PROJECT TITLE	The Student Grouping Study: investigating the effects of setting and mixed attainment grouping
DEVELOPER (INSTITUTION)	n/a
EVALUATOR (INSTITUTION)	UCL Institute of Education
PRINCIPAL INVESTIGATOR(S)	Professor Jeremy Hodgen Dr Becky Taylor
TRIAL (CHIEF) STATISTICIAN	Professor Jeremy Hodgen
STUDY PLAN AUTHOR(S)	Professor Jeremy Hodgen Dr Becky Taylor Dr Jake Anders Dr Antonina Tereshchenko Professor Becky Francis

Study Plan version history

VERSION	DATE	REASON FOR REVISION
1.2 [*latest*]		
1.1		
1.0 [*original*]	1/11/2019	*[leave blank for the original version]*

Table of contents

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 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 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).

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

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 C.1).

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 all 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 as pedagogical content knowledge (PCK) 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. 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.

Our previous study, Best Practice on Grouping Students, 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 the prior study has 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. 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.

Provide detailed evidence of how grouping practices are implemented. The Student Grouping Study will pay close attention to how grouping practices are implemented in different schools. As a result, the study will characterise effective practices so that schools can understand how to implement setting and mixed attainment in order to make the greatest impact. This will include developing measures of 'opportunity to learn' mathematics (OTL) and teacher quality (pedagogical content knowledge, PCK). 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.

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 in Years 7-8 attending schools that use mixed attainment grouping for mathematics, and low-attaining students in Years 7-8 attending a similar group of schools that use setting?

2. What difference is there, if any, in the attainment of *all* students in Years 7-8 attending schools that use mixed attainment grouping for mathematics, and *all* students in Years 7-8 attending a similar group of schools that use setting?

3. What difference is there, if any, in the mathematics self-confidence of low-attaining students in Years 7-8 attending schools that use mixed attainment grouping for mathematics, and low-attaining students in Years 7-8 attending a similar group of schools that use setting?

4. What difference is there, if any, in the mathematics self-confidence of *all* students in Years 7-8 attending schools that use mixed attainment grouping for mathematics, and *all* students in Years 7-8 attending a similar group of schools that use setting?

5. To what extent do opportunity to learn and teacher quality (as assessed by pedagogical content knowledge) 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. To what extent do schools' specific current grouping practices vary within the groups using setting and mixed attainment? What are the reasons for these practices?

7. How do students with low prior attainment experience different grouping practices? What are the beneficial and detrimental effects of these experiences associated with setting and mixed attainment?

8. What pedagogic practices do teachers use in Year 7 and Year 8 mathematics lessons, for different grouping types? 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 and schools' decisions regarding grouping?

Note: For the purposes of this study, low attaining students will be defined as those students who do not achieve at the expected level at KS2 (as per the DfE NPD).

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, low-prior attainers and high-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

Design type		Matched design study in a natural context
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

Participants will be all students starting Year 7 in September 2019 in recruited schools. They will participate in the study until they finish Year 8, in July 2021. 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 will be 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 will be eligible to participate in the mixed attainment group if they meet the following criteria:

- State-funded secondary or middle 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.[3]

Schools will be eligible to participate in the 'Setting' comparison group if they meet the following criteria:

[3] 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).

- State-funded secondary or middle 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 will not be 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.

We intend to recruit matched schools, with a ratio of 1:2 Mixed Attainment to Setting schools (see sample size calculations below). The 'quasi-randomisation date' will be the date at which recruitment to the mixed attainment group is complete and the matching procedure to identify matched setting schools is carried out. The exact date will depend on the progress of recruitment to the mixed attainment group but is anticipated to be around the middle of May 2019.

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[4] 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.

[4] See Francis et al. (2017) for further information about the self-confidence survey items.

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, we intend to develop and validate an instrument to measure teachers' mathematical knowledge. We will draw on the measures 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). This measure consisted of three sub-scales: knowledge of different approaches to solving tasks, knowledge of students' misconceptions, difficulties and learning, and knowledge of representations and explanation used in instruction. A key challenge will be 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. We anticipate validating the measure through at least two pilot administrations, each with a sufficient sample to conduct factor analysis and Item Response Theory (IRT) analysis. We note that we considered other approaches to measuring teacher quality, specifically using historical average student gains associated with the teacher (although this was judged to be too costly and burdensome to schools).

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.[5]

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

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

As noted above, we conjecture that moderating factors are likely to include those relating to the student (SES, 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).

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
- 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 and based on a 2-level model with students clustered within schools. We show the MDES calculation assuming a pre-post test correlation of 0.75 (at the student-level). We 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.

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).

		Study Plan	
		OVERALL	**FSM**
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	100
Number of schools	Intervention	40	40
	comparison	80	40
	Total	120	120
Number of pupils	Intervention	4000	1000
	comparison	8000	2000
	total	12000	3000

Selection of the comparison group and identification assumptions

We will recruit schools to the mixed attainment group first. We estimate 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 do not know the identity of all these schools in advance, recruitment will take 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 anticipate that recruitment may be difficult. After mixed attainment schools have been recruited, we will undertake to identify a group of setted schools matched to each mixed attainment school.

Matching will be conducted at the school-level. There are strong theoretical reasons for matching at school-level since the decision on grouping students is likely to be have taken at a school-level (Hallam & Ireson, 2003; Cook, Shadish, & Wong, 2008). Moreover, given the timescale for recruitment, it will not be practically possible to gain access to the NPD prior to recruitment and thus collect student-level data in time for matching.

We initially considered the advantages and disadvantages of two broad approaches to matching: propensity score matching (PSM) and coarsened exact matching (CEM). Both approaches involve the loss of recruited schools through the matching process, either through the imposition of common support in PSM or the failure to identify matched comparison schools within the tolerances chosen for each variable in CEM. Our initial preference was for PSM and we investigated the use of nearest neighbour and optimal matching strategies (Anders et al., 2017). However, it was found that CEM would produce better balance on the key attainment variables. We will consider imposing common support to ensure sufficient balance across the groups. The recruited sample of mixed attainment schools included some middle schools, which were excluded from the study because there are very few middle schools nationally.

Matching will be on variables that simultaneously influence the outcomes and the decision to adopt the treatment, but should not be themselves be affected by the intervention (Caliendo & Kopeinig, 2008). Hence, we propose to use a range of school-level variables focusing on the intake and the school itself, including historical variables for attainment. Specifically, the matching variables we will include the following (the detailed rationale for these choices can be found in Appendix C).

- Mean prior attainment (KS2) (current and historical)
- Spread of prior attainment or % high and low prior attainers
- % free school meals (FSM6 on entry to Year 7)
- % EAL
- % in low and high IDACI neighbourhoods
- Size of school
- School region (urban/rural)

Having identified this pool of matched comparators, we initially aimed to recruit 2 of these to the study, prioritising with the closest matches. Our previous study indicates that recruiting schools to the mixed-attainment group will be considerably more difficult, because these schools are relatively rare (Taylor et al., Submitted). In order to increase power conditional on that constraint, we aimed to recruit a greater number of schools using setting practices. Our estimate that, on average, we would need to approach 10 “matched” schools in order to actually recruit two suitable matches proved to be realistic.[6] However, because the recruitment of eligible mixed attainment schools proved more difficult than anticipated and a smaller sample of these was recruited, a larger sample of setted schools was recruited.

