this data and this will be respected, contact details of the organisation, and categories of data that we will be processing and that the data processing will be compliant with the GDPR and data protection legislation. Further details on the lawful basis for data processing are available on request.

The evaluation team at UCL have carried out a data protection impact assessment and will put in place a data management plan. As part of this data management plan, data will be checked and cleaned to ensure the GDPR principle (d) of accuracy is met.

UCL, QUB and Brunel University will sign a data sharing agreement outlining data security and protection issues. Schools will sign a data sharing agreement with UCL.

UCL has produced specifications on what GDPR means for research conducted within UCL and how project teams can prepare for the GDPR. The Essentials Guide will guide our project set up:

https://www.ucl.ac.uk/legal-services/guidance/general-data-protection-regulation-gdpr

Link to UCL Data Protection Policy:

https://www.ucl.ac.uk/drupal/legal-services/sites/legal-services/files/migrated-files/DataProtectionPolicy1016.pdf

Data security

All personal data collected or obtained as part of this project will be treated as "Highly Restricted" under UCL Data Protection classification guidance. Personal data (pupil names, UPNs, dates of birth, FSM eligibility, sex, national test results, class and teacher, as well as teacher names and survey data) will be 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 will be pseudonymised. Once pseudonymised it will be 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 recording will be stored in a locked filing cabinet within a locked office at UCL to which only the research team will have access.

Some data transfer will be required between collaborators on this project at UCL and QUB. This will be conducted by making a secure remote connection (e.g. VPN) between the university networks and transferring data across this. In addition, the data will be encrypted before sharing using a password shared between research team members by separate communication.

Data from the National Pupil Database will be accessed through the ONS Secure Research Service (ONS) at an approved Safe Setting. The procedures for this access are currently being developed by the ONS and the DfE. Our understanding is that we will submit the pupil level data securely to the ONS system using end-to-end encryption. These data will then by matched to the NPD by the DfE and an extract will then be made available for analysis on the ONS secure system. UCL has ONS approval for limited institutional access to the ONS system via an onsite safe room. Access to the ONS system will only be available to ONS approved researchers.

Schools will be required to submit personal data to UCL. This will be conducted via the Data Safe Haven's direct data transfer portal. Schools will be provided with clear guidance on securely submitting and protecting this data.

Online surveys for teachers will be administered through UCL's REDCap survey system whereby data is uploaded directly to the DSH in an encrypted form (if this is possible for the entire survey).

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

The DSH environment is certified to ISO27001:2013 with BSI – certificate number: IS 612909. The most recent external audit was in May 2017. 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 on-going 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 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/TLS with data retained via policy and systems that prevent data leakage.

Data will be kept for at least the duration of the project, until successful submission of the data to the EEF's data archive has been agreed by the funder. We may keep anonymised data beyond this period for the purpose of supporting submissions and revisions to submissions to academic journals. They will be kept for no longer than 10 years in line with UCL's guidance on retention of records for research.

Personnel

The team is based at IOE, UCL's Faculty of Education and Society and led by Professor Hodgen and Dr Taylor.

Professor Jeremy Hodgen: Project leadership and strategic management, contribution to all aspects of study.

Dr Becky Taylor: Project leadership and management; recruitment and retention (including matching), quantitative analysis, instrument design & analysis.

Dr Antonina Tereshchenko (Brunel University): implementation analysis and case studies, recruitment and retention of schools.

Dr Jake Anders: Additional statistical, design and methodological expertise & advice: Statistical design and recruitment (matching); quantitative analysis.

Dr Maria Cockerill (Queens University Belfast): recruitment of schools, implementation analysis and case studies.

Dr Rosa Kwok: Research fellow developing and validating the teacher and student surveys and provide assistance with administering the teacher survey in Year 7.

Dr Laurie Jacques: Research fellow providing expertise in qualitative data collection and assist with carrying out the qualitative component of the implementation analysis and subsequent coding and preliminary analysis of the data.

Dr Piers Saunders: Lecturer in mathematics education at IOE, contributing to case study data collection.

An administrator: An administrator will provide day-to-day support for the project, including supporting recruitment and data collection.

Steering Board

A Steering Board will provide independent oversight of the research and advice on the research design and methods. See Appendix G for the terms of reference and membership for this group.

Risks

Table 8. Risks and mitigations.

Risk	Likelihood	Impact	Action
Failure to recruit	Moderate	High	• Establish timeline for recruitment involving a variety of methods • Regular review of recruitment processes • Commit staffing to recruitment • Allow schools to join the mixed attainment group that have a nurture group. • Should it not be possible to recruit 40 mixed attainment mathematics schools by an agreed deadline, recruit 40 mixed attainment English schools instead.
Failure to gain data from schools	Low	High	• Training of recruitment and administrative staff to handle GDPR-related concerns from schools. • Appropriate financial incentives, timed to be issued following outcome data collection timepoints.
Attrition of schools	Moderate	Moderate / High	• Over-recruit schools • Appropriate financial incentives • Regular contact with participating schools • Allocate staff time to school liaison at key data collection points • Collect data at two potential exit points (end of Year 7 and end of Year 8)
Some mixed attainment schools introduce setting (and vice versa)	Moderate	Moderate	• Over-recruit schools • Investigate through the IA
Loss of staff	Low / Moderate	Low	• UCL IOE has a large staff team and would reallocate staff

Timeline

Table 9. Project timeline

Date	Activity	Staff responsible/ leading
Nov – Dec 2018	Ethical & data protection approval sought Data security Protocol written Recruitment manager and administrator appointed	BT/AT/JH/JA
Jan 2019	MOUs agreed	BT/AT
Feb – Jul 2019	Recruitment & matching Prepare student baseline survey Prepare school practices survey	MC/BT BT/AT BT/AT
Aug – Sep 2019	Statistical analysis plan written	JH/JA
Sep – Oct 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
Sep 2019 – Aug 2020	Survey instrument design and validation Qualitative instrument design & piloting (Sep – Feb) Qualitative data collection (Mar – Jul)	RA1 AT AT/RA2
Study interrupted due to COVID-19 pandemic and restarted in 2022		
Jan 2022	Ethics and data protection revised	BT
Feb – Jul 2022	Recruitment & matching Update and prepare student baseline survey Update and prepare school practices survey	Admin/MC/JA BT/AT BT/AT
Sep – Oct 2022	Contacting schools about collecting data, processing opt out forms Receive Year 7 student information from schools (student names, dates of birth, School ID and Unique Pupil Numbers (UPN)) Baseline student self-confidence survey	Admin Admin BT/Admin
Sep 2022 – Jul 2023	Qualitative instrument design & piloting (Sep – Feb) Wave 1 qualitative data collection (Mar – Jul)	AT AT/LJ
March – April 2023	Heads of Maths online survey	BT
March 2023	Case study schools selected and recruited	AT/LJ
May – July 2023	*Case study school phase 1 visits Year 7*: Classroom observation and face-to-face interviews with students, teachers and Heads of Maths; 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; documents (lesson materials, schemes of learning).	AT/LJ
May – June 2024	Senior leader survey	BT

May – July 2024	Mathematics teacher survey	BT
May – July 2024	Outcome measure data collection: - Student endline survey (outcome self-confidence, opportunity to learn, perceptions of mathematics teaching) - Progress Test in Mathematics Teacher survey Senior Leader survey	BT/Admin
Jul 2024 – Jan 2025	Outcome analysis Implementation analysis Report writing	JH/BT/JA AT/BT/JH All

References

Anders, J., Brown, C., Ehren, M., Greany, T., Nelson, R., Heal, J., . . . Allen, R. (2017). *Evaluation of complex whole-school interventions: methodological and practical considerations*. Retrieved from London: https://educationendowmentfoundation.org.uk/public/files/Evaluation/Setting_up_an_Evaluation/EEF_CWSI_RESOURCE_FINAL_25.10.17.pdf

Archer, L., DeWitt, J., Osborne, J., Dillon, J., Willis, B., & 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. doi:10.3102/0002831211433290

Baron, Reuben M., and David A. Kenny. 1986. "Mod- erator-Mediator Variables Distinction in Social Psy- chological Research: Conceptual, Strategic, and Statistical Considerations." *Journal of Personality and Social Psychology* 51(6):1173–82.

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., et al. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal*, 47, 133-180. http://dx.doi.org/10.3102/0002831209345157.

Boaler, J., Wiliam, D., & Brown, M. (2000). Students' Experiences of Ability Grouping - disaffection, polarisation and the construction of failure. *British Educational Research Journal*, 26(5), 631–648.

Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qual. Res. Sport Exerc. Health,* 11(4), 589-597. https://doi.org/10.1080/2159676X.2019.1628806

Burris C.C., Heubert J.P., & Levin H.M. (2006). Accelerating mathematics achievement using heterogeneous grouping. *Am Educ Res J.*, 43(1), 137–54.

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys, 22*(1), 31-72. doi:10.1111/j.1467-6419.2007.00527.x

Cook, T., Shadish, W., & Wong, V. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management, 27*(4), 724-750.

Cooper, P., & 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. Retrieved from https://www.edweek.org/ew/articles/2015/01/07/differentiation-doesnt-work.html

Dracup, T. (2014). The Politics of Setting. Retrieved from https://giftedphoenix.wordpress.com/2014/11/12/the-politics-of-setting/

Dunne, M., Humphreys, S., Dyson, A., Sebba, J., Gallannaugh, F., & 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

Dunne, M., Humphreys, S., Sebba, J., Dyson, A., Gallannaugh, F., & Muijs, D. (2007). *Effective teaching and learning for pupils in low attaining groups*. London.

EEF. (2017). Improving mathematics at Key Stages Two and Three: guidance report. Education Endowment Foundation.

EEF (2021). *Teaching and Learning Toolkit*. London: Education Endowment Foundation. https://educationendowmentfoundation.org.uk/education-evidence/teaching-learning-toolkit

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

Francis, B., Archer, L., Hodgen, J., Pepper, D., Taylor, B., & 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.

Francis, B., Connolly, P., Archer, L., Hodgen, J., Mazenod, A., Pepper, D., . . . 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. doi:https://doi.org/10.1016/j.ijer.2017.09.001

Francis, B., Craig, N., Archer, L., Hodgen, J., Mazenod, A., Taylor, B., & Tereshchenko, A. (2019). Teacher 'quality' and attainment grouping: the role of within-school teacher deployment in social and educational inequality. *Teaching and Teacher Education*.

Gutman, L. M., & Schoon, I. (2013). *The impact of non-cognitive skills on outcomes for young people: Literature review*. Retrieved from London: https://educationendowmentfoundation.org.uk/public/files/Publications/EEF_Lit_Review_Non-CognitiveSkills.pdf

Hallam, S. and J. Ireson (2003). "Secondary school teachers' attitudes towards and beliefs about ability grouping." British Journal of Educational Psychology **73**(3): 343-356.

Hayes, A. F. (2013). Introduction to mediation, moderation and conditional process analysis: A regression-based approach. New York: Guildford Press.

Higgins, S., Katsipataki, M., Coleman, R., Henderson, P., Major, L., Coe, R., & Mason, D. (2018). *Education Endowment Foundation Teaching and Learning Toolkit*. London: Education Endowment Foundation.

Hill, H., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal, 42*(2), 371-406.

Ho, D., Imai, K., King, G., & Stuart, E. (2011). MatchIt: nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software, 42*(8), 1-28.

Hodgen, J., Taylor, B., Francis, B., Archer, L., Mazenod, A., & Tereshchenko, A. (forthcoming). Resisting homogeneous 'ability' grouping practices: how some schools buck the trend and implement mixed attainment teaching.

Hodgen, J., Foster, C., Coe, R., Brown, M., Higgins, S., & Küchemann, D. (2020). Low attainment in mathematics: An investigation focusing on Year 9 students in England. Technical Report. UCL Institute of Education.

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

Hodgen, J., Foster, C., Brown, M. *et al.* (2023a). Low-Attaining Secondary School Mathematics Students' Perspectives on Recommended Teaching Strategies. *Int J of Sci and Math Educ*. https://doi-org.ezproxy.brunel.ac.uk/10.1007/s10763-023-10420-8

Hodgen, J., Taylor, B., Anders, J., Tereshchenko, A., & Francis, B. (2019) The Student Grouping Study: investigating the effects of setting and mixed attainment grouping: Study Plan. London: Education Endowment Foundation. https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/Students_Grouping_Study_Study_Plan_2019.11.19.pdf?v=1718801483

Hodgen, J.; Taylor, B., Francis, B., Craig, N., Bretscher, N., Tereshchenko, A., Connolly, P., & 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. 10.1002/berj.3838.

Humphrey, N., Ledrum, A., Ashworth, E., Frearson, K., Buck, R., & Kerr, K. (2016). *Implementation and process evaluation (IPE) for interventions in education settings: An introductory handbook*. Retrieved from London: https://educationendowmentfoundation.org.uk/public/files/Evaluation/Setting_up_an_Evaluation/IPE_Guidance_Final.pdf

Imai, Kosuke, Luke Keele, and Dustin Tingley. 2010. "A General Approach to Causal Mediation Analysis." *Psychological Methods* 15(4):309–334.

Jackson, B. (1964). *Streaming: An education system in minature*. 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, & A. Kurz (Eds.), *The handbook of accessible achievement tests for all students: Bridging the gaps between research, practice, and policy* (pp. 99–129). Springer.

Maxwell, B., Stevens, A., Demack, S., Coldwell, M., Wolstenholme, C., Reaney-Wood, S., Stiell, B., & Lortie-Forgues, H. (2021). *Review: EEF Implementation and Process Evaluation (IPE) Quality Pilot.* EEF.

Mayring, P. (2020). Qualitative Content Analysis: Demarcation, Varieties, Developments [30 paragraphs]. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, 20(3), Art. 16. http://dx.doi.org/10.17169/fqs-20.3.3343.

Mazenod, A., Francis, B., Archer, L., Hodgen, J., Taylor, B., Tereshchenko, A., & Pepper, D. (2018). Nurturing Learning or Encouraging Dependency? Teacher Constructions of Students in Lower Attainment Groups in English Secondary Schools. *Cambridge Journal of Education*.

Miles, M. B., Huberman, A. M., & Saldaña, J. (2019). *Qualitative data analysis: A methods sourcebook* (4th edn.). London: Sage.

Nitz, S., Prechtl, H., & Nerdel, C. (2014). Survey of classroom use of representations: development, field test and multilevel analysis. *Learning Environments Research, 17*(3), 401-422. doi:10.1007/s10984-014-9166-x

OECD. (2013). Selecting and grouping students *PISA 2012 results: What makes schools successful? Resources, Policies and Practices* (Vol. IV, pp. 71-92). Paris: OECD Publishing.

Old, A. & Reddy, B.. (2015). Mixed ability works debate. Retrieved from https://youtu.be/YUa_CI-stRo

Praetorius, A.-K., & Charalambous, C. Y. (2018). Classroom observation frameworks for studying instructional quality: looking back and looking forward. *ZDM, 50*(3), 535-553. doi:10.1007/s11858-018-0946-0

Preacher, K. J. (2015). Advances in Mediation Analysis: A Survey and Synthesis of New Developments. *Annual Review of Psychology, 66*(1), 825-852. doi:10.1146/annurev-psych-010814-015258

Roy, P., Styles, B., Walker, M., Morrison, J., Nelson, J. & Kettlewell, K. (2018) Best Practice in Grouping Students Intervention A: Best Practice in Setting Evaluation report and executive summary. London: Education Endowment Foundation.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*(3), 164-183. doi:10.1111/j.1756-5391.2009.01032.x

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., & the Teaching for Robust Understanding Project. (2016). *An Introduction to the Teaching for Robust Understanding (TRU) Framework*. Berkeley, CA: Graduate School of Education. http://truframework.org or http://map.mathshell.org/trumath.php.

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

Schwartz W (1995) Opportunity to learn standards: Their impact on urban students. ERIC/CUE Digest 110: 1–9.

Slavin, R. E. (1990). Achievement effects of ability grouping in secondary schools: A best-evidence synthesis. *Review of educational research, 60*(3), 471-499.

Sobel, Michael E. 1986. "Some New Results on Indi- rect Effects and Their Standard Errors in Covari- ance Structure Models." *Sociological Methodology* 16:159–86.

Steenbergen-Hu, S., Makel, M. C., & 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. doi: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. doi:10.1177/1745499917711549

Taylor, B., Francis, B., Archer, L., Hodgen, J., Pepper, D., Tereshchenko, A., & Travers, M.-C. (2017). Factors deterring schools from mixed attainment teaching practice. *Pedagogy, Culture & Society, 25*(3), 327-345. doi:10.1080/14681366.2016.1256908

Taylor, B., Hodgen, J., Tereshchenko, A., & Gutiérrez, G. (2020). Attainment grouping in English secondary schools: A national survey of current practices. *Research Papers in Education*, 37(2), 199-220. doi:10.1080/02671522.2020.1836517.