In order to achieve this, it will be necessary to identify a greater number of initial matched setted schools for each mixed attainment school. We have simulated the response for PSM matched involving combinations of matching variables above, different numbers of initial matched setted schools and an estimated probability of responding positively of 0.2, which is slightly more conservative than our estimate above. (See Appendix D for a discussion of this process.) The simulation exercise is based on our initial set of 43 schools know to be using mixed attainment grouping in mathematics and is based on a PSM based on the following matching variables: average KS2 attainment at 2018, 2017 and 2016, % FSM, % high attainers, % low attainers, school size, IDACI, OFSTED rating, academy status, region and urban/rural.

Based on this simulation exercise, 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 using a robust matching model as part of this project, given the small number of mixed attainment schools available. Our aim

[6] On the basis of a 1:4 contact-to-recruitment ratio and that some schools will not meet our “setting” eligibility criteria for the comparison group.

is to achieve a good balance on prior attainment, proportion of students eligible for FSM, proportion of high and low attainers and school size in order to justify the conditional independence assumption. We will review this in the light of the actual recruited sample of mixed attainment schools.

Recruitment strategy

In order to speed up the recruitment process, we will target schools who we know to be teaching mathematics to mixed attainment groups since prior research shows that mixed attainment grouping is much less prevalent in mathematics than in English (Taylor et al., 2017). Nevertheless we believe that a sample will be attainable by using networks such as the Mixed Attainment Maths conference group. We will attempt to over-recruit to the treatment group (and, hence, to the matched comparison group as a logical corollary of this) to allow for attrition through school dropout and the potential exclusion of schools falling outside the range of common support.

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.

We will conduct and present a graphical exploration of the density of the area of common support and also report the proportion of the sample discarded due to a) caliper width and b) imposition of common support.

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

Given the exploratory nature of this naturalistic study we do not consider it appropriate to specify the regression model in advance. Indeed, the EEF's "standard" approach, a 2-level MLM, relies on strong assumptions that may not hold within the PSM framework. Rather, we propose to follow the approach proposed by Ho, Imai, King, and Stuart (2011) and use either the MatchIt or optmatch packages in R to build a more robust inferential regression model using the matching as a pre-processing stage. We intend to include covariates in the model as this is likely to further reduce bias in the estimates of impact in addition to the precision. These covariates will include KS2 attainment and several other variables, which be specified in detail in the SAP after the two groups are recruited. Precedence will be given to those variables where imbalance is found. We intend to report both Average Treatment Effects on the Treated and Average Treatment Effects. We will explore how to enable comparability of the ES estimates with those from other trials.

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.

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.

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

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.

Sub-group 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. If the coefficient on the interaction term is found to be statistically significant at the 5% level then we will 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.

Mediation and moderation analyses

We intend to conduct mediation and moderation analysis using mediators and moderators identified in the logic model:

Mediators: Opportunity to learn (OTL), teacher quality (measured as teacher pedagogical knowledge), class size, and (if possible) students' attitudes (such as liking for school or engagement).

Moderators: School factors and student and teacher demographic characteristics

This will follow the principles of mediation and moderation 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). Most of these mediation and moderation analyses will be treated as exploratory but will aim to provide indicative evidence of the potential factors that drive differences in our outcomes. However, 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). We will review this as part of the statistical analysis plan (SAP) where we will specify the models in detail.

Missing data

We will adopt an approach that we have used previously (see Anders and Shure, 2018), which will be fully specified in the SAP. 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 5% of data in the model is missing or if more than 10% of data for a single school is missing.

Effect size calculation

Effect sizes will be calculated using the Cohen's *d* ES for cluster randomised trials as per the current EEF (2018) 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, will be fully specified in the SAP.

Analysis of the implementation of setting and mixed attainment

We regard the implementation analysis for this study as particularly important, given the likely diversity of practice within the Mixed Attainment and Setting groups. We will therefore carry out a multi-phase mixed methods implementation analysis (Anders et al., 2017; Humphrey et al., 2016) and aim to collect a range of data throughout the duration of the study including:

- Collection and analysis of school policies relevant to grouping practices.
- Survey of Heads of Mathematics to identify schools for qualitative data collection.
- Survey of mathematics teachers in setting and mixed attainment schools.
- Interviews with a sub-sample of teachers and senior leaders.

- Surveys of the whole cohort of students, and focus groups with a sub-sample of students. Separate focus groups in each school will include students with high, middle and low prior attainment.
- Observations of mathematics lessons with a focus on (a) pedagogic practices and (b) the student experience.

Anders et al. (2017) note the importance of an iterative approach to evaluation in the case of complex whole school interventions. The implementation analysis will consist of both quantitative and qualitative elements. We intend to collect some basic information about grouping practices early on, in order to inform and target later qualitative data collection activities. We envisage a total of six intensive case studies in each group in order to capture the range of practice.

Case study schools will be carefully selected on the basis of current grouping practices and school characteristics, with the intention of representing a variety of schools. School practices will be established through the Head of Mathematics survey in September/October 2019. We will spend up to a total of five days in each case study school at two time points. At the first time point, we will collect relevant school policies, interview at least one member of the senior leadership team, the Head of Mathematics and a minimum of three mathematics teachers. We will also carry out three focus groups with students and observe a number of mathematics lessons. Focus groups will consist of four students, with one group of each of high, middle and low prior attainment in each school. At the second time point, we will conduct follow-up interviews informed by the analysis of the first round of data collection.

In order to understand how and why grouping practices impact on attainment and self-confidence outcomes, we need to understand how and why schools are using setting and mixed attainment grouping practices (partly in order to justify the assumptions underlying the PSM model), the variation in these practices and how they influence classroom pedagogy and so the learner experience.

With attention to our Logic Model (Appendix A, Table A.1), our priority areas will be to monitor practices in the two groups and examine closely the different ways that grouping practices are enacted in schools.

Decision factors: We have proposed that a range of factors will have influenced schools' decision to use setting or mixed attainment grouping. Certain factors have been carefully selected for use in the matching process (see Appendix C and also the decision logic model, Appendix A, Table A.2). In the qualitative element of the implementation analysis, we will interview (and survey) senior and middle leaders to identify the factors that have actually influenced their decisions. We will additionally ask about the stability of practices: how long they have been in use (including variations in practice) and whether there are any plans to change.

We anticipate that a small number of schools may decide to change their grouping practices during the course of the study. Where possible, we will interview the key decision makers in these schools.

Grouping strategy: Detailed understanding of grouping strategies will be established through interviews with senior leaders and the Head of Mathematics. To what extent do differences in practice *between* ‘mixed attainment’ and ‘setting’ schools mean that they constitute two meaningfully distinct groups? To what extent are the practices *within* each group (mixed attainment or setting) similar enough to be meaningfully treated as a group? This analysis will draw on the analysis of the other implementation dimensions below and will be used to inform the analysis of compliance. To what extent do the schools’ actual grouping practices meet the study definition of setting or mixed attainment grouping? Are students taught in sets or mixed attainment groups for one year (Year 7 only) or two years (Year 7 and Year 8)? Is there any within-class attainment grouping, either explicit (e.g. ‘ability tables’) or implicit (e.g. differentiation by task with fixed groups)? For schools operating mixed attainment grouping, do all classes reflect the full range of prior attainment in the year group? Are there any attainment groups alongside mixed attainment groups, e.g. nurture groups for very low attaining and SEND students, groups for ‘gifted and talented’ students or ‘catch up’ classes for lower attainers? In what ways are nurture groups for students with very low prior attainment organised? What information is used to allocate students to attainment groups? How frequently does movement between groups take place and on what basis? How many set ‘levels’ are in operation and how many classes are there at each level? What size are the classes and does class size vary with set ‘level’? Are there any elements of streaming introduced, for example through mathematics sharing grouping decisions with another subject?

Teacher/pedagogic/school mediators: These will be investigated through surveys of teachers and explored in greater detail with interviews in the case study schools. To what extent do teachers adhere to setting or mixed attainment grouping practices as defined by their school? Do teachers have personal preferences for setting or mixed attainment teaching? What are teachers’ expectations of students? Do these vary with student prior attainment/set level taught? What pedagogies are used with sets and with mixed attainment groups? Does pedagogy differ between set ‘levels’? To what extent does the curriculum vary between and within classes, and to what extent does opportunity to learn capture this variation?

We will also investigate the influence of school factors such as school leadership ethos (which will be measured through a short survey to be completed by a member of the SLT. We intend to pilot and validate this survey.

Student factors: We anticipate that student attitudes to setting and mixed attainment grouping may also influence their engagement and measures of these will be included in the student survey. Student focus groups in case study schools will explore in greater depth how these factors are related to specific school grouping practices.

We will also examine relevant implementation factors:

Preplanning and foundations: Why do schools use particular grouping practices for particular year groups and subjects? What beliefs do teachers hold about the stability and suitability of these practices?

Implementation support system: What training, support and resources are available to teachers to ensure high quality grouping practices? What training, support and resources are available to teachers to ensure high quality pedagogy for mixed attainment or set groups. To what extent is there evidence to support the perception amongst teachers that mixed attainment grouping takes more time than setting?

Implementation environment: How are the school context and the school characteristics related to choice of grouping practices?

Implementer factors: How are teacher characteristics, attitudes and beliefs related to the grouping practices in their school?

Implementation data collection and analysis

Questionnaires and Surveys

- Questionnaire to be completed by the Head of Mathematics in each study school, administered online to gather data on schools' specific current grouping practices in Year 7 and then in Year 8, as well as the reasons for these practices in each Year.
- Survey of Senior Leaders from each study school to gather data on beliefs around attainment grouping, equity, school leadership ethos and organisational factors/constraints.
- Survey to be completed by all mathematics teachers in the study, administered online to gather data on: their perceptions of reasons for grouping practices; personal beliefs about attainment grouping and mixed attainment, as well as effects on students; pedagogic practices; opportunity to learn; and mathematical pedagogical content knowledge. The full survey will be piloted and validated on a sample of teachers before use. The validation of the measure of mathematical pedagogical content knowledge is described above.
- Survey of all students at the end of the study adapted from the survey used in the BPGS study to collect data on attitudes to, and experiences of, respective grouping practices and mathematics teaching; opportunity to learn; and aspirations / expectations. We will validate this survey using statistical techniques (factor analysis and item response theory modelling) and further piloting in order to create attitude and opportunity to learn scales for use in modelling. This will be further specified in the SAP.

Case Studies

We will conduct case studies of 12 schools, including 6 schools with setting and 6 with mixed attainment grouping. Schools will be selected randomly. Data collected will include:

- Interviews with Head of Mathematics in each school (n=12). We will conduct them in Y7.
- Interviews with two Year 7 mathematics teachers in each school (n=24).
- Interviews with two Year 8 mathematics teachers in each school (n=24).
- Schemes of work and curriculum plans for mathematics lessons to use as prompts for interviews and to examine differences in the curriculum offered, and opportunity to learn afforded, to different students.
- Systematic observation of 4 full mathematics lessons for relevant teachers and students (n=48), including two lessons in Year 7 and two in Year 8. We will observe 4 lessons in each of the mixed attainment schools. We will observe 4 low set lessons in

each of the setting schools. Additionally, to establish a contrast with other sets, we will aim to observe parts of lessons in top and middle sets in each of the 6 case study schools with setting. At this stage, we anticipate that we will focus on several lessons with the same class by the same teacher using a structured observation approach. There are at least 11 structured observation protocols that have been validated and used for the systematic observation of mathematics lessons and these have been reviewed and compared in a recent study (Praetorius & Charalambous, 2018). We will review and evaluate these in order to decide on the optimal approach for the current study. We anticipate that our focus will be on the teacher, the tasks / mathematics and the pedagogy, although we will collect some classroom vignettes as prompts for student interviews. We will also collect student work.

- Brief interviews (n=8) with 4 focal students with low prior attainment after observed lessons in Year 7.
- Brief interviews (n=8) with 4 focal students with low prior attainment after observed lessons in Year 8.
- Focus groups (n=12) with students in low set and students with low prior attainment in mixed attainment class. A total of 48 students with low prior attainment (4 per school balanced by gender) will be involved. We will conduct these in either at the end of Year 7 or during Year 8.
- Focus groups (n=12) with a total of 24 students in top sets and with 24 students at high prior attainment. Students will be balanced by gender as above.
- Focus groups (n=12) with 24 students in middle sets and with 24 middle prior attainment. Students will be balanced by gender as above.

Observations and interviews will involve structured and semi-structured elements. All methods will be piloted in one school prior to implementation analysis.

Table 3: Overview of how data addresses implementation research questions

Implementation Analysis RQs	**Data**	**Analysis**
RQ6.	Survey to SLs. Questionnaire to Heads of Mathematics in Y7. Interviews with Heads of Mathematics.	Descriptive statistics. Qualitative thematic analysis.
RQ7 (and also further analysis of RQ5).	Systematic observations of mathematics lessons with a focus on pre-selected students with low prior attainment. Individual brief interviews with each focal student in Year 7 and Year 8 after observed lessons, with lesson observations used as individual interview prompts. Focus group with focal students with low prior attainment. Focus groups with students with high and middle prior attainment. Teacher interviews. Teacher surveys. Student survey. Schemes of work.	Statistical techniques. Qualitative thematic analysis Descriptive statistics. Documentary analysis of schemes of work.
RQ8.	Surveys with teachers. Systematic observations of mathematics lessons. Interviews with teachers.	Descriptive statistics. Qualitative thematic analysis.
RQ9.	Survey to SLs. Questionnaire to Heads of Mathematics in Y8. Interviews with Heads of Mathematics.	Descriptive statistics. Qualitative thematic analysis.

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 themeselves 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 and QUB will sign a data sharing agreement outlining data security and protection issues.

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) to 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 proposed team would be based at UCL Institute of Education 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.