Taylor, B., Francis, B., Craig, N., Archer, L., Hodgen, J., Mazenod, A., . . . Pepper, D. (2018). Why is it difficult for schools to establish equitable practices in allocating students to attainment 'sets'? *British Journal of Educational Studies*. doi:10.1080/00071005.2018.1424317

Taylor, B., Hodgen, J., Francis, B., Archer, L., Mazenod, A., Tereshchenko, A., & Travers, M.-C. (forthcoming). Can a mixed attainment grouping intervention increase equity in English secondary schools?

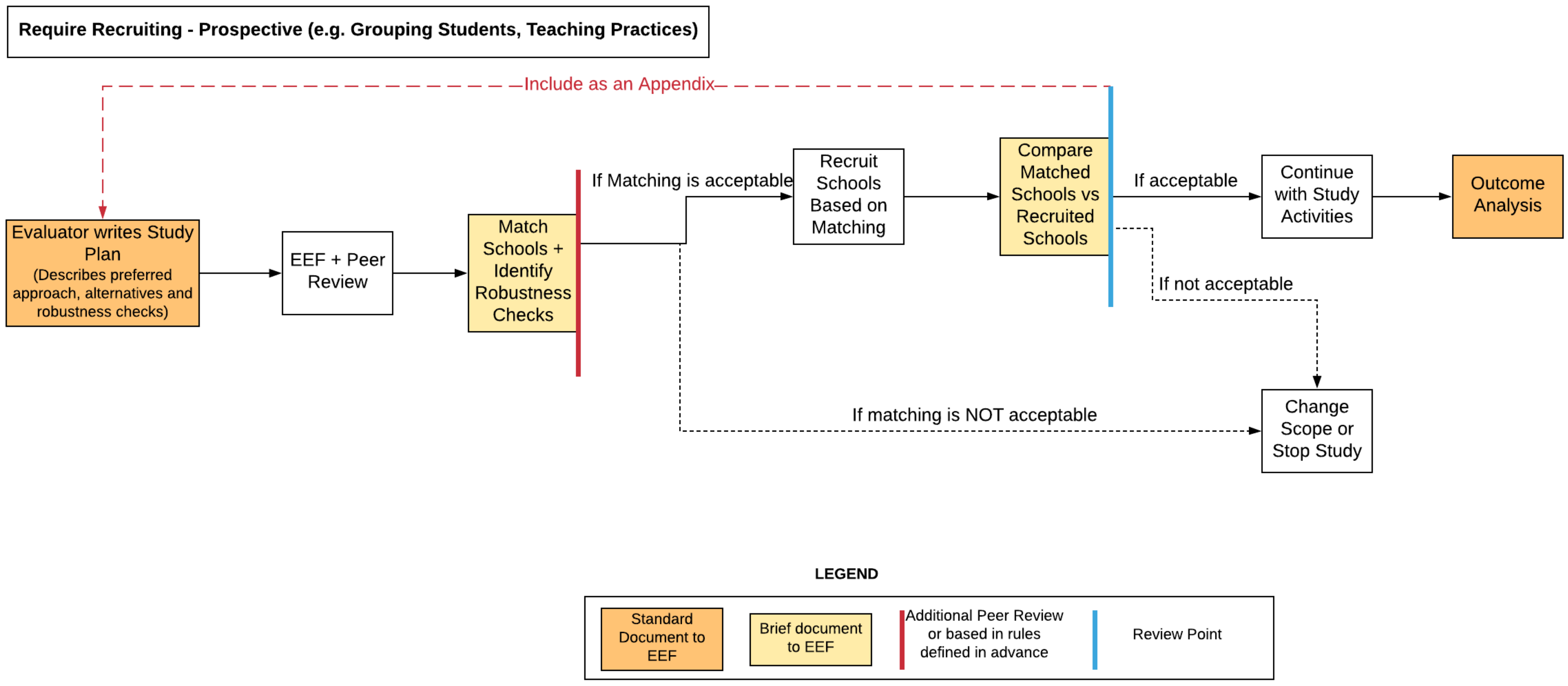
Tereshchenko, A., Francis, B., Archer, L., Hodgen, J., Mazenod, A., Taylor, B., Pepper, D., & Travers, M.-C. (2018). Learners’ attitudes to mixed-attainment grouping: examining the views of students of high, middle and low attainment. *Research Papers in Education*, 1-20. https://doi.org/10.1080/02671522.2018.1452962

Acknowledgments

We would like to thank the members of the project Steering Board for their useful comments and suggestions to inform and shape this document. The Group is composed of Rob Coe, Richard Dorsett, Bronwen Maxwell, Fatima Hussain and Stefan Speckesser. Heather Rolfe was previously a member of the steering group.

APPENDIX A: Review process

The standard process for studies recruiting comparator groups is illustrated below. We propose to write a detailed statistical analysis plan after recruitment on the basis of the actual matching.



APPENDIX B: Grouping Study Research Design Options

In designing this study, we considered the pros and cons of a number of six research design options to investigate to impact of mixed attainment grouping & to address the broad questions identified by EEF as of interest to teachers and schools:

- Is high quality mixed-ability teaching is more effective than setting in secondary schools?
- If successful, to what extent is it the mixed-ability teaching that leads to improvement, or the effects of a high quality PD intervention (and could the PD intervention also pay dividends in schools that set)?

The options include a standard EEF RCT [option A]. Option C described the preferred design.

Option	Issues	Pros	Cons
A: Standard EEF RCT design: schools randomly assigned to mixed attainment intervention, business as usual control	Pilot study indicates that mixed attainment teaching is a complex whole-school intervention, where standard RCT designs are not appropriate (Anders et al, 2017, Report to EEF), because the intervention: • Reduces heads options to allocate students and teachers to classes • Requires substantial pedagogy & curriculum change	• On the basis of the pilot, it is difficult to see any advantages to this option.	• Recruitment difficult (as we know from the pilot), because 'difference' between intervention and the control is huge • Counterfactual very problematic – from existing pilot, we know the schools who sign will adopt some form of mixed attainment practice ('always compliers') whether assigned to the intervention or not.
B: Head-to-head 'competitive' trial: Recruit two groups setting v mixed attainment on a quasi-experimental basis	• PD required for both groups • Ideally a year of lead-in time, particularly for the mixed attainment group.	• Very clear counterfactual • Good match to EEF's core question • Addresses the PD v mixed attainment at least partially (both groups would have good PD)	• Need to over-recruit to achieve well-matched groups

Option	Issues	Pros	Cons
C: Naturalistic design: Recruit two groups of school - ones who are already doing setting and ones who are already doing mixed attainment	• No need for PD. • Need to focus on English.	• Easy to organise • Sidesteps the PD question • Recruitment in English less problematic (although may need incentives for both groups)	• Less robust design • Need to tightly define eligibility criteria in order to avoid practice within the two groups being too varied (and thus confounding the counterfactual). • Recruitment may be difficult in mathematics. • Need to over-recruit to get well-matched groups (and more so than for Option B). • Incentives required for both groups. • Does not address the PD question.
D: Within-school student-level RCT: recruit large schools prepared to try out mixed attainment in some classes, whilst setting in others	• PD required. • Ideally a year of lead-in time	• Very clear counterfactual • Fewer schools need to achieve adequate power (although recruitment would still be a problem because recruitment would be from a restricted population of schools) • Smaller number of schools so ensuring compliance easier	• Difficult to do in English, recruitment might be difficult (needs large schools that organise each year into equivalent 'bands') • Would need strong school leadership to sell to parents • Doesn't isolate the PD issue
E: Recruit schools to a mixed attainment trial and use previous years cohorts as a control	• PD required. • Ideally a year of lead-in time	• Avoids the 'always compliers' problem in the control • Schools well-matched (but students less well-matched)	• Recruitment still a problem (and it might be worse as we would need to recruit heavy 'setters' who were prepared to quickly change to heavy mixed-attainment practice • Less robust design • Doesn't isolate the PD issue

Option	Issues	Pros	Cons
F: Focus on low attainers, RCT design, randomisation at school level: Randomly allocate low attainers to the middle attainment groups, with control 'business as usual' low attaining groups.	• Recruit schools that have a relatively small number of set levels (say, 3 or 4, including low attaining sets). • Provide high-quality catch up on core subjects (reading, writing, mental calculation / number) to enable low attainers to access the curriculum.	• Addresses low attaining and disadvantaged students	• Potentially a difficult sell to schools (and might not be attractive to the many schools who use a 'nurture' group) • Recruitment could be a problem • Requires a slightly larger sample to achieve power (for MLM, critical value for within-school student sample size is ~50 • Sidesteps the PD issue

APPENDIX C: Logic models

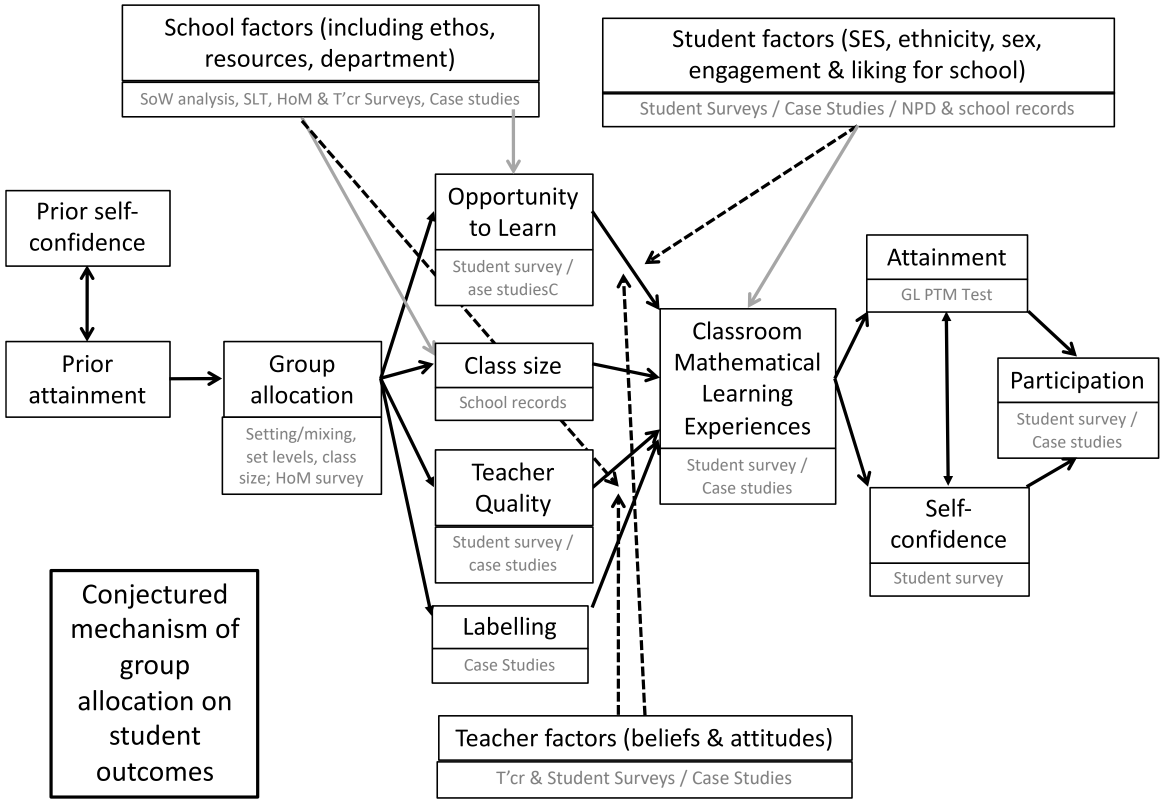


Figure C.1: Logic model illustrating the conjectured mechanism for the effects of group allocation on student outcomes. Bold arrows indicate direct relationships (including mediators); dashed arrows indicate potential moderating relationships; grey arrows indicate other indirect relationships (which will be explored in the implementation analysis)

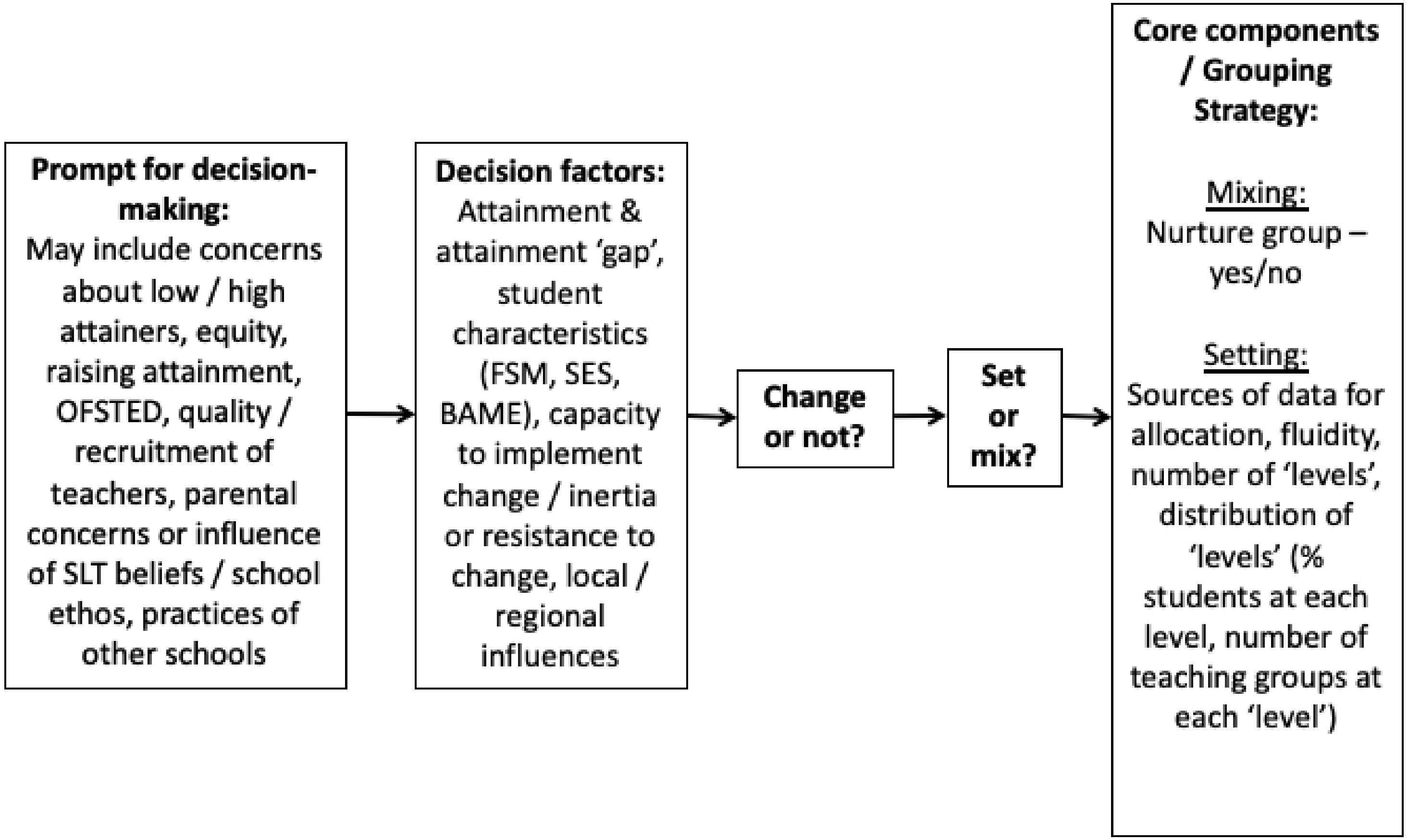


Figure C.2: Logic model illustrating the possible factors in school decisions about whether to group by attainment or in mixed attainment classes.

APPENDIX D: Matching Factors: what factors influence (or are correlated with) school decisions to set/mix?

Table D.1: Overview of potential matching factors

Factor	Justification	Indicator(s)	Issues	Influence decision?	Impact on outcomes?	Influenced by outcomes?	Use for matching?
Prior attainment	Aim to raise attainment across the board	Average KS2	Include historic data	Yes	Yes	No	Yes
	Low prior attainers – aim to close attainment gap / raise attainment of low prior attainers	Spread (SD) of prior attainment Or % of low prior attainers		Yes	Yes	No	Yes
	High prior attainers – maintain / raise attainment of high prior attainers	Spread (SD) of prior attainment Or % of high prior attainers		Yes	Yes	No	Yes
	Aim to close attainment gap between FSM and non-FSM at entry	FSM gap at KS2		Yes	Yes	No	Yes

Factor	Justification	Indicator(s)	Issues	Influence decision?	Impact on outcomes?	Influenced by outcomes?	Use for matching?
Attainment outcomes (KS4)	Aim to raise attainment outcomes across the board	A8 P8 En/Ma GCSE	Outcome (e.g., GCSE) data would affect the decision, but, as above, is not exogenous to treatment. Over a period of time this might be influenced by outcomes but probably over a long period of time.	Yes	Yes	Yes	No
	Low prior attainers – aim to close attainment gap / raise attainment of low prior attainers	Low prior attainers: A8 P8 En/Ma GCSE Spread (SD) of GCSE outcomes	As above.	Yes	Yes	Yes	No
	High prior attainers; Maintain / raise attainment of high prior attainers	High prior attainers: A8 P8 En/Ma GCSE Spread (SD) of attainment	As above	Yes	Yes	Yes	No

Factor	Justification	Indicator(s)	Issues	Influence decision?	Impact on outcomes?	Influenced by outcomes?	Use for matching?
Student characteristics	Aim to close attainment gaps	% FSM6 at intake	Over time, proportion of disadvantaged students could be affected by outcomes (e.g., as a result of changes to value added, or overall GCSE attainment, the school could become more or less attractive to disadvantaged or advantaged parents). To avoid this, take an average measure over time or a measure at some time point in the past (e.g. 5 years previously). Concern for FSM outcomes in high % FSM schools *may* make mixing more likely to avoid stigmatising low-attaining students receiving FSM or because of focus on raising attainment of FSM.	Yes	Yes	No	Yes
		% BAME	As for FSM. An indicator of greater diversity – can drive segregation (or forced integration)	Yes	Yes	No	Yes
		% EAL	As for FSM. Schools may choose to group learners by different stages of learning English.	Yes	Yes	No	Yes

Factor	Justification	Indicator(s)	Issues	Influence decision?	Impact on outcomes?	Influenced by outcomes?	Use for matching?
		% SEND	As for FSM. Parents of some SEND students do have choice, so this may be influenced by the outcomes (though v small numbers – only those with EHCP). But we could avoid this by taking a measure over time or at some point in the past as for FSM? BUT would it be a major factor in a school's decision?	Yes	Yes	May have small effect	No
		% in high and low IDACI neighbourhoods		Yes	Yes	No	Yes
		Gender ratio	Influence on decision unclear. Possible influence for co-educational schools with highly unbalanced intake, but otherwise may not be an influence. May be difficult for stakeholders to understand if we don't match on this.	Unlikely	Yes	No	No
School capacity to implement change	Schools unlikely to *change* practices or take on 'riskier' MA grouping without capacity to do so. Lots of unobservables, but some observable	Age range (sixth form or not)	May be an indicator of capacity to change, but unlikely to affect decision of grouping practices at KS3	Yes	Yes	No	No
		Size of school	Cannot set if cohort very small. Need at least 75 students/3 teaching groups in year group to make setting at 3 levels possible – implications for recruitment criteria.	Yes	Yes	No	Yes

Factor	Justification	Indicator(s)	Issues	Influence decision?	Impact on outcomes?	Influenced by outcomes?	Use for matching?
	factors are potential proxies	OFSTED grade	May be an indicator of capacity to change, but OFSTED grade may be influenced by outcomes / value added	Yes	Yes	??	Yes
		Academy Status	May be an indicator of capacity, and willingness, to change	Yes	??	No	Possibly? Yes
		MAT membership	May be linked to capacity or impetus to improve.	Yes	Maybe	Yes	No
		Urban / rural	Level of competition between schools likely to influence decisions on grouping	Yes	No	No	Yes
Ethos / Values of SLT	MA (or setting) could be a values-driven decision	School policy documents Mission statement Soft Ofsted grades	Difficult to observe without cost Also likely endogenous variable	Yes	Yes	Maybe	No

APPENDIX E: Additional power calculations

Table E.1: Power calculations for primary analysis indicating MDES estimates. Recommended option highlighted (40 mixed-attainment, 80 setting).

Mixed attainment		30	30	40	40	45	50	50
	Setting	90	120	80	120	90	50	100
Correlation between pre-test measures (KS2 & other covariates) & post-test		N=120	N=150	N=120	N=160	N=135	N=100	N=150
Pupil-level	School-level	MDES						
0.65	0.33	0.222	0.215	0.204	0.192	0.192	0.211	0.182
0.75	0.38	0.216	0.210	0.199	0.187	0.188	0.206	0.178

Table E.2: Power calculations for sub-group analysis (low attaining students, FSM) indicating MDES estimates. Recommended option highlighted (40 mixed-attainment, 80 setting).

Mixed attainment		40	45
	Setting	80	90
Correlation between pre-test measures (KS2 & other covariates) & post-test		N=120	N=135
Pupil-level	School-level	MDES	
0.65	0.33	0.214	0.202
0.75	0.38	0.207	0.195

APPENDIX F: Grouping Study Matching and Simulated Response Exercise

To support our decisions regarding matching approach, we have conducted a matching and simulated response exercise to guide our approach set out in the study plan.

Propensity score estimation

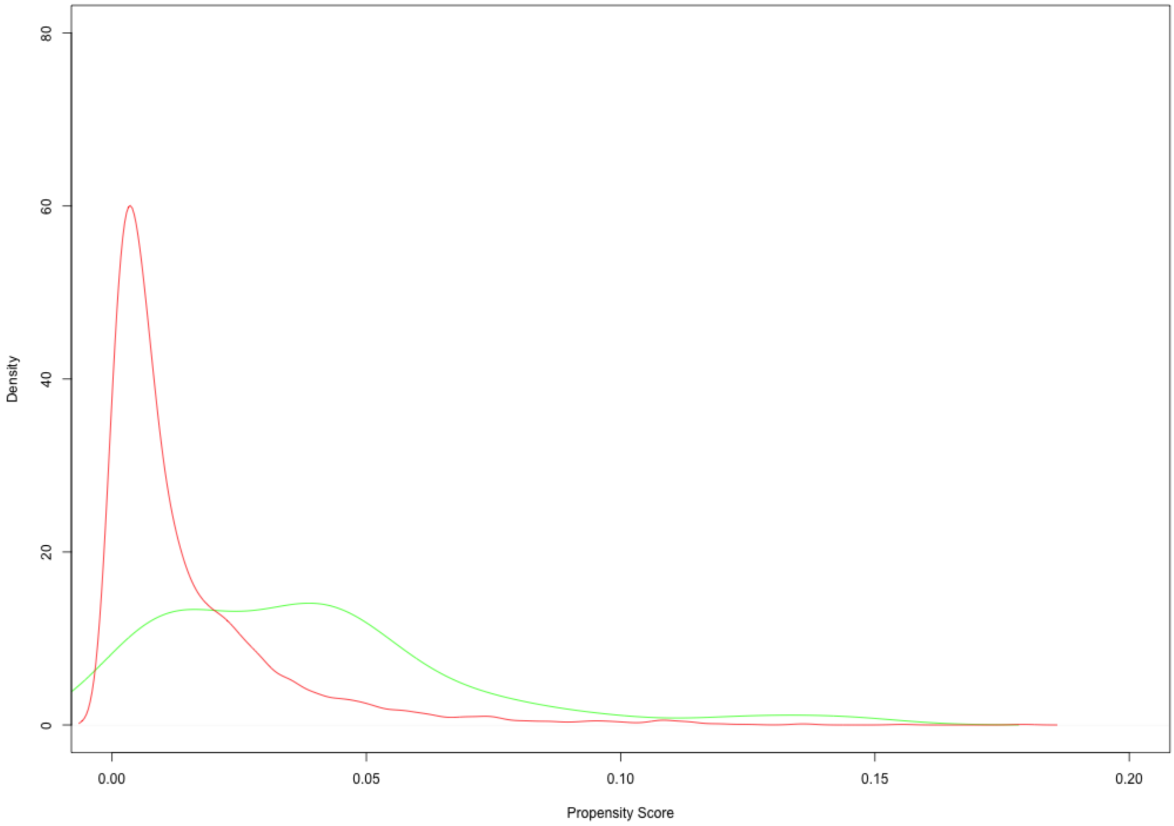
Matches are found based on a treatment propensity score estimated from the following generalised linear model with a probit link function:

$$\begin{aligned} probit(MixedAttain_i) &= \beta_0 + \beta_1 Prop.FSM_{i2017} \\ &+ \beta_2 Prop.HighPrior_{i2017} + \beta_3 Prop.LowPrior_{i2017} + \beta_4 Academy_{i2017} \\ &+ \beta_5 KS2_{i2018} + \beta_6 KS2_{i2017} + \beta_7 KS2_{i2016} + \beta_8 No.Pupils_{i2017} + \beta' \boldsymbol{IDACI_{i2017}} \\ &+ \beta' \boldsymbol{Ofsted_{i2017}} + \beta' \boldsymbol{Region_i} + \beta' \boldsymbol{Urban_i} + \varepsilon_{it} \end{aligned}$$

where $MixedAttain_i$ is our 0/1 indicator of whether school i is a mixed attainment school, $Prop.FSM_{i2017}$ is proportion of the school eligible for Free School Meals in 2017, $Prop.HighPrior_{i2017}$, is proportion of the school's cohort identified as "high attainment" in 2017, $Prop.LowPrior_{i2017}$, is proportion of the school's cohort identified as "low attainment" in 2017, $KS2_{it}$ is average KS2 score of the school's intake in year t, $No.Pupils_{i2017}$ is the size of the cohort, $\boldsymbol{IDACI_{i2017}}$ is a vector of binary variables indicating the quintile group into which the school falls in terms of the average Index of Deprivation Affecting Children and Infants (IDACI) of its intake, $\boldsymbol{Ofsted_i}$ is a vector of binary variables indicating the school's most recent Ofsted overall effectiveness grade, $\boldsymbol{Region_i}$ is a vector of binary variables indicating the government office region in which the school is located, $\boldsymbol{Urban_i}$ is a vector of binary variables indicating the urban/rural setting of the school, and ε_i is an idiosyncratic error term.

This model was based on iterative testing of model fit of available matching variables. Given the small number of treatment schools (i.e. 43 mixed attainment schools), there a risk of instability from use of more complicated models. However, this model produces substantially improved balance compared to a simpler version.

Figure F1. Density plot of distribution of propensity scores for mixed attainment and all other schools (unmatched)



Notes. Green line plots density for mixed attainment schools; red line plots density for all other schools.

Matching approach

For the purpose of this exercise, matches are found using an optimal matching algorithm with no replacement (in practice, allowing replacement makes no difference in this application, seemingly because there are plenty of potential matched comparators available) using the MatchIt package in R. We explore identifying both 20 and 25 potential matched comparators (from which recruitment will be attempted) for each mixed attainment school to explore the pros and cons of each.

We start with the hypothesis that 25 matches will have slightly worse match quality on average, however with 20 matches we are more likely not to successfully recruit one of these schools as part of this initial matching exercise (which would require identification of further potential matches, which would likely be of lower quality).

Match quality

In Figures 2 and 3 we plot an illustration of the distribution of the schools by their treatment status and whether they are identified as matched comparison schools. In Figure 2 this is plotted for 1:20 matching and in Figure 3 this is plotted for 1:25 matching. These plots demonstrate a good spread of matched comparison schools across the propensity score range of the matched treatment schools, which is not appreciably different depending on the number of matched schools identified.

In Table 1 we report the standardised differences between the treatment and potential matched comparison samples for 1:20 and 1:25 matching, along with these statistics for the unmatched sample.

Figure F2. Plot of propensity score distribution of schools by treatment and matching status (1:20 matching)

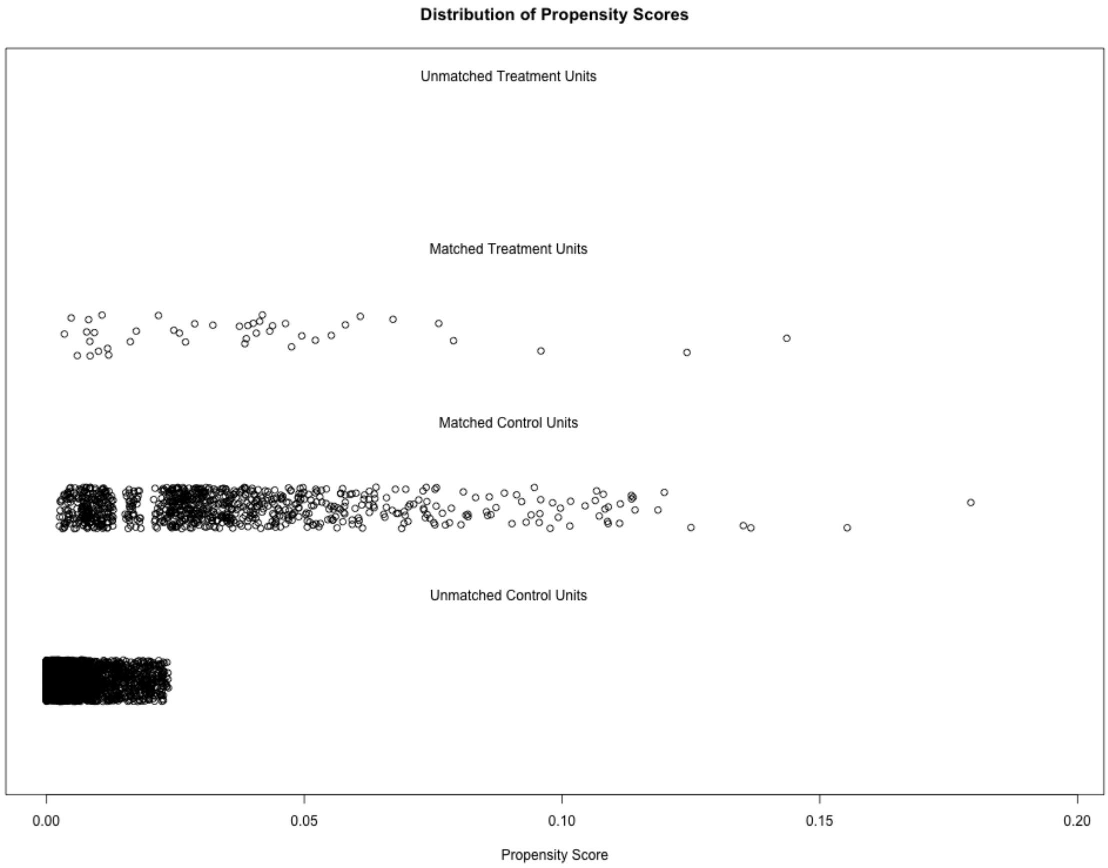


Figure F3. Plot of propensity score distribution of schools by treatment and matching status (1:25 matching)