Professor Becky Francis: Project leadership and strategic management, project advocacy, contribution to all aspects of study.

Dr Antonina Tereshchenko: implementation analysis and case studies, recruitment and retention of schools.

Professor Louise Archer Additional design and methodological expertise & advice: qualitative analysis, and instrument design.

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.

Professor Martin Mills: Additional design and methodological expertise and advice: qualitative analysis, classroom processes & instrument design.

A Research Associate (RA1) will develop and validate the teacher and student surveys and provide assistance with administering the teacher survey in Year 7.

A Research Associate (RA2) will provide 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.

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

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

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	SAP 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 to identify schools for IA Baseline student self-confidence survey	Admin Admin BT/AT BT/Admin
Nov 2018	UPNs submitted for NPD data matching	AT/Admin
Sep 2019 – Aug 2020	Survey instrument design and validation Qualitative instrument design & piloting (Sep – Feb) Qualitative data collection (Mar – Jul) Teacher Year 7 survey	RA1 AT AT/RA2 BT/RA1
Oct 2020 – Mar 2021	Qualitative data collection Preparation for outcome testing administration	AT/RA2 Admin/AT
Apr – Jul 2020	Outcome measure data collection: - Progress Test in Mathematics - Student surveys Teacher survey (year 8)	TA/Admin/BT
Jul 2020 – Jan 2021	Report writing	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.

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.

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.

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.

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*.

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

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., Gutierrez, G., & Tereshchenko, A. (Submitted). Attainment grouping in English secondary schools: A national survey of current practices.

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

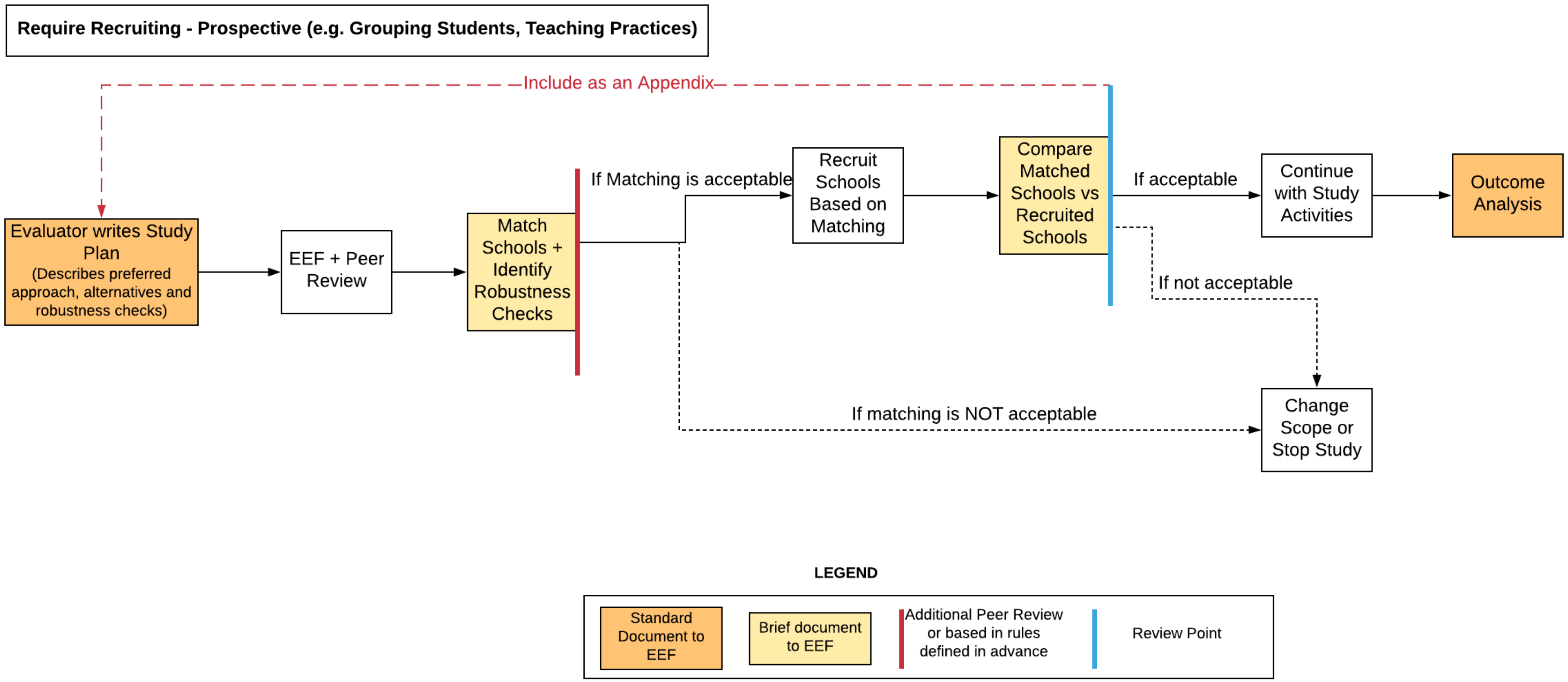
Taylor, B., Hodgen, J., Francis, B., Archer, L., Mazenod, A., Tereshchenko, A., & Travers, M.-C. (forthcoming). Can a mixed attainment grouping intervention increase equity in English secondary schools?

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, Heather Rolfe and Stefan Speckesser .

APPENDIX A: Review process

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


Require Recruiting - Prospective (e.g. Grouping Students, Teaching Practices)
Include as an Appendix
Evaluator writes Study Plan (Describes preferred approach, alternatives and robustness checks)
EEF + Peer Review
Match Schools + Identify Robustness Checks
If Matching is acceptable
Recruit Schools Based on Matching
Compare Matched Schools vs Recruited Schools
If acceptable
Continue with Study Activities
Outcome Analysis
If not acceptable
If matching is NOT acceptable
Change Scope or Stop Study
LEGEND
Standard Document to EEF
Brief document to EEF
Additional Peer Review or based in rules defined in advance
Review Point

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)?

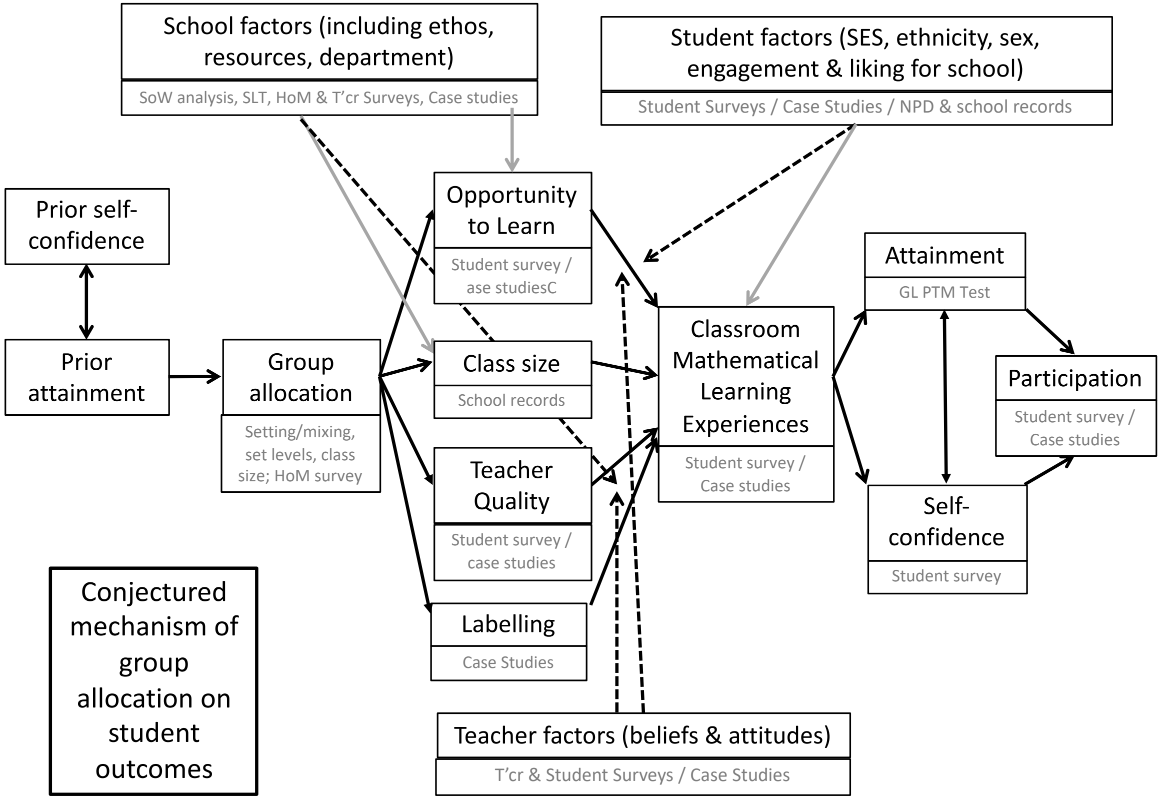
The options include a standard EEF RCT [option A]. Option C described the preferred design..

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

Option	Issues	Pros	Cons
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
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



School factors (including ethos, resources, department)
SoW analysis, SLT, HoM & T'cr Surveys, Case studies
Student factors (SES, ethnicity, sex, engagement & liking for school)
Student Surveys / Case Studies / NPD & school records
Prior self-confidence
Prior attainment
Group allocation
Setting/mixing, set levels, class size; HoM survey
Opportunity to Learn
Student survey / ase studiesC
Class size
School records
Teacher Quality
Student survey / case studies
Labelling
Case Studies
Classroom Mathematical Learning Experiences
Student survey / Case studies
Attainment
GL PTM Test
Participation
Student survey / Case studies
Self-confidence
Student survey
Teacher factors (beliefs & attitudes)
T'cr & Student Surveys / Case Studies
Conjectured mechanism of group allocation on student outcomes

Figure C.1: Logic model / theory of change 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)

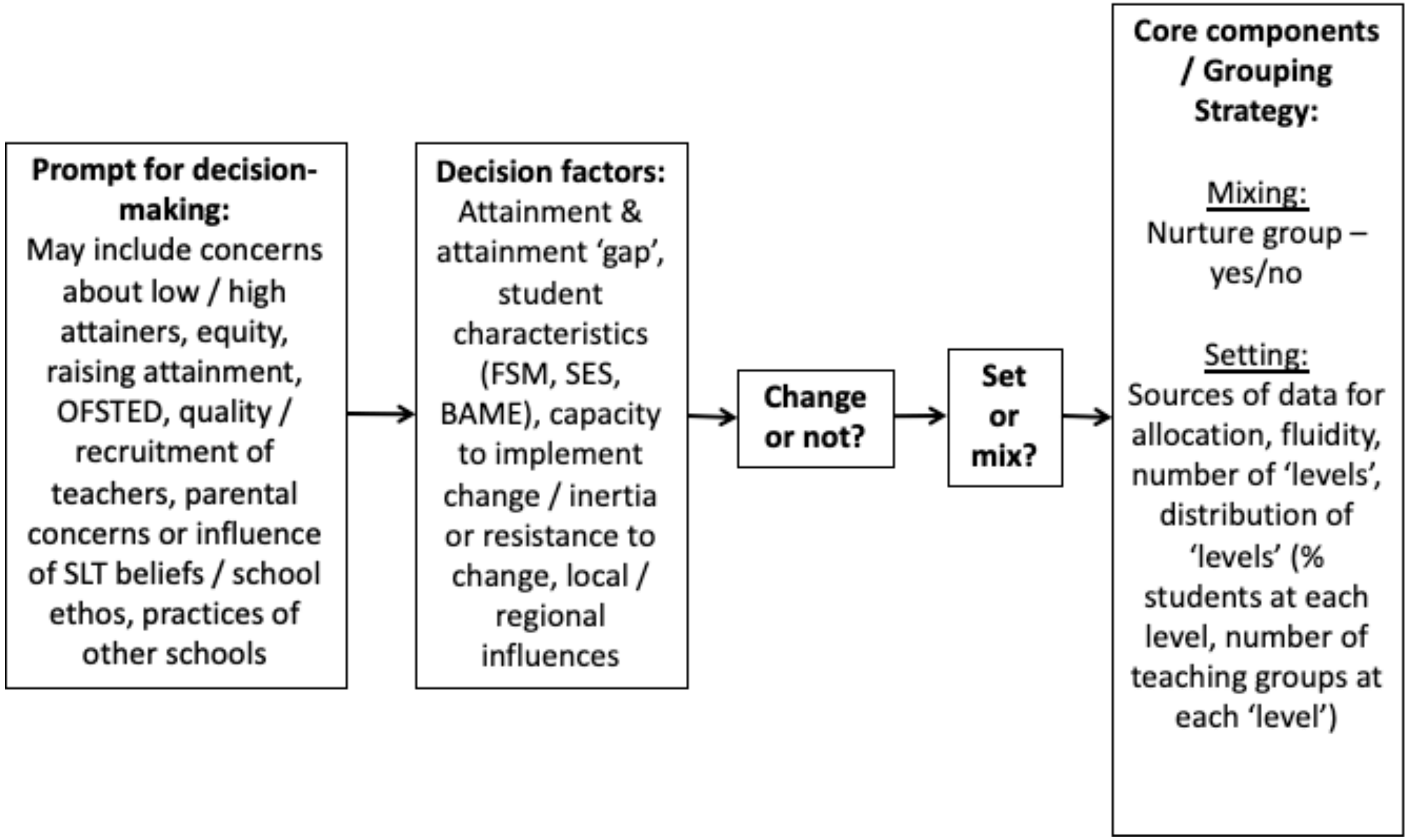

Prompt for decision-making:
May include concerns about low / high attainers, equity, raising attainment, OFSTED, quality / recruitment of teachers, parental concerns or influence of SLT beliefs / school ethos, practices of other schools
Decision factors:
Attainment & attainment 'gap', student characteristics (FSM, SES, BAME), capacity to implement change / inertia or resistance to change, local / regional influences
Change or not?
Set or mix?
Core components / Grouping Strategy:
Mixing:
Nurture group – yes/no
Setting:
Sources of data for allocation, fluidity, number of 'levels', distribution of 'levels' (% students at each level, number of teaching groups at each 'level')