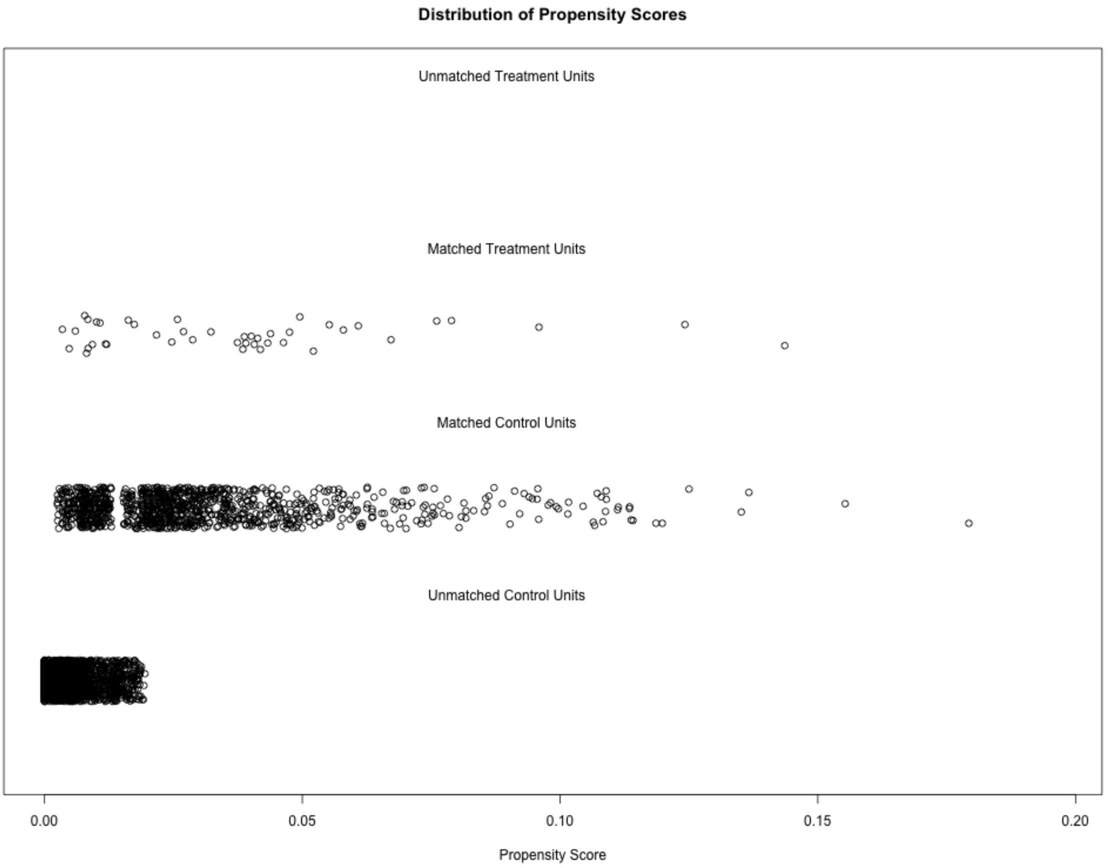


Table F1. Standardised differences in characteristics by matching methods

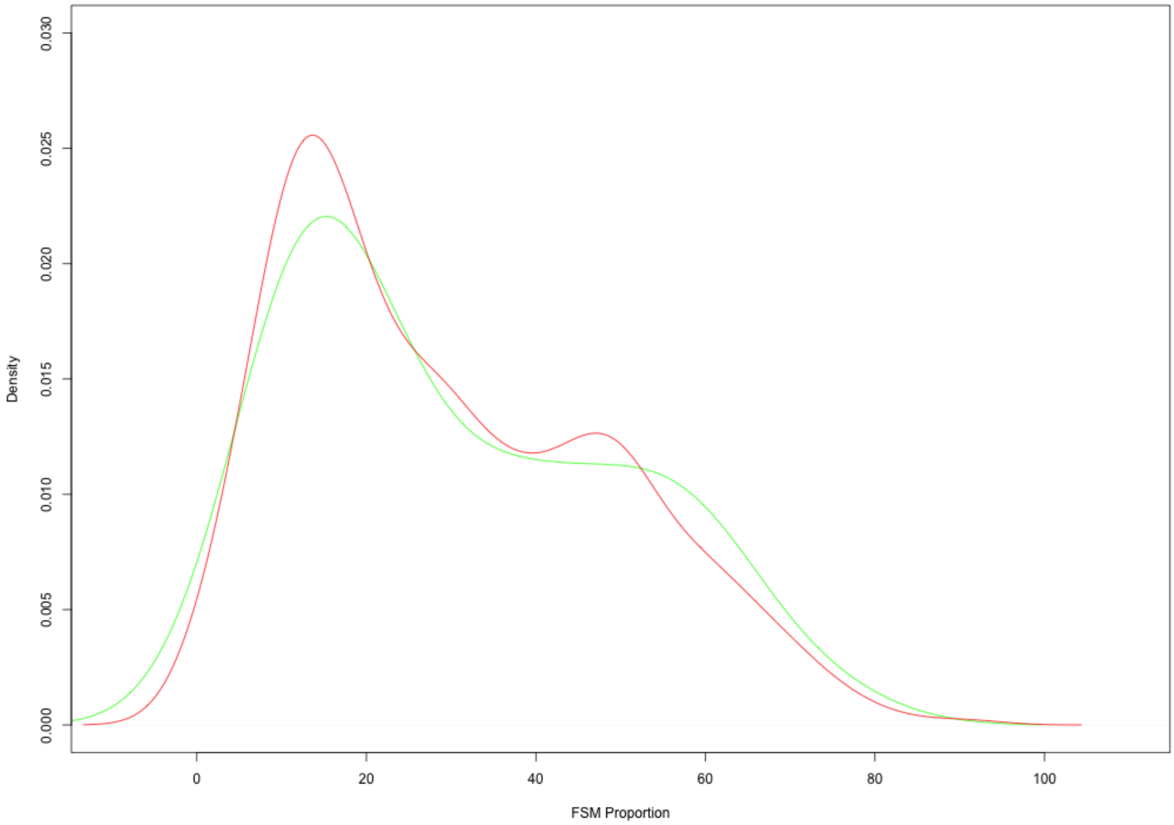
Characteristics	20 Matches	25 Matches	Unmatched
No. Pupils	0.10	0.08	0.28
Academy Proportion	-0.26	-0.28	-0.10
FSM Proportion	0.05	0.12	0.21
KS2 APS 2018	-0.01	-0.03	0.19
KS2 APS 2017	-0.02	-0.06	0.08
KS2 APS 2016	-0.02	-0.05	0.11
Low Attainers Prop.	-0.02	0.00	-0.21
High Attainers Prop.	-0.02	-0.04	0.18
IDACI Q1 Prop.	0.04	-0.01	0.32
IDACI Q2 Prop.	0.00	-0.01	-0.16
IDACI Q4 Prop.	-0.01	-0.01	-0.21
IDACI Q5 Prop.	0.03	0.08	0.27
Ofsted Outstanding Prop.	0.03	0.00	0.21
Ofsted Good Prop.	0.04	0.05	-0.06
Urban Setting Prop.	-0.05	-0.04	0.14
East Mids Prop.	0.04	0.03	0.12
East of England Prop.	0.09	0.02	-0.21
London Prop.	0.05	0.13	0.71
North West Prop.	-0.02	-0.05	-0.15
South East Prop.	0.00	-0.06	-0.11
South West Prop.	-0.04	-0.03	-0.11
West Mids Prop.	-0.10	-0.09	-0.07
Yorks/Humb Prop.	-0.01	-0.01	-0.08
Average	0.05	0.06	0.19

Notes. Reporting “Std. Diff” between treated and comparison schools identified by each matching method described. Standard differences calculated by dividing means by overall sample standard deviation. “Mean Abs. Std. Diff” = Mean absolute standard difference calculated across characteristics in table. IDACI Quintile 5 and Ofsted: Inadequate categories excluded since these are determined by the remainder of the other categories of this variable.

We consider the distribution of selected continuous characteristics and how this differs between treatment and matched comparison groups in the following plots. In Figures 4 and 5, we plot the density of the proportion of FSM pupils in the school for 1:20 and 1:25 matching, respectively. In Figures 6 and 7, we do the same for average KS2 points score on intake. In Figures 8 and 9, this is repeated for the proportion of the school's intake identified as low attainment by the DfE, while Figure 10 and 11 do the same for the proportion of the school's intake identified as high attainment.

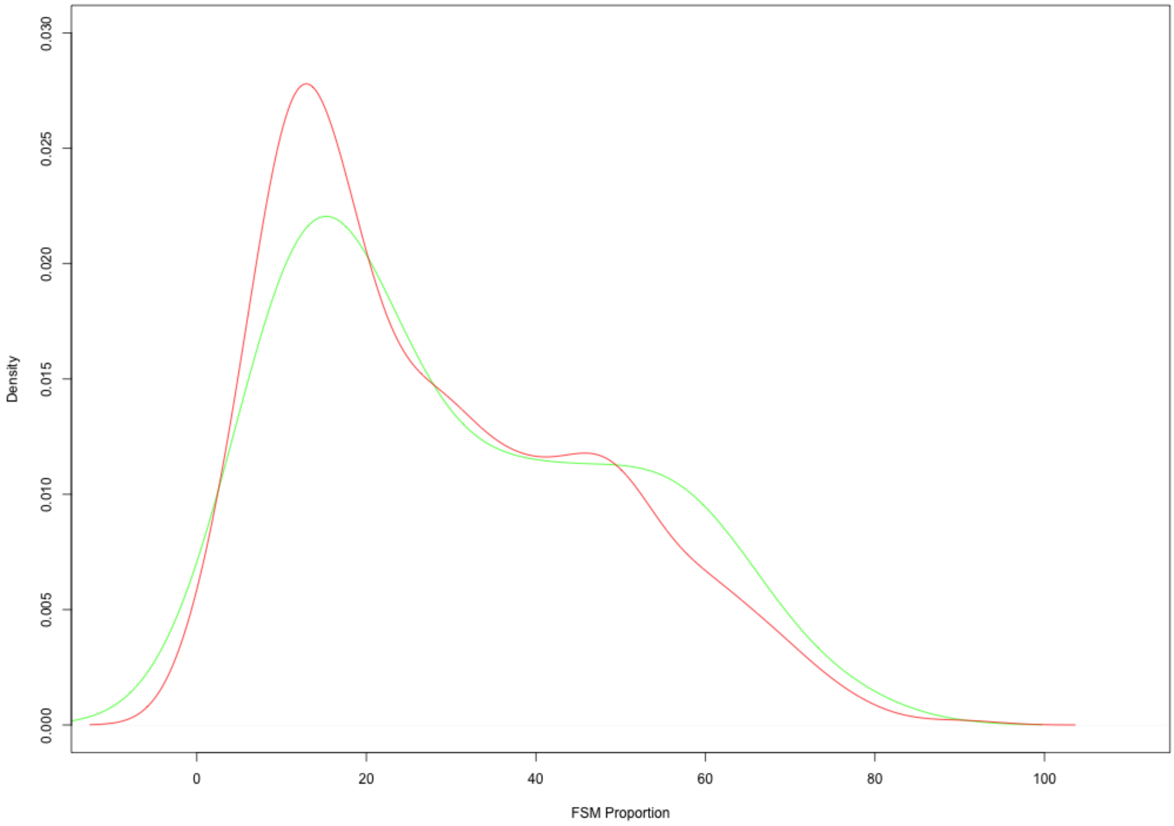
Overall, it is unclear that balance is substantially worse among the potential comparators in the case of 1:25 matching. Later in this document, we check that our simulated responses patterns among these potential comparators does not alter this picture.

Figure F4. Distribution of proportion of pupils identified as FSM in treatment and potential comparison groups (1:20 matching)



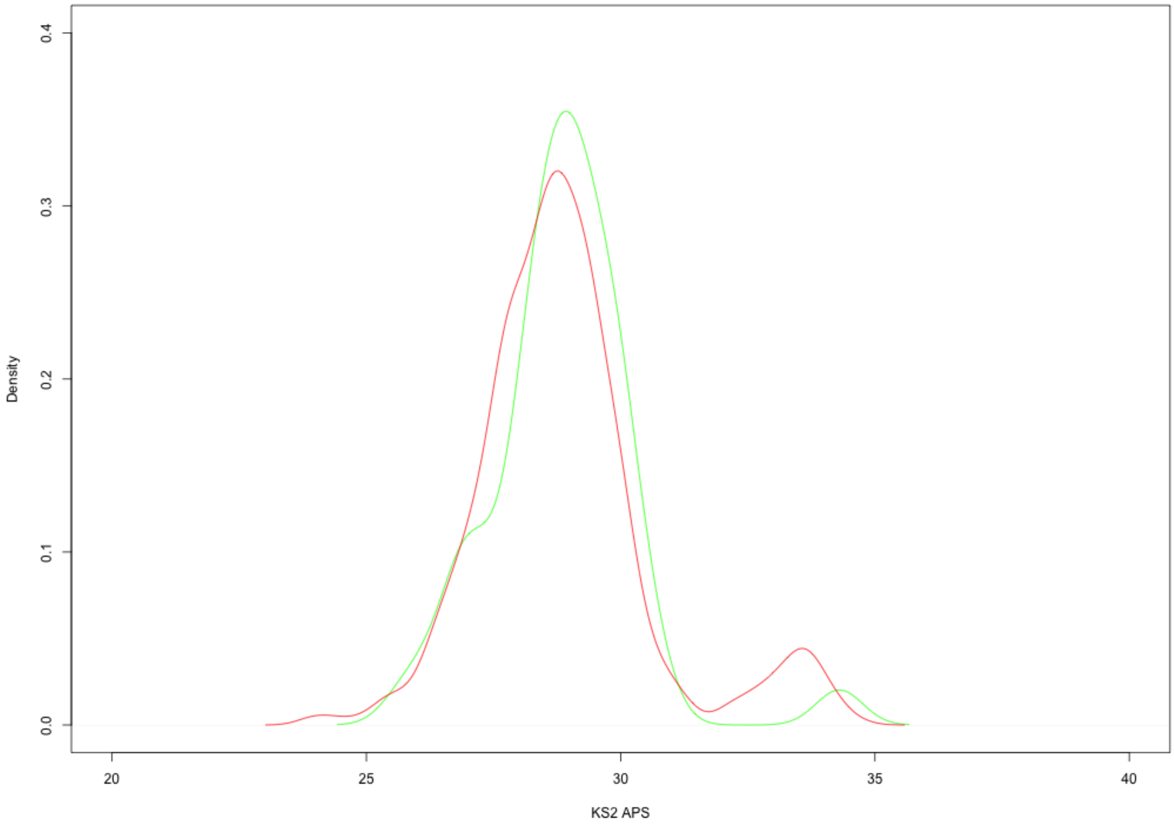
Notes. Kernel density plot of school FSM proportion for treated (green) and comparison (red) schools.

Figure F5. Distribution of proportion of pupils identified as FSM in treatment and potential comparison groups (1:25 matching)



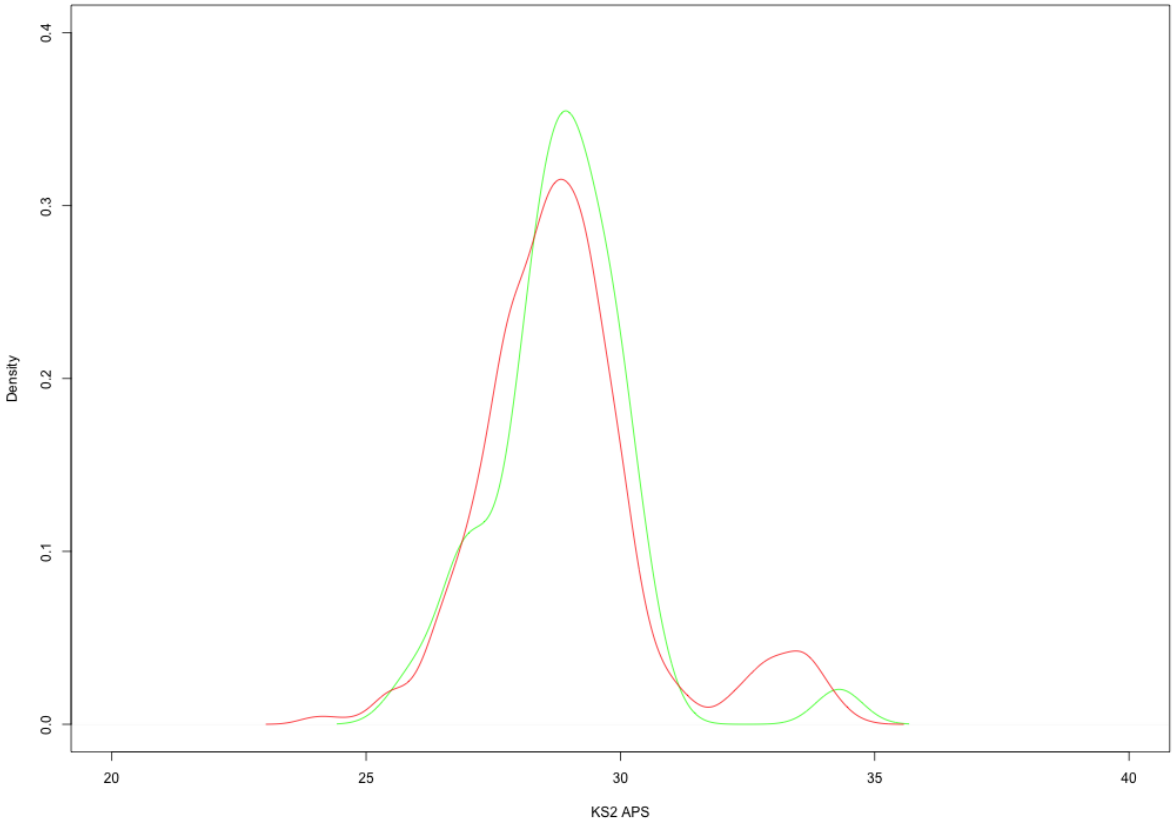
Notes. Kernel density plot of school FSM proportion for treated (green) and comparison (red) schools.

Figure F6. Distribution of average KS2 prior attainment in treatment and potential comparison groups (1:20 matching)



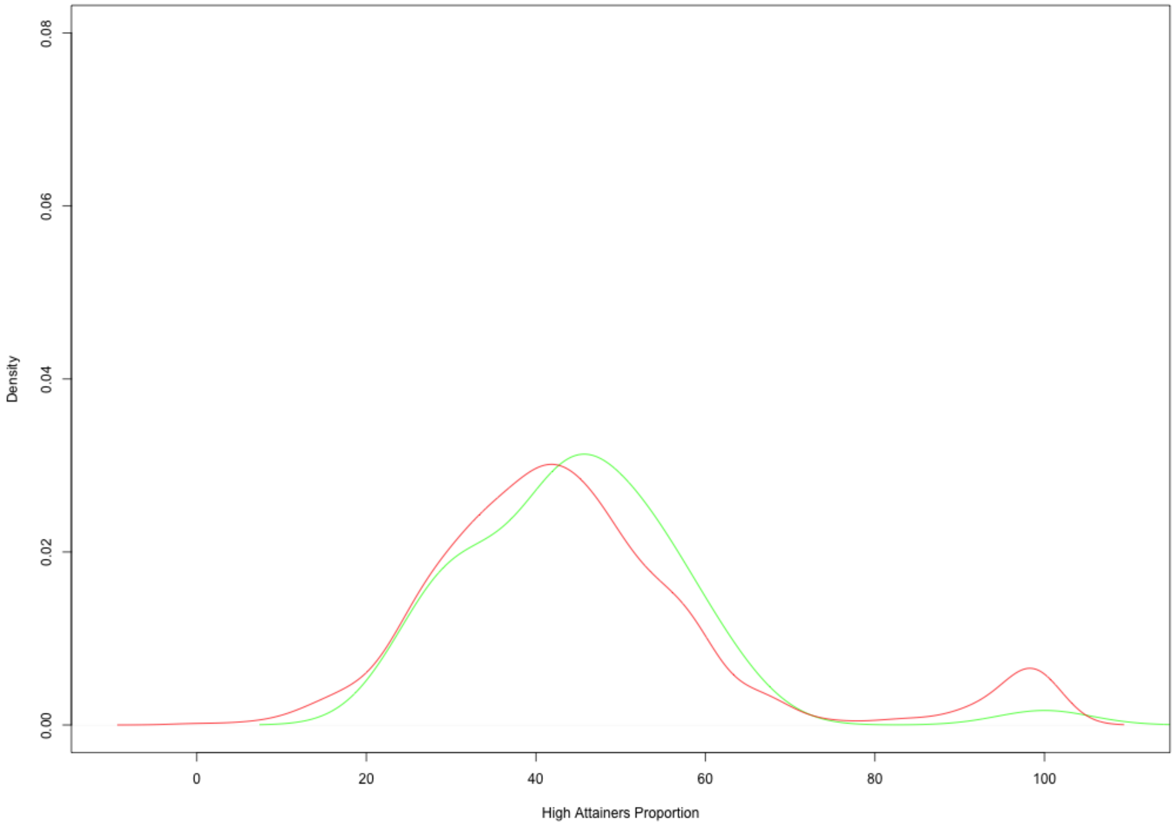
Notes. Kernel density plot of school average KS2 prior attainment for treated (green) and comparison (red) schools.

Figure F7. Distribution of average KS2 prior attainment in treatment and potential comparison groups (1:25 matching)



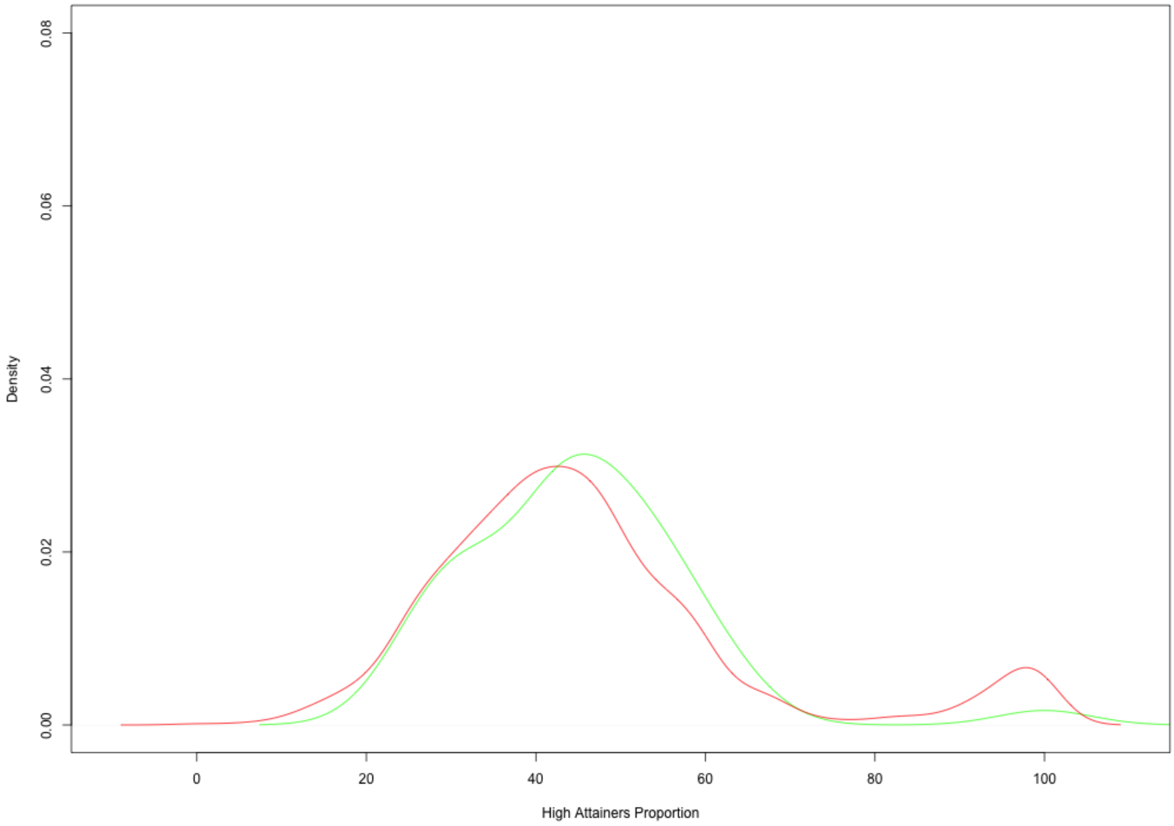
Notes. Kernel density plot of school average KS2 prior attainment for treated (green) and comparison (red) schools.

Figure F8. Distribution of proportion of pupils identified as high attainment on intake in treatment and potential comparison groups (1:20 matching)



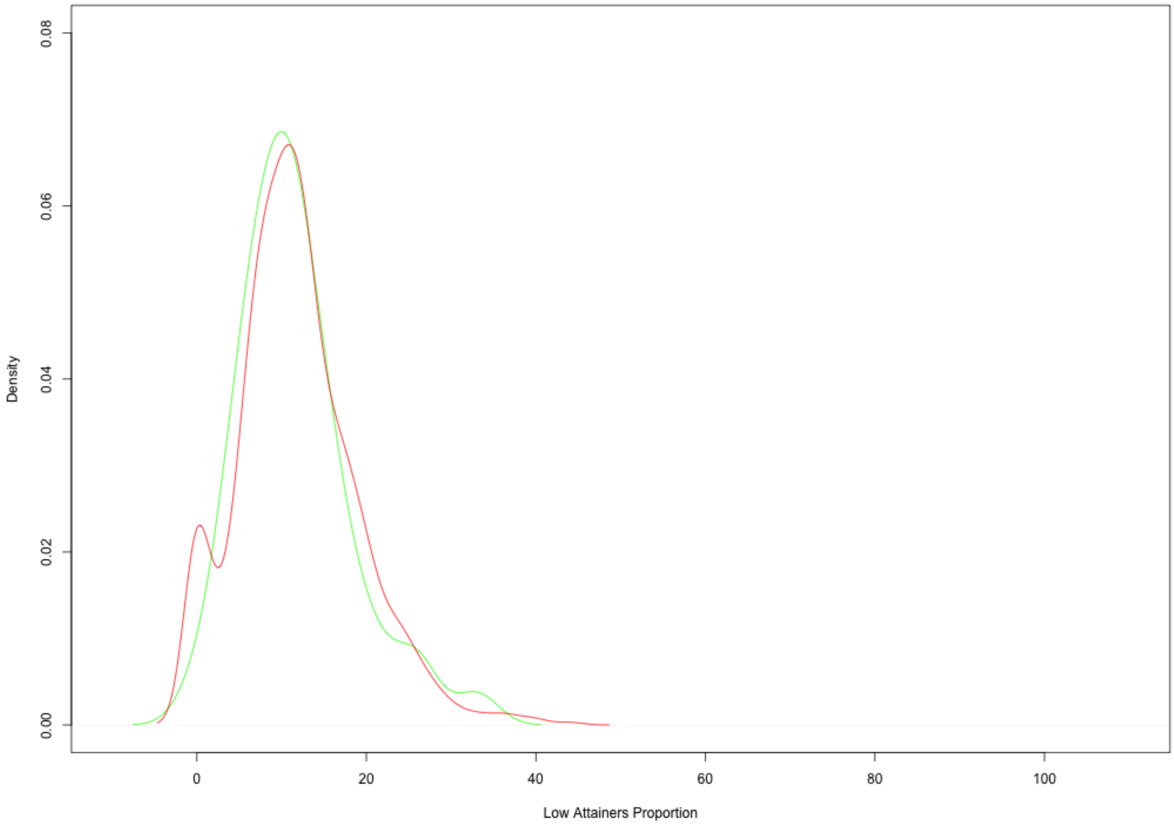
Notes. Kernel density plot of proportion of pupils identified as high attainment on intake by DfE for treated (green) and comparison (red) schools.

Figure F9. Distribution of proportion of pupils identified as high attainment on intake in treatment and potential comparison groups (1:25 matching)



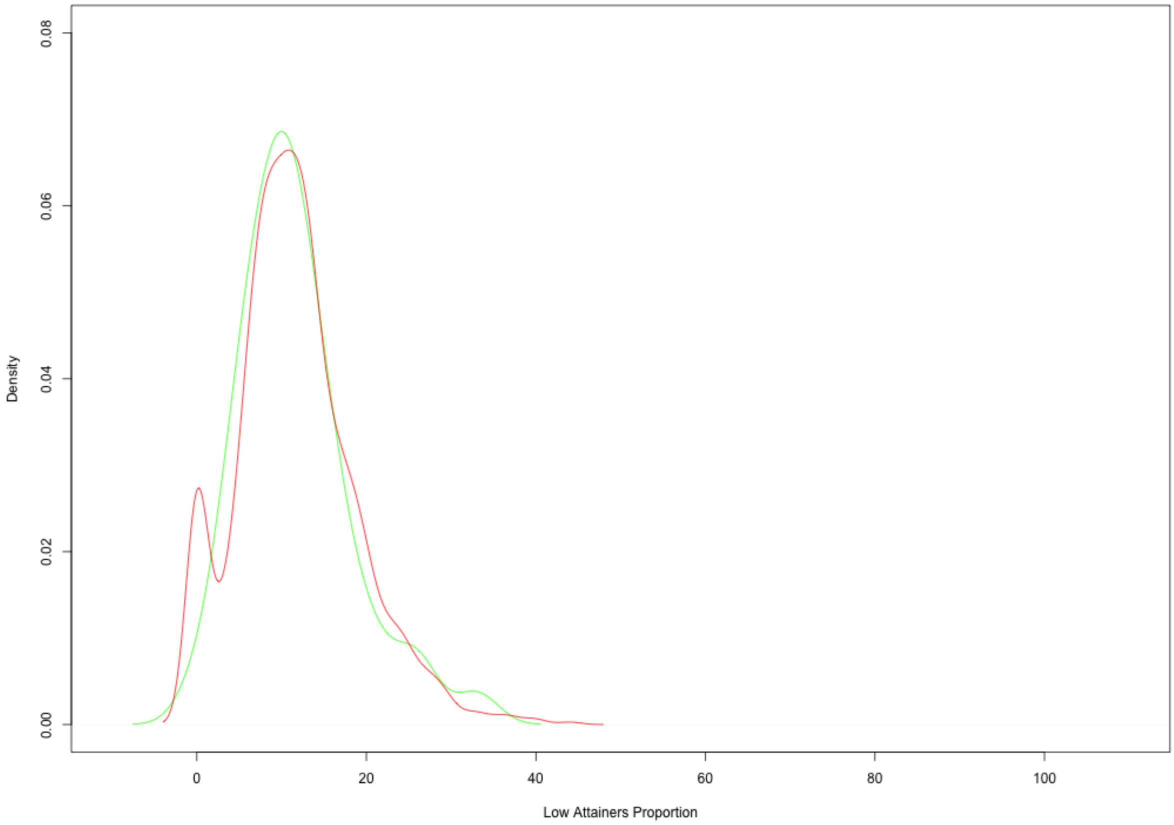
Notes. Kernel density plot of proportion of pupils identified as high attainment on intake by DfE for treated (green) and comparison (red) schools.

Figure F10. Distribution of proportion of pupils identified as low attainment on intake in treatment and potential comparison groups (1:20 matching)



Notes. Kernel density plot of proportion of pupils identified as low attainment on intake by DfE for treated (green) and comparison (red) schools.

Figure F11. Distribution of proportion of pupils identified as low attainment on intake in treatment and potential comparison groups (1:25 matching)



Notes. Kernel density plot of proportion of pupils identified as low attainment on intake by DfE for treated (green) and comparison (red) schools.

Simulated response: failure to recruit

For this study, it is necessary to actively recruit schools, since the data needed for this evaluation cannot be extracted entirely from administrative datasets. As such, we carry out a basic simulation of this recruitment process, as follows.

In 1000 simulations, we assign all schools identified as matched comparators a response probability drawn randomly from a uniform distribution between 0 and 1. We assume that those schools with response probabilities above 0.8 will join the study if contacted to do so. We then treat as recruited the two schools with a response probability above 0.8 with the smallest difference in propensity score from its respective treated school for each mixed attainment school. In doing so, we mimic the recruitment process in which we will work systematically through a matched recruitment list for each mixed attainment school sorted in the same way, continuing until two schools have been recruited or the list has been exhausted.

In the same way, in some simulations it is the case that there is only one school, or even no schools, with a response probability about 0.8 in the potential matched comparator list for each mixed attainment school. This is more likely to be the case when only 20 potential matched comparators are identified, rather than 25, which we demonstrate with the following analysis.

Table F2. Proportion of simulations in which the column title number of schools achieves the row title number of responses – 1:20 matching

	0	1	2	3	4					
No responses	0.612	0.315	0.065	0.007	0.001					
	0	1	2	3	4	5	6	7	8	
One responder	0.082	0.208	0.27	0.212	0.135	0.055	0.026	0.01	0.002	
	34	35	36	37	38	39	40	41	42	43
Two responders	0.002	0.004	0.024	0.045	0.079	0.178	0.241	0.215	0.163	0.049

Table F3. Proportion of simulations in which the column title number of schools achieves the row title number of responses – 1:25 matching

	0	1	2	3		
No responses	0.841	0.147	0.01	0.002		
	0	1	2	3	4	5
One responder	0.328	0.394	0.183	0.073	0.02	0.002
	38	39	40	41	42	43
Two responders	0.006	0.036	0.094	0.208	0.37	0.286

We note that there are limitations to this approach. The 0.8 probability cut off is an assumption (based on an estimated recruitment probability of 0.2) and this simple process makes the assumption of no correlation between school characteristics and response probability. Note, however, that because the list is worked through systematically from the potential matched comparator for each mixed attainment school with the smallest difference in propensity scores from the treated school to the one with the largest.

Simulated response: effects on imbalance

Response patterns will also have an effect on imbalance, relative to the matched sample. We can use our simulations to explore these. Given the recruitment strategy we intend to follow, i.e. prioritising those with the most similar propensity scores to the mixed attainment schools, our simulations suggest that, if anything this process is likely to reduce imbalance relative to the initial matched sample. This is perhaps unsurprising, given the large size of the matched sample that we generate.

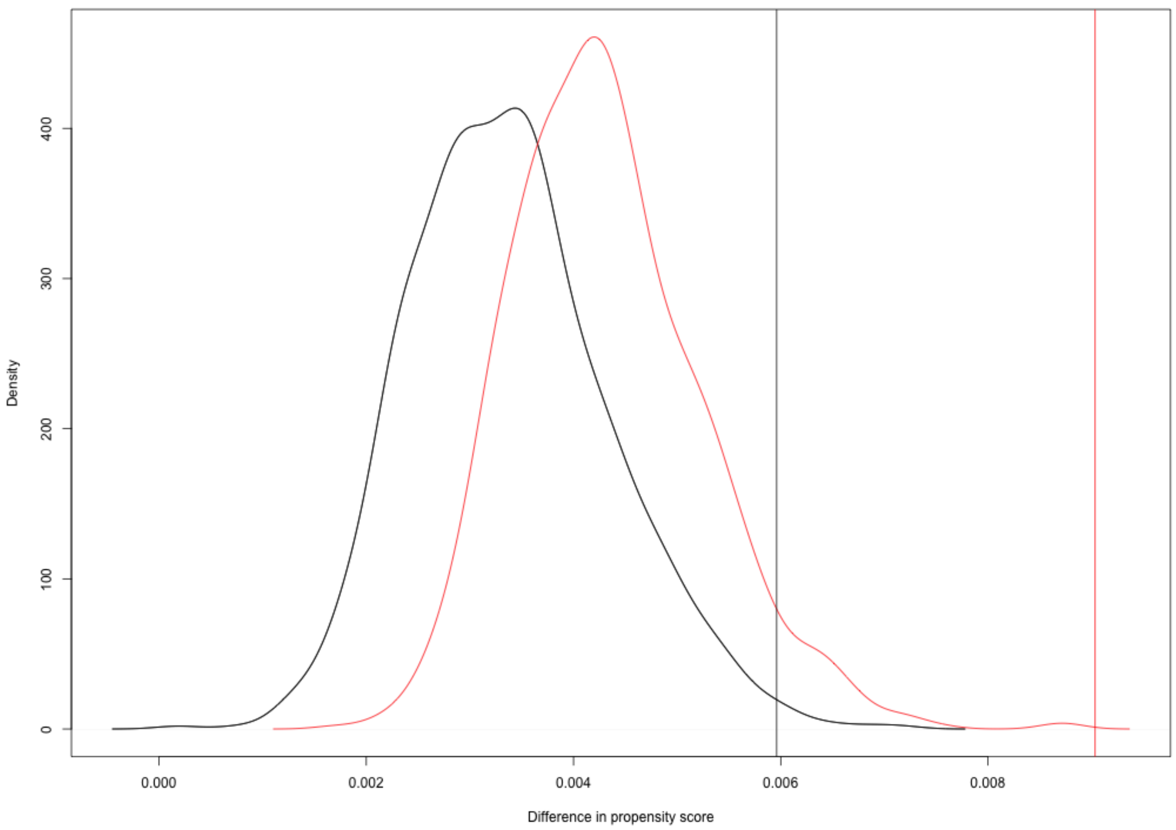
Figure 12 plots the distribution of the difference in propensity score between treatment and simulated recruited schools by whether 1:20 or 1:25 matching was carried out in the initial matching process. Figures 13 to 18 repeat this but for standardised difference measures of imbalance in key characteristics. Overall, we judge that it is not particularly the case that there is systematically better balance between the simulated recruited samples and the mixed attainment schools in the case of 1:20 matching compared to 1:25 matching.