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

Factor	Justification	Indicator(s)	Issues	Influence decision?	Impact on outcomes?	Influenced by outcomes?	Use for matching?
	Aim to close attainment gap between FSM and non-FSM at entry	FSM gap at KS2		Yes	Yes	No	Yes
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

Factor	Justification	Indicator(s)	Issues	Influence decision?	Impact on outcomes?	Influenced by outcomes?	Use for matching?
	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~~

Factor	Justification	Indicator(s)	Issues	Influence decision?	Impact on outcomes?	Influenced by outcomes?	Use for matching?
		% EAL	As for FSM. Schools may choose to group learners by different stages of learning English.	Yes	Yes	No	Yes
		% 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

Factor	Justification	Indicator(s)	Issues	Influence decision?	Impact on outcomes?	Influenced by outcomes?	Use for matching?
		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 factors are potential proxies	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
		OFSTED grade	May be an indicator of capacity to change, but OFSTED grade may be influenced by outcomes / value added	Yes	Yes	??	~~Possibly?~~ 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

Factor	Justification	Indicator(s)	Issues	Influence decision?	Impact on outcomes?	Influenced by outcomes?	Use for matching?
		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

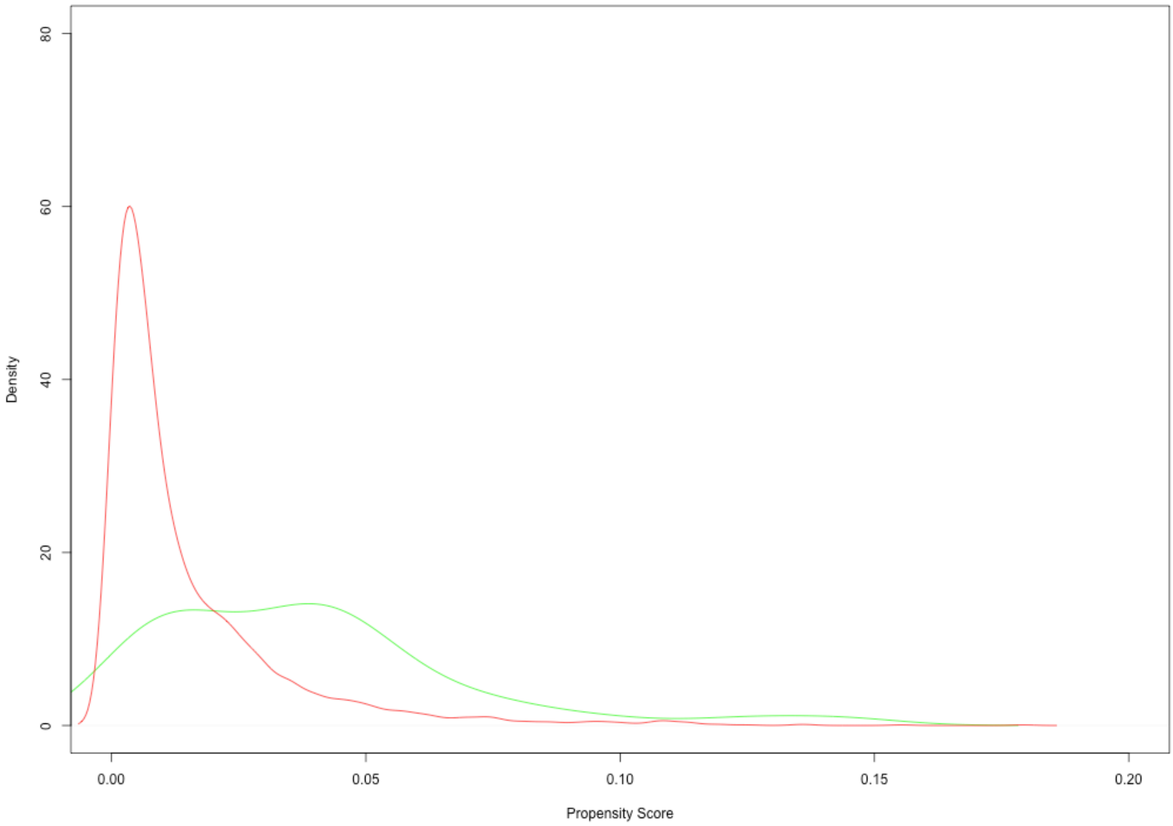
Matches are found based on a treatment propensity score estimated from the following generalised linear model with a probit link function:

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

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

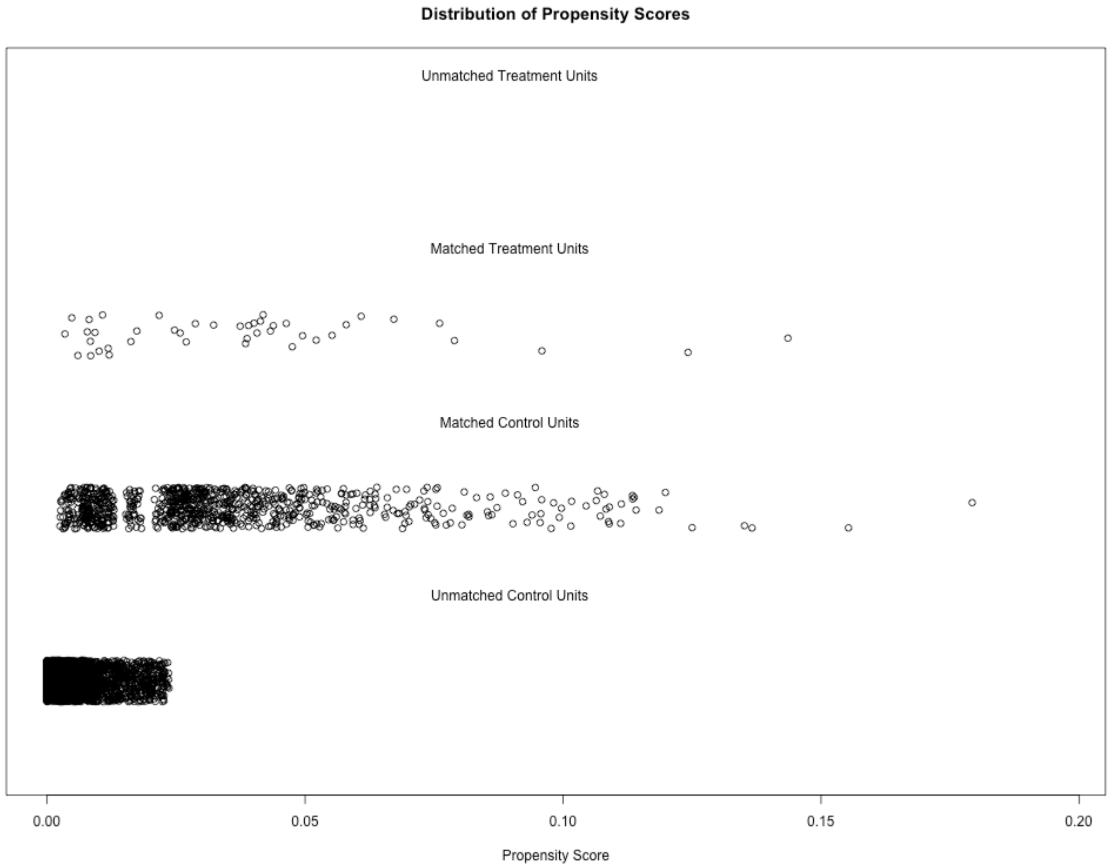


Figure 3. Plot of propensity score distribution of schools by treatment and matching status (1:25 matching)