Conclusions

Based on the above analysis, we are minded to adopt a 1:25 approach to matching given the increased probability that this will lead to successful recruitment within the initial matched sample, without evidence of this compromising the match quality of the finally recruited sample.

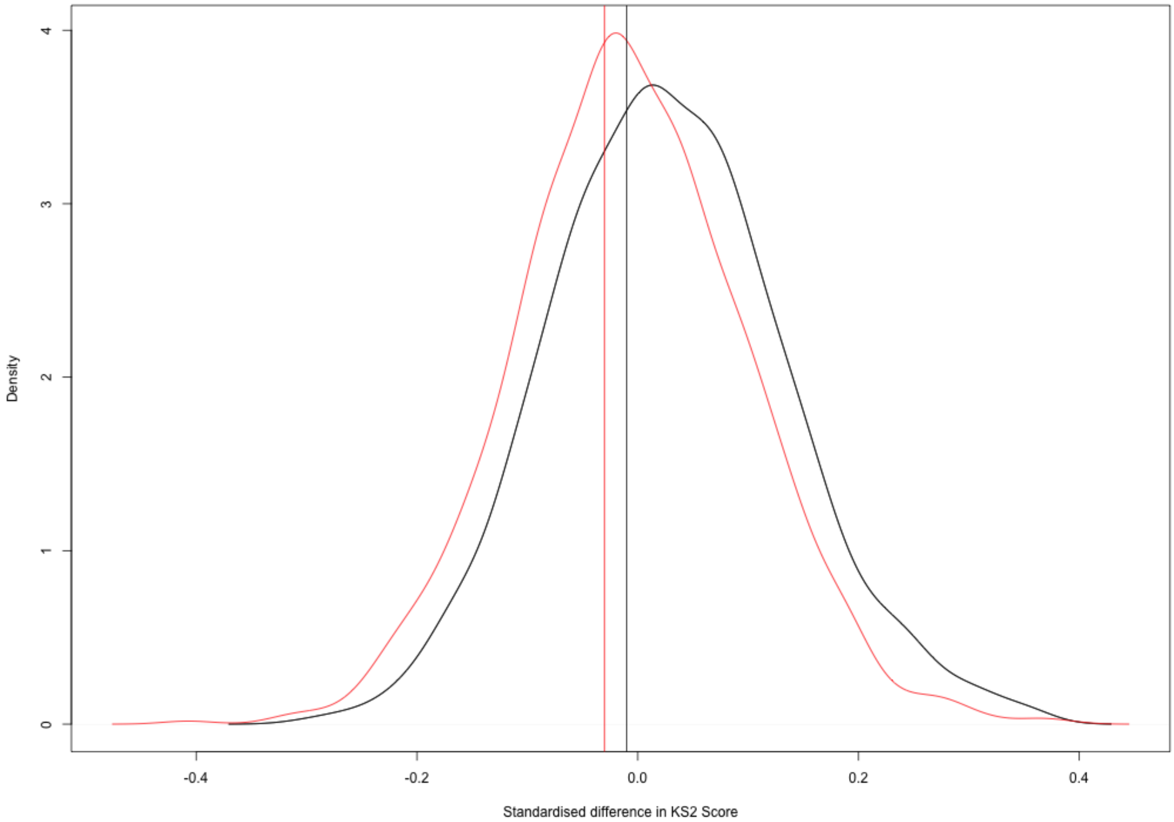
The exercise has also emphasised the particular importance of estimating propensity scores using a parsimonious model as part of this project, given the small number of mixed attainment schools available. This makes prioritisation of characteristics on which we need to achieve a good match to have confidence in the estimates from this project a particularly key issue.

Figure F12. Simulated density of imbalance in propensity score measures after response: comparing 1:20 (black) to 1:25 (red) matching



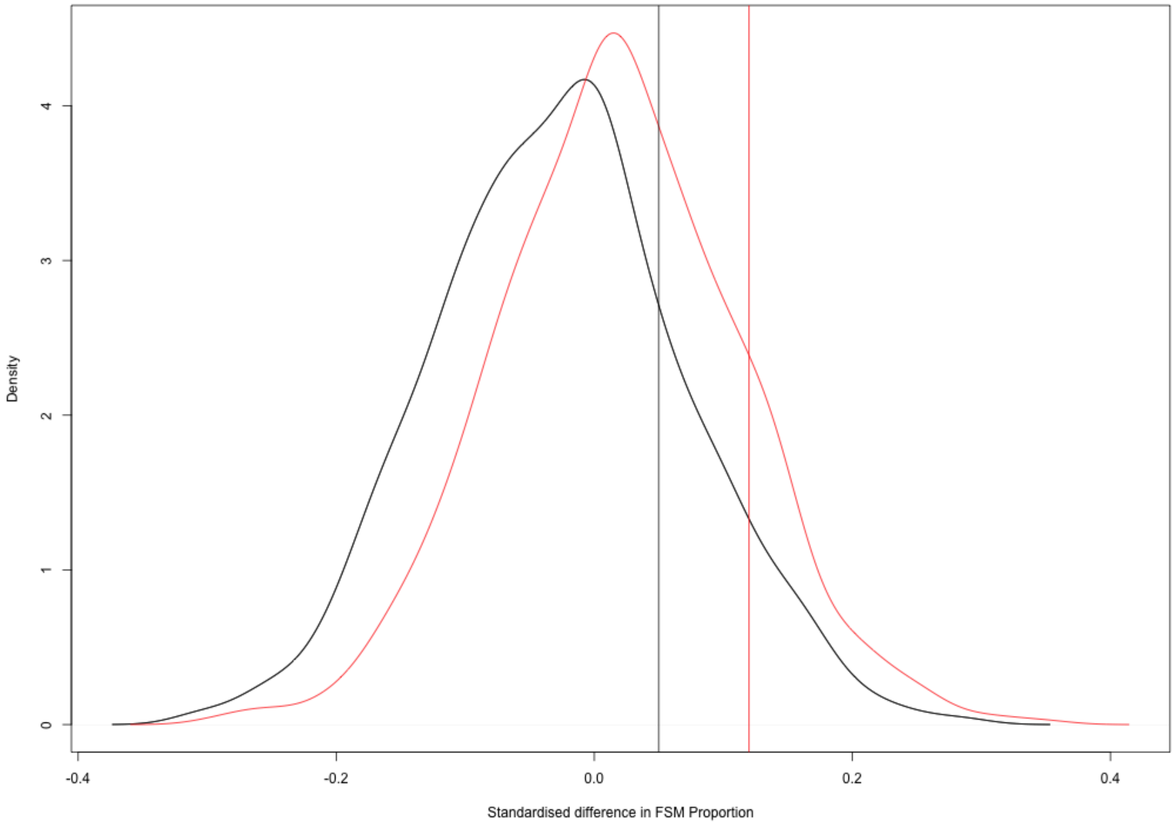
Notes. Density plots showing difference in propensity score between treated and simulated responders among the matched comparison sample. Simulations based on 1:20 matched sample plotted in black; simulations based on 1:25 matched sample plotted in red. Simple imbalance from full sample of potential matches plotted as vertical line (1:20 matched sample plotted in black; 1:25 matched sample plotted in red).

Figure F13. Simulated density of standardised imbalance in KS2 score after response: comparing 1:20 (black) to 1:25 (red) matching



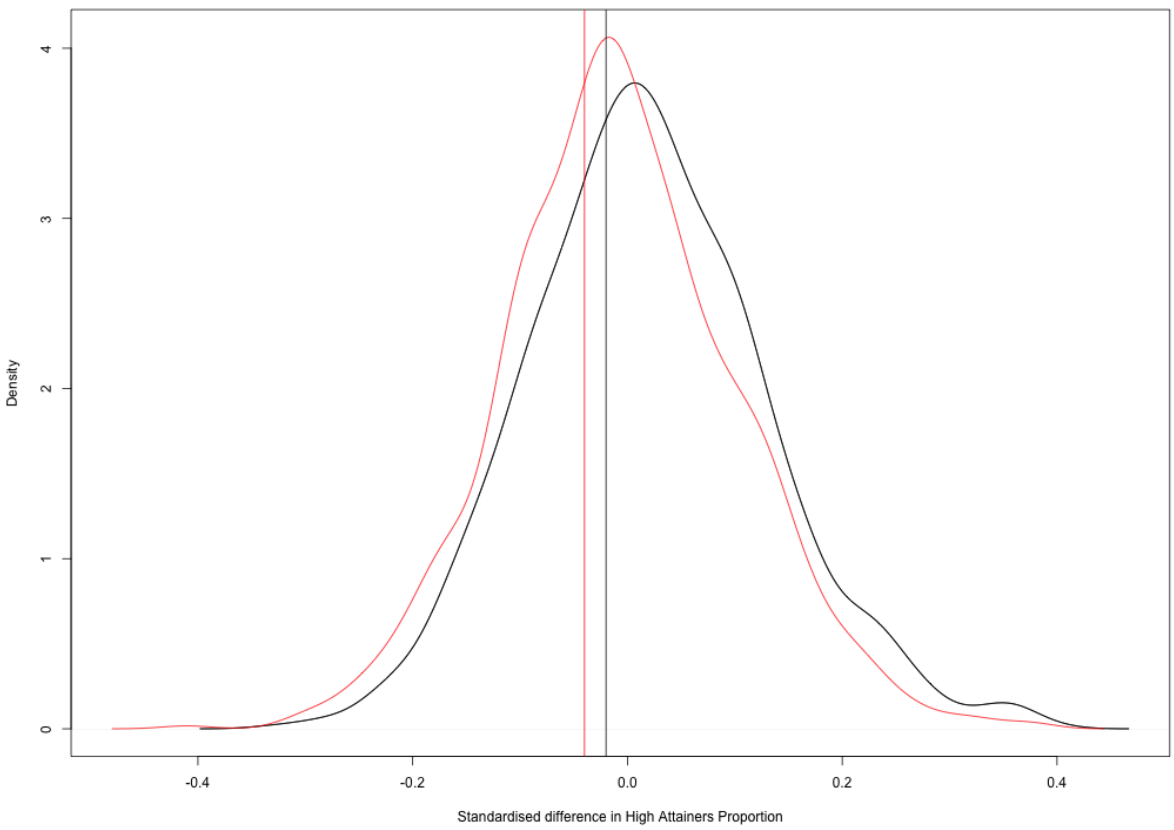
Notes. Density plots showing standardised difference in KS2 score between treated and simulated responders among the matched comparison sample. Simulations based on 1:20 matched sample plotted in black; simulations based on 1:25 matched sample plotted in red. Simple imbalance from full sample of potential matches plotted as vertical line (1:20 matched sample plotted in black; 1:25 matched sample plotted in red).

Figure F14. Simulated density of standardised imbalance in FSM proportion after response: comparing 1:20 (black) to 1:25 (red) matching



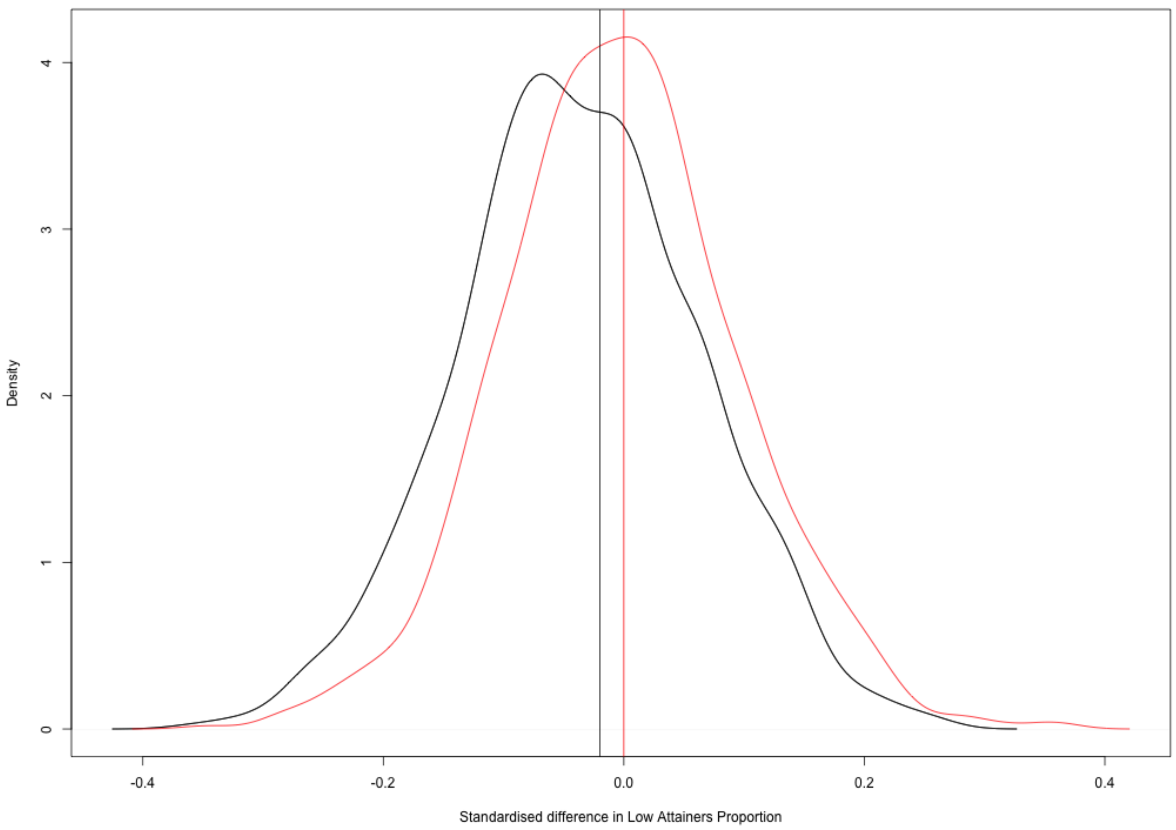
Notes. Density plots showing standardised difference in FSM proportion between treated and simulated responders among the matched comparison sample. Simulations based on 1:20 matched sample plotted in black; simulations based on 1:25 matched sample plotted in red. Simple imbalance from full sample of potential matches plotted as vertical line (1:20 matched sample plotted in black; 1:25 matched sample plotted in red).

Figure F15. Simulated density of standardised imbalance in proportion of high attainers after response: comparing 1:20 (black) to 1:25 (red) matching



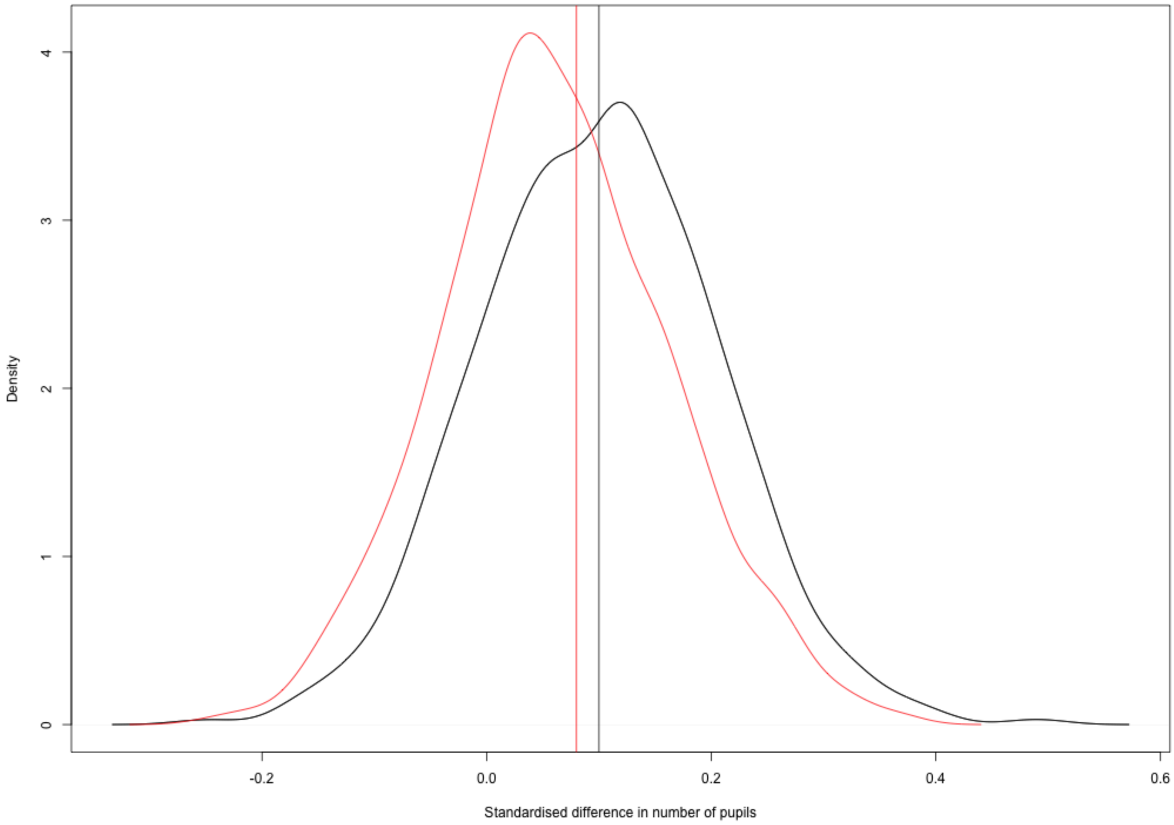
Notes. Density plots showing standardised difference in proportion of high attainers between treated and simulated responders among the matched comparison sample. Simulations based on 1:20 matched sample plotted in black; simulations based on 1:25 matched sample plotted in red. Simple imbalance from full sample of potential matches plotted as vertical line (1:20 matched sample plotted in black; 1:25 matched sample plotted in red).

Figure F16. Simulated density of standardised imbalance in proportion of low attainers after response: comparing 1:20 (black) to 1:25 (red) matching



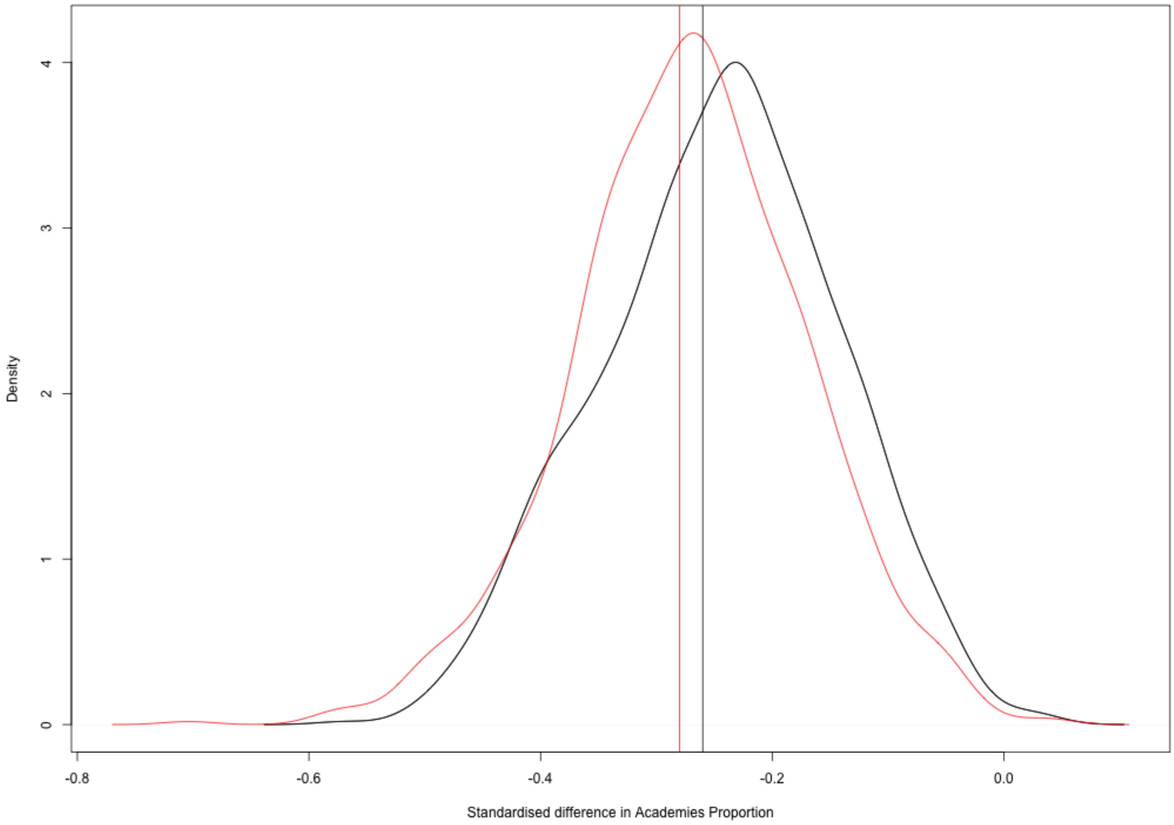
Notes. Density plots showing standardised difference in proportion of low attainers between treated and simulated responders among the matched comparison sample. Simulations based on 1:20 matched sample plotted in black; simulations based on 1:25 matched sample plotted in red. Simple imbalance from full sample of potential matches plotted as vertical line (1:20 matched sample plotted in black; 1:25 matched sample plotted in red).

Figure F17. Simulated density of standardised imbalance in number of pupils after response: comparing 1:20 (black) to 1:25 (red) matching



Notes. Density plots showing standardised difference in number of pupils between treated and simulated responders among the matched comparison sample. Simulations based on 1:20 matched sample plotted in black; simulations based on 1:25 matched sample plotted in red. Simple imbalance from full sample of potential matches plotted as vertical line (1:20 matched sample plotted in black; 1:25 matched sample plotted in red).

Figure F18. Simulated density of standardised imbalance in proportion of academies after response: comparing 1:20 (black) to 1:25 (red) matching



Notes. Density plots showing standardised difference in proportion of academies between treated and simulated responders among the matched comparison sample. Simulations based on 1:20 matched sample plotted in black; simulations based on 1:25 matched sample plotted in red. Simple imbalance from full sample of potential matches plotted as vertical line (1:20 matched sample plotted in black; 1:25 matched sample plotted in red).

APPENDIX G: Steering Board Terms of Reference

Role

The Steering Board will provide independent oversight of the research and advice on the research design and methods. Specifically, the Steering Board will discuss the project study plan, statistical analysis plan and the final report; and make recommendations to EEF at each of this stages.

Membership

The members of the Steering Board will be appointed by the Education Endowment Foundation to serve over the lifetime of the project (2018 until 2022). Members will be chosen so that there is appropriate breadth and balance of methodological expertise covering statistical and qualitative methods. The initial membership will be: Rob Coe (EEF); Stefan Speckesser, Heather Rolfe (NIESR); Bronwen Maxwell (Sheffield Hallam University); and Richard Dorsett (University of Westminster).

Reporting

The Steering Board will report to the EEF

Frequency of meetings

It is anticipated that the Steering Board will have five meetings: in November 2018, March 2019, May 2019, July 2020 and November 2021. The EEF, or the project research team, may ask for advice outside of these meetings on an ad hoc basis and, if necessary, the EEF may call an additional meeting of the group.

Ways of working

The Steering Board will receive reports from the project research team. Notes of meetings will be taken and circulated by EEF staff.

EEF staff and project research team members will be in attendance at meetings.

APPENDIX H: Head of Mathematics Survey

Table H1. Head of Mathematics survey questions.

About grouping in maths at your school	
Do you have a nurture group in Year 7 for maths?	Yes/No
Briefly, why do you have a nurture group?	[free text]
How many maths classes are there in Year 7?	Options 2-16
How many different ability or attainment level sets are there? 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.	Options 2-12
Briefly, what is the main reason why you use your current approach to grouping for maths in Year 7?	[free text]
Whose decision is it to group in this way?	Head of Maths Senior Leadership Team Multi-Academy Trust Governors/Trustees By mutual agreement within the maths department Other (please specify)
If you would like to provide any additional information about grouping in maths at your school, please do so here	[free text]
Grouping information	
In some schools, pupils are not allocated to their maths sets straight away at the beginning of the autumn term. When were pupils in your school allocated to their maths groups?	Straight away In September, but after the beginning of term October After October half term Later in the school year
Which sources of information were/will be used to allocate students to Year 7 maths groups this year? Please tick all that apply, but exclude exceptional cases such as students arriving from overseas.	National Curriculum Key Stage 2 test results National Curriculum Key Stage 2 teacher assessments A commercially-available test (e.g. CATS/MidYIS) The school's own test of student attainment Teacher judgements of students' abilities Teacher observations of student behaviour A parent's judgement of their child's ability Information from students' feeder schools

	The results of random allocation of students Other (please specify)
When are commercially-available tests conducted?	Before the start of Year 7 (e.g. 'moving up' day) In the first two weeks of term Later in the first half-term After the autumn half-term holiday
When is the school's own test conducted?	Before the start of Year 7 (e.g. 'moving up' day) In the first two weeks of term Later in the first half-term After the autumn half-term holiday
Briefly, what is the main reason why you use these sources of information for allocating pupils to classes?	[free text]
[Setting schools only] How did you allocate Year 7 teachers to maths classes for this school year 2022-23? Please tick all that apply.	Subject expert teachers tend to be placed with the higher ability classes Subject expert teachers tend to be placed with the lower ability classes Teachers are rotated between ability classes and year groups from year to year Teachers are randomly allocated to different classes I have little choice over how teachers are allocated to Year 7 classes Teachers are allocated to classes based on their personal preferences Teachers are allocated to classes based on their particular strengths Experienced teachers tend to be placed with the higher ability classes Experienced teachers tend to be placed with the lower ability classes Experienced teachers tend to be placed with the middle classes Inexperienced teachers tend to be placed with the higher ability classes Inexperienced teachers tend to be placed with the lower ability classes Inexperienced teachers tend to be placed with the middle classes Teachers are allocated first to GCSE/A level classes and later to Year 7 Teachers are allocated so that they have a balanced set of classes (and ages) Other (please specify)

[Setting schools only] Which teacher strengths are taken into consideration when allocating teachers to groups?	Teaching pupils with SEND Teaching pupils with high attainment Managing challenging behaviour Building relationships with vulnerable pupils Subject knowledge Other strengths (please specify)
[Mixed attainment schools only] How did you allocate Year 7 teachers to maths classes for this school year 2022-23? Please tick all that apply.	Teachers are randomly allocated to different classes I have little choice over how teachers are allocated to Year 7 classes Teachers are allocated to classes based on their personal preferences Teachers are allocated to classes based on their particular strengths Teachers are allocated first to GCSE/A level classes and later to Year 7 Other

APPENDIX I: TRU Scoring rubric for case study field visits (an extract)

Date:	School code:		Teacher:	Observers:
Time:	Year Group:	Set/ MA:	No. of pupils:	
Notes for interview (transfer to interview schedule)				
Note: Text on white background refers to whole class and text on grey background refers to small group				

Domain Mathematical Content	Agreed evidence observed (Whole class)	Agreed evidence (Small group)
1 Classroom activities are unfocused or skills--oriented, lacking opportunities for engagement with key grade level content (as specified in the Common Core Standards)		
1 The mathematics discussed is not at grade level; **OR** discussions are aimed at "answer getting." Explanations, if they appear, are largely procedural.		

2 Activities are at grade level but are primarily skills---oriented, with few opportunities for making connections (e.g., between procedures and concepts) or for mathematical coherence (see glossary).		
2 Discussions are at grade level but are primarily skills---oriented, with few opportunities for making connections (e.g., between procedures and concepts) or for mathematical coherence (see glossary).		
3 Classroom activities support meaningful connections between procedures, concepts and contexts (where appropriate) and provide opportunities for building a coherent view of mathematics.		
3 Explanation of and justification for central grade level mathematical ideas is coherent.		

APPENDIX J: Student paired interview after the observed lesson

TRU: Mathematics content

W/C 1. How did you find the lesson today?

Probe: Tell me what you liked/disliked about this lesson?

2. What was the lesson about?

Probe: Was the maths similar to any maths in previous maths lessons?

TRU: Cognitive Demand

General 3. What did you learn today that you didn't know about before today?

W/C 4. Which parts in the today's lesson did you find most challenging?

Prompt: think about a task you worked on in the lesson

Probe: Did you manage to work it out? How?
Which did you find the easiest?
Learnt the most from?

5. What were you expected to do to be successful in your maths lesson today?

Prompts: give correct answers, solve problems using the steps the teacher showed me, explain my thinking, listen and make sense of other students' reasoning

Probe: Is that different to other maths lessons?
What would other students in your class say? Which ones? How do you know?

S/G [If no paired/group work, ask if they ever do it. If yes, ask how they like it, then ask 7 and 8 only and skip remaining S/G questions]

6. What was the purpose of your small group work today?

Prompts: solve a problem using the steps our teacher showed us, check with my group members to see if my answers were correct, investigate a mathematical idea, share the different ways students in my group were solving a problem

Probe: Is that what small group work in your maths lessons is usually for?

7. Do all groups usually work on the same/different task/activity, like today?

Prompt: we usually all have the same task, the teacher gives some groups different tasks, we choose the task we want to work on

8. Do you usually work with the same students for group work?

Prompt: the teacher tells us who to work with, we always sit in the same groups, and why do you think that is

9. How useful did you find other students' ideas for your learning during small group/paired work today? And other students, how useful do you think your ideas were for their learning?

10. Did your group struggle with the task today? Y: What did you do help yourselves?

Prompt: ask the teacher, ask someone in a different group, listen to each other's ideas, try out different ways and compare them, wait for the teacher to come

Probe: Is that what normally happens?

TRU: Access

W/C 11. How comfortable were you sharing your ideas in the lesson today?
Probe: Is that usually how you feel in your maths lessons (e.g., comfortable or not sharing)?

Probe: For 'Yes' – Can you give an example of something that you shared in the lesson today?
Prompts: How did the teacher or other students react to what you shared? How was what you shared used by the teacher or other students?

Probe: For 'No' – Can you give an example of something that you shared (or wanted to share?) in the lesson today but felt uncomfortable doing so? What was your reason for feeling uncomfortable?
Prompt: How did you think your teacher/ or other students might react if you shared your thinking?
12. Did listening to other students in today's whole class discussion help make your thinking better?
Prompt (if yes): Can you give an example [thinking about an activity or a problem in today's lesson]?
Prompt (if no): Did anything another student said in the lesson confuse you? Can you give an example [thinking about an activity or a problem in today's lesson]?

S/G 13. Did you participate in your small group today?
Probe: Is that what it's usually like in your group work?

14. What happens in small group/paired work if you disagree with another student in your group?
Prompt: Did this happen today? How did you work through it? Share your answer and your reason? Compare your reason with other students. Check the calculation again using the teacher's way? Say nothing and assume you were wrong? Say nothing and assume you were right? It would depend on who the other student was.

TRU: Agency, ownership and identity

W/C 15. Which students talked the most in the lesson today?
Prompts: students who knew the right answer; who had ideas to share; had questions to ask
Probe: Is that what it's usually like in your maths lessons?
Which students usually share their ideas more / less than others? Ask more questions than others? Students that are good or bad at maths?
And you, when do you talk in maths lessons?

16. What do you think was that the purpose of today's whole class discussion?
Prompts: share how we solved problems using the steps our teacher showed us, learn the way the teacher showed us to solve the problem, learn different ways that work to solve a problem from other students, share a mathematical idea we came up with on our own, check to see if our answers are correct
Probe: What would happen if *you* had a different answer from the teacher or another student?

S/G 17. How comfortable were you sharing your thinking in your small group/with your partner today?

Probe: Is that usually how you feel in your group work?

Probe: Y – How did other students react to what you shared today?

Prompt: Did the students agree with what you shared? Did your group use your idea?
Y – [if the teacher was with the group] How did the teacher react to what you shared?

N – What was your reason for not feeling comfortable to share your thinking in today's group work?
Prompt: How did you think your group might react if you shared your thinking?