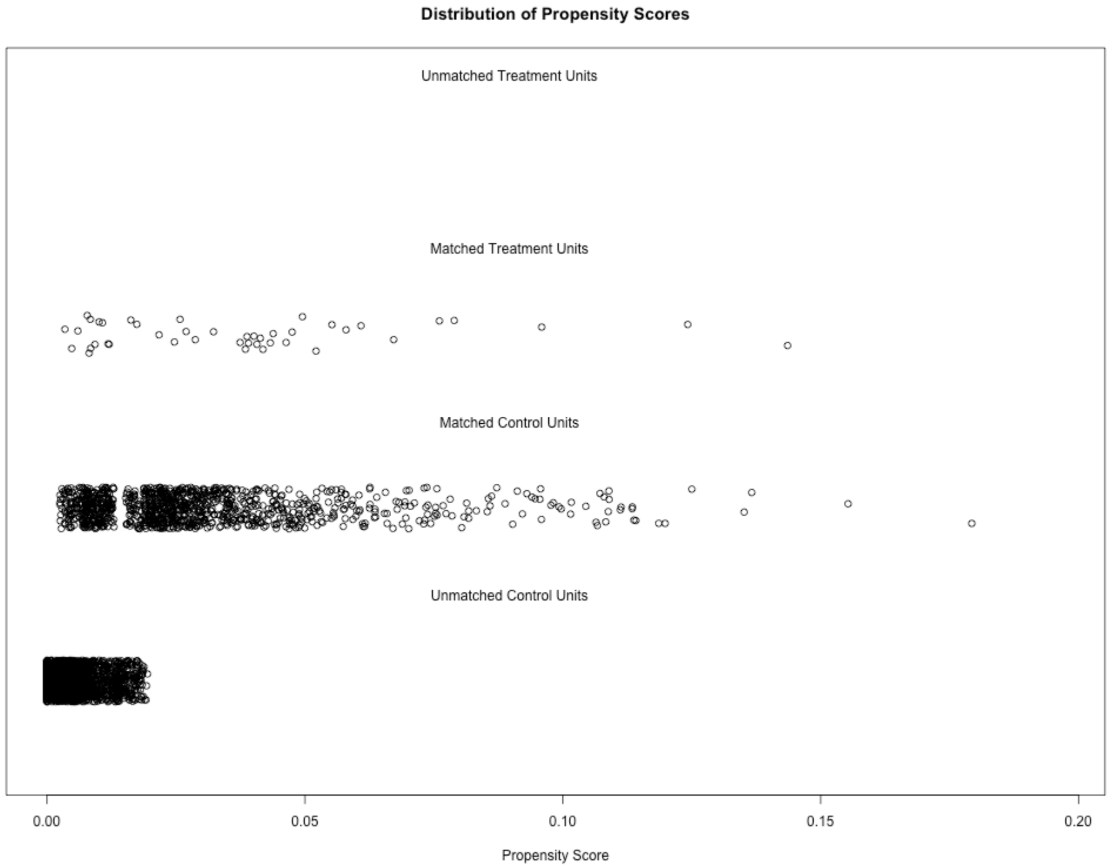


Table 1. Standardised differences in characteristics by matching methods

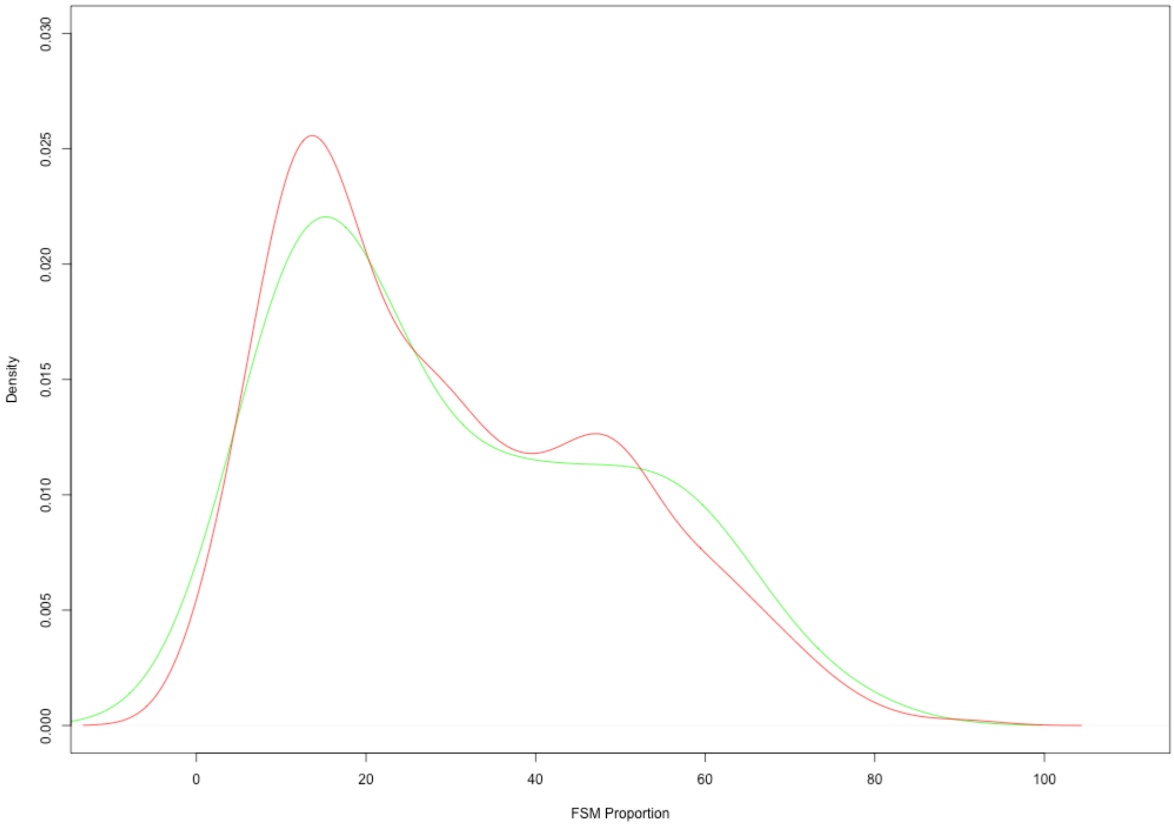
Characteristics	20 Matches	25 Matches	Unmatched
No. Pupils	0.10	0.08	0.28
Academy Proportion	-0.26	-0.28	-0.10
FSM Proportion	0.05	0.12	0.21
KS2 APS 2018	-0.01	-0.03	0.19
KS2 APS 2017	-0.02	-0.06	0.08
KS2 APS 2016	-0.02	-0.05	0.11
Low Attainers Prop.	-0.02	0.00	-0.21
High Attainers Prop.	-0.02	-0.04	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.