18. Who talked the most in your group today?

Prompt: Is it because of the task you worked on or is it not activity-related? Some students talk more than others, students that are good or bad at maths?
Probe: Why do you think those students had more to say?

TRU: Formative assessment

W/C 19. How well do you think you understood the maths in today's lesson?

Prompt: very well, some parts were difficult, hard.

Probe: How did how well you understood the maths make you feel?

20. Do you think your teacher knows if you understood the lesson? How do you know that?
Prompt: he/she checked my work in the lesson, he/she asked me harder questions, he/she asked me to share my work with others.

S/G 21. And when you work in groups/pairs, does the teacher come to your group?
Prompt: why or why not (e.g. the teacher thinks we can do the work without him/her), who does the teacher go to, how the teacher chooses a group he/she is going to work with, do you mind if the teacher doesn't come to you?

W/C 22. What do you think the students in other sets/classes might have been learning in today's maths lesson?
Prompt: same maths, harder/easier maths

APPENDIX K: Teacher post lesson observation interview (mixed attainment)

* Lesson Specific Clarification Questions

1. How does mixed attainment grouping work in your school?

Prompts: which year groups, how many maths classes in each year group

Probe: How long have you been teaching students using mixed attainment grouping practice? (including ITE)

How do you feel about this approach to grouping students?

TRU: Mathematics

2. What were your mathematical goals for this lesson?

Prompts: New learning? Key mathematical ideas? Concept? Procedure? Prior learning? Assessment?

Probe: Were the goals the same for all students in the class?

Prompts: Did all students achieve the mathematical goal for your lesson today (e.g. the focal students in particularly?) If so, how do you know? If not, why do you think that?

What mathematical connections do you think students made to prior learning?

What mathematical goals for future learning, will today's learning lead on to?

TRU: Cognitive Demand

3. How challenging did you imagine the students might find the mathematics content in the lesson?

Prompt: Misconceptions, common errors

Probe: What kinds of mathematical thinking did you want your students to experience in today's lesson?

Prompt: remember facts, vocabulary, use or apply strategies, make connections, develop partial understandings, draw on prior knowledge, interpret/use

Probe: How well did you feel the students managed the level of challenge in the lesson today?

Prompts: Would you say that is the same for all students, with LPA and HPA, e.g. focal students?

How did students manage any challenges they had? (e.g. asking the teacher a question? Asking/questioning another student?)

How do you see your role in managing these challenges?

TRU: **Access**

4. In what ways did you ensure that all students in the class today accessed the mathematics?

Prompt: ways did you ‘differentiate’

To what extent are the low attaining students considered in your lesson?

The lesson required the students to… talk, write, lean in, listen hard, manipulate symbols, make diagrams, interpret text, use manipulatives, connect different ideas, etc

Teacher: used prompting questions, scaffolding questions, giving clues, redirecting student attention to a critical feature?

Tasks: asking open ended/single answer, offering low threshold-high ceiling/narrow focus, providing opportunity for multiple solution strategies/ routine practice of worked example

Probe: Were there any students who participated more or less than others in the lesson? Why do you think that was?

Prompt: What can you tell me about the students who contributed most/less in the lesson today? Specifically, LPA/ HPA students (e.g. our focal students maybe)? Typical patterns of participation?

TRU: Agency, ownership and identity

5. In what ways did you anticipate how different groups of students might contribute during the lesson?

Prompt: Were there parts of your lesson where you had considered how to involve LPA/ HPA?

Probe: Which students contributed to the whole class lesson today?

Prompt: Lower PA/ Higher PA

Probe: Were there **any particular things** that one or more students contributed today that influenced what happed next in the lesson?

Prompt: Were these Lower PA/Higher PA students? How typical are these patterns of interaction for these students?

(Or draw attention to an example observed in the lesson)

TRU: Formative assessment

6. What did you do differently from what you had planned to do today? Why?

Prompt: student responses? Timings? Questions you asked? Student difficulties?

Probe: What did you learn in today’s lesson about your students’ understanding of the mathematics in this lesson?

Prompt: how LPAs different from HPAs

Prompt: What did students say or do that revealed what they understood about the mathematical ideas in today’s lesson?

7. Considering the lessons today, can you imagine in what ways it would be different if you were using setting?

Prompts: in terms of: a) your students' experiences of the mathematics (Specifically for LPA / HPA); b) your teaching practices (e.g. with LPA/ HPA)

8. How many of the Y7 classes do you teach? (out of how many classes?)

Probe: Do you teach other year groups?

Probe: Have you every taught this mathematical content to a different class grouping? How was it the same/ different?

Prompts:

Student experiences: less/ more opportunity … to discuss ideas with other students/ teacher, to work independently/ in a group, to use multiple representations, to solve authentic problems

Teacher practice: different content, seating arrangements, timing, tasks/exercises, resources, representations, cognitive demand, pace

9. Finally, I'd like to ask about your background as a teacher:

- How many academic years have you taught mathematics since you qualified?
- What's your highest mathematics qualification?
- What was your route to achieve QTS?

APPENDIX L: Focus group schedule

Sorting Task 1 Mathematical Content

Sorting statements

Introducing the card sort

- We would like to know more about learning maths in your Y8 lessons and how your Y8 teacher helps you. We have some statements that we'd like you to look at together as a group and discuss where you think you would place the cards under each of the three headings: Most of the time, some of the time, rarely/never
- We'd like you to say out loud to one another your reasons for your choices. If you can't agree, that is fine (in fact it is good). Just try to explain why *you* think the card goes where you would place it and give some examples to back up your suggestion and we'll work out where the card could go.
- Do you have any questions before we begin?

- But before we do the sorting task, we'd like to ask you - what makes maths enjoyable or less enjoyable for you. [Allow students to respond individually].

[*Prompt*: when it's easy/ hard, getting it right, making you think, eventually working out the answer, working with friends]

- OK, we'll take a few cards at a time now and I will ask you a few other questions to help you think about the statements on the cards.

Notes for interviewer:

1. Make it clear for the recording where they are positioned e.g. "So you would put 1.1 with rarely/never?".
2. Describe the theme for the cards as you lay them out. E.g. for 1. Mathematical facts and procedures – "These cards are about learning mathematical facts and procedures".
3. If you sense a strong opinion about the statement, invite the students to suggest whether they would prefer it to be in a different position and to state which, aloud (for the recording).

1. Mathematical Facts and procedures

Lay **1.1** and **1.2** out. Invite a student to read each one aloud. "So where would you put these? Can you discuss with one another?"

1. Learning maths is about remembering facts (e.g. times tables, formulas, definitions) and methods (e.g. long multiplication, finding areas of a 2D shape.)
2. Learning maths is about copying and using the teacher's method from the board.

Probes

i. What facts/methods do you have to remember? How does your teacher help you to remember these facts and methods?

[*prompt:* times tables? formulas? Definitions?]

ii. Are there any facts or methods that are hard for you to remember? Why do you think that might be?

[*prompt*: practice at the beginning of a lesson, short tests, homework]

iii. What happens when you go wrong with a method or can't remember a fact?

Lay out **1.3** and **1.4**. Read the statements to the students. "So where would you put these? Can you discuss with one another again?"

3. Learning maths is about using facts to find other facts or comparing different methods to solve a problem.
4. Learning maths is about finding your own methods to solve problems.

Probes

iv. Can you give an example of how you might use a fact to solve another fact? [e.g. if you know 8 x 2 = 16, what other problems could this fact help you to solve?]. Does your teacher ever ask you think in that way?
v. Does your teacher ever use students' own methods to help you to learn?

2. Mathematical concepts

Lay out **2.1**. Read aloud to the students. "So where would you put this?"

1. Learning maths is about the teacher *telling us about* mathematical ideas e.g. telling us what ratio is.

Probes

vi. Can you give an example of some maths that you feel you understand well. So, what makes you think you understand it well?

[*prompt*: can you do it without thinking? explain why or how it works? It makes sense to you? know when to use it? do it quickly?]

vii. What does your teacher do to help you understand the maths you are working on in your lessons?

[*prompt:* reminds you of what you've learned before, gives examples of when you might use the maths, explains things clearly, prompts you to say what you already know? uses students' mistakes]

3. Mathematical language

Lay out **3.1** and **3.2**. Read aloud to the students. "So where would you put these?"

1. Learning maths is about using correct mathematical words.

2. Learning maths is about explaining your ideas in your own words.

Probes

viii. What mathematical words have you been using recently to help you learn maths?
ix. Can you give some examples of when you've been asked to explain your ideas to the class?

[prompt: What happens after you've explained your ideas?]

x. Do you think there are some students who get chosen more often than others to explain their ideas? Why do you think that is?

Lay out **4.1** and **4.2**. (Mathematical representations). Read aloud to the students. "So where would you put these?"

4. Mathematical representation

1. Learning maths is about reading and writing equations and using mathematical symbols.
2. Learning maths is about representing our ideas in different ways. E.g. drawing a bar model, using cubes.

Probes

xi. Can you give some examples of how you use different representations to help you learn the maths in your lesson?

[*prompt*: drawing something? Moving something around? Using technology – e.g., phone, ipad, computer app]

xii. Does your teacher help to link the maths you are learning to real life? Can you give some examples?

5. Mathematics as a discipline

Lay out **6.1** and **6.2**. Read aloud to the students. "So where would you put these?"

1. Maths is about getting the right answer.
2. Mathematics is about looking for and using patterns and rules.

Probes

xiii. Do you ever get asked to explain why an answer is correct? Can you give an example?
xiv. What do you think it means to be good at maths?

[*prompt*: you? think of someone in your class]

Sorting Task 2 Cognitive demand

Sorting statements

Preparing for the card sort

- This time we would like to know more about your experiences of struggle in your Y8 maths lessons. Like before, we have some statements that we'd like you to look at together.

- But before we do the sorting task we'd like to ask you what you understand by the term "struggle". [Allow students to respond].

[*Prompt*: being stuck, making you think hard, feeling challenged, you are learning/ not learning]

Probes

I. Do you think it is a good or a bad thing to struggle in your maths lessons? Why do you think that? [*prompt*: been given right/wrong level of work]

II. What kinds of things do you struggle with in maths?

[*prompt*: remembering things? Making sense of a problem? Understanding the teacher/ the students? Keeping up with others?]

III. Which students struggle in your maths class?

[*prompt*: students who are better at maths/ not as good at maths] How do you know when they are struggling?

Beginning the sorting task

- Here are some cards about struggle, we'd like you to discuss where you think you would place the cards under each of these headings.
- We'll take a few cards at a time and sometimes I will ask you a few other questions to help you think about the statements on the card.

Notes for interviewer:

1. Make it clear for the recording where they are positioned e.g. "So you would put 1.1 with rarely/never?".
2. Describe the theme for the cards as you lay them out. E.g. for 1. Mathematical facts and procedures – "These cards are about learning mathematical facts and procedures".
3. If you sense a strong opinion about the statement, invite the students to suggest whether they would prefer it to be in a different position and to state which aloud (for the recording).

Most of the time, some of the time, rarely/never

6. Teachers managing struggle

Lay **6.1** and **6.2** out. Read each one aloud. "So where would you put these? Can you discuss with one another?"

1.	When we struggle our teacher asks us to explain what we are thinking.
2.	The teacher shares things we are struggling with, with the whole class.

Probes

I. Do you think your teacher wants all students to struggle in your maths lessons? Why?

[*prompt*: no, a little bit, all students need to]

II. How does your teacher respond when someone struggles in your lesson?

[*prompt*; the teacher … reassures/ encourages you, gives you a hint, repeats instructions, uses your idea with the class, asks other students to help you, gives different work]

III. How does your teacher know when students are struggling or when they find the work easy?

7. Students managing struggle

Lay **7.1** and **7.2** out. Read each one aloud. “So where would you put these? Can you discuss with one another?”

1.	When we are struggling, we put up our hand for our teacher to come to us.
2.	We share our answers even if we aren’t sure we are correct.

Probes

IV. Can you give me an example of when you have been stuck on something in maths. How did you try and get ‘unstuck’? Was that alone or were you working with others?

[*prompt*: hands up? Anything before hands up? Look back at book, check board, discuss with others around me; use resources/ draw something?]

V. Does explaining your ideas or listening to others’ ideas help you to makes sense of the maths in your class? Why/how?

APPENDIX M: Head of Mathematics interview

*

* Notes from HOM survey/ Observations from lesson – specific queries

1. I'd first like to ask you some background information

   - How many academic years have you been a HOM?

   - In this school?

   - Others?

   - Did they also use similar grouping practices?

   - How many academic years have you been a teacher of mathematics?

   - Do you/ have you had any other roles in the school?

   - What is your highest mathematics qualification? (*GCSE, A level, mathematical degree?*)

   - What was your route to achieving QTS? (*PGCE, GPT, TF other?*)

   - Which year groups are you teaching maths this year? (*sets? Nurture group?*)

About the type of grouping practice

2. Tell me about how you group students in Y7 and Y8 for mathematics in your school.

   Prompts: How many sets/ classes? Nurture groups? Who decides who goes into which set/class?

   Probe: Are there any other provisions for students in Y7/8 for different groups of students (e.g. Lower PA / higher PA?)

   Prompts: homework club, interventions, enrichment activities?

3. What are the reasons for using this approach to grouping your students in Y7 and 8?

   Prompt: Similar grouping practices in older year? Why? Why not? Certain student characteristics of one/all cohorts? Who decided to group students this way?

   Probe: How well does setting /mixed attainment teaching work?

   Prompt: What are the benefits? Challenges?

4. How do you feel about leading a department that uses this approach to grouping students?

   Probe: Do you think this is the best way to group students? Why?

   Prompt: Have you ever considered grouping students in other ways? Have you worked in schools that have grouped students in a different way?

   Reflecting on the values and beliefs underpinning the grouping approach

   TRU Equitable access

5. How do you support *__all learners__* in Y7, so that they have an opportunity to make progress in mathematics in your school?

Prompt: what about for those students with lower prior attainment? Higher prior attainment? Teacher allocation, student-to-class allocation?

Probe: How do you know whether your (LPA/ HPA) students' progress benefit from your grouping approach?

Prompt: Have there been any particular challenges to creating equitable access to mathematical learning? For LPAs or HPAs?

Probe: (Setting only) What would your advice be about doing setting as equitably as possible?

TRU: Agency, ownership and identity

6. From your perspective as a HOM overseeing the department, what do teachers do to enable all students to *participate* in learning mathematics in Y7/8?

Prompt: work in groups, discuss, explain, listen to teacher, work in silence, volunteer answers

Probe: (Setting) How would student agency compare in different sets?

Prompt: lower/higher sets; agency is where students have a voice in their classroom

Probe: So, how do you think your grouping approach promotes student agency in their mathematics learning?

7. To what extent would Y7 students in your school see themselves as maths people?

Prompt: (MA) Would Lower/higher attainers see themselves in the same way or (setting) in higher or lower sets

Probe: What do think for? your students might mean to be good at maths?

TRU: Mathematical content

8. Tell us about the origins of your scheme of work?

Prompt: SoW written by…school? MAT? NCETM? White Rose? Purchased from… so would you say you use a "mastery" approach?

Probe: (MA) So how do teachers use this SoW to set learning objectives that account for different prior attainment in the class?

Prompt: Lower prior attainers / higher prior attainers?

Probe: (Setting) So how do teachers use this SOW to set learning objectives for different sets?

Prompt: experience the same/different mathematical content? If no, how do you decide what content is included/omitted?

TRU: Cognitive Demand

9. How is challenge included in your scheme of work?

Prompt: different types of tasks – open ended, reasoning, different contexts

Probe: What would you expect *challenge* to look like in mathematics lessons in Y7/8?

Probe: Would challenge look different in different sets? / for students with different prior attainment?

TRU: Formative assessment

10. How do you expect teachers in Y7/8 to use formative assessment in the classroom?

Prompt: Clear learning objectives for lessons? Indicators of success/ difficulty? Assessment policy? Marking policy? Feeding back to students?

Probe: How might teachers in Y7/8 in your school monitor the learning of students with differing prior attainment in their lessons?

Probe: How well do you feel teachers' use of formative assessment impacts on LPA students and HPA students in Y7/8? / students in different sets?

Reflecting on the quality of teaching for different groups of students

11. What do you think are important qualities for teaching mathematics?

Prompting: teachers' subject knowledge, pedagogical knowledge (how to teach), pedagogical content knowledge (content knowledge for teaching); knowledge about learners; school T&L policy?

Probe: Are there different qualities needed for teaching different students in different grouping arrangements? mixed attainment classes/top set/bottom set?

Prompt: Teacher allocation – revisit if relevant – allocating teachers with different skills/qualities to particular class/sets

Probe: (MA only) I guess you'll have a range of teachers with different expertise in Y7 and Y8. How are they supported to teach mixed attainment mathematics in your school?

12. (MA) What advice would you give to a HOM considering adopting mixed attainment grouping in a school that currently sets?