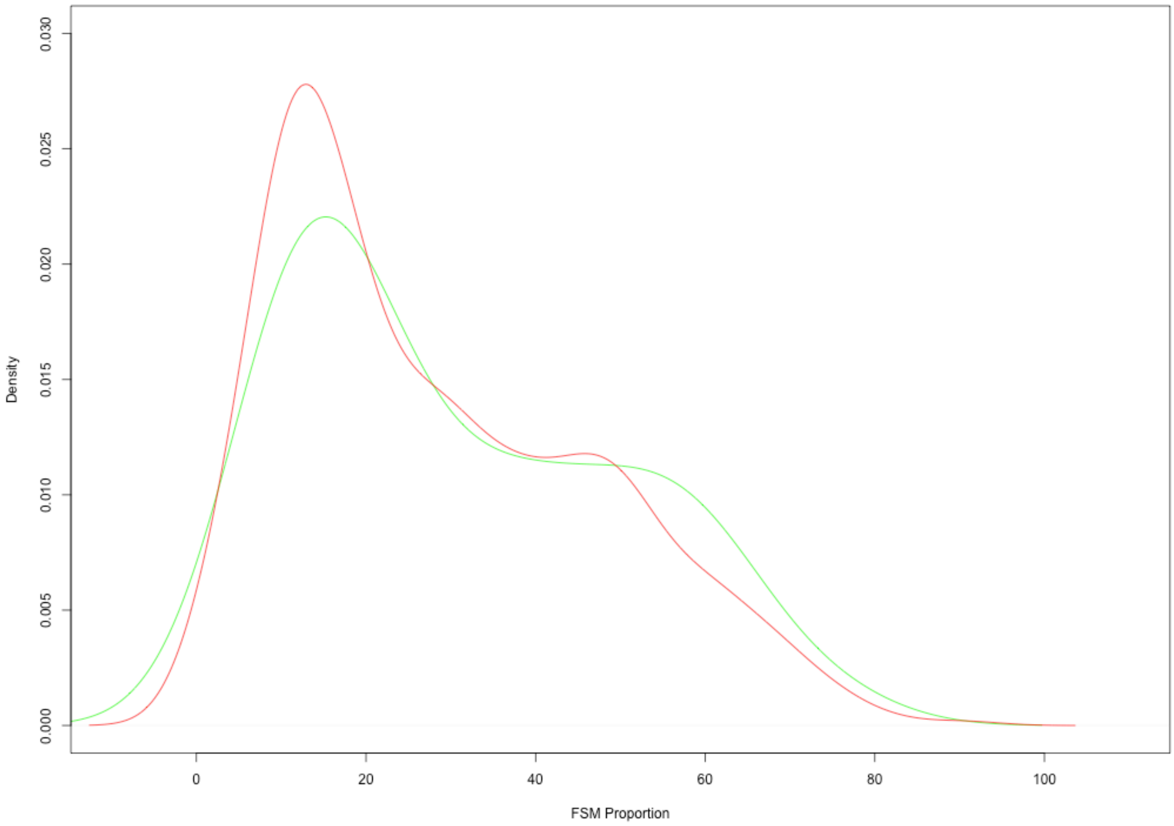
Overall, it is unclear that balance is substantially worse among the potential comparators in the case of 1:25 matching. Later in this document, we check that our simulated responses patterns among these potential comparators does not alter this picture.

Figure 4. Distribution of proportion of pupils identified as FSM in treatment and potential comparison groups (1:20 matching)



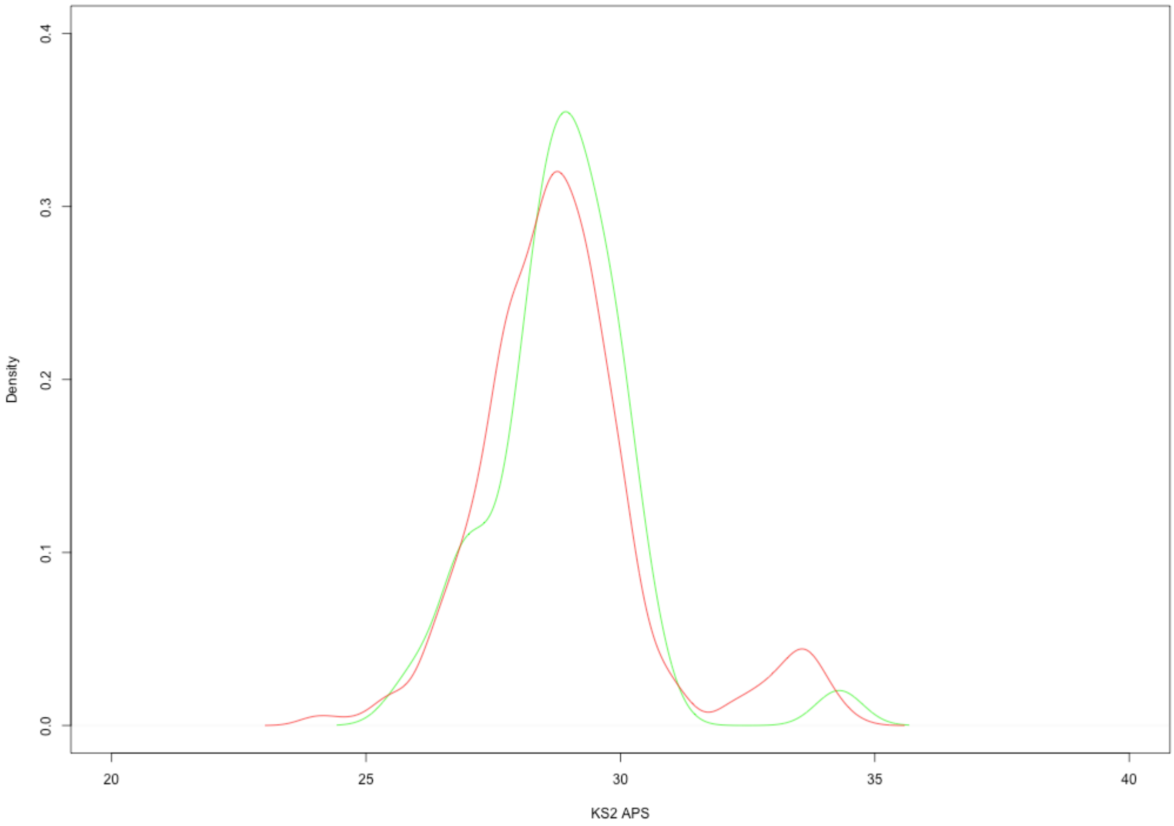
Notes. Kernel density plot of school FSM proportion for treated (green) and comparison (red) schools.

Figure 5. Distribution of proportion of pupils identified as FSM in treatment and potential comparison groups (1:25 matching)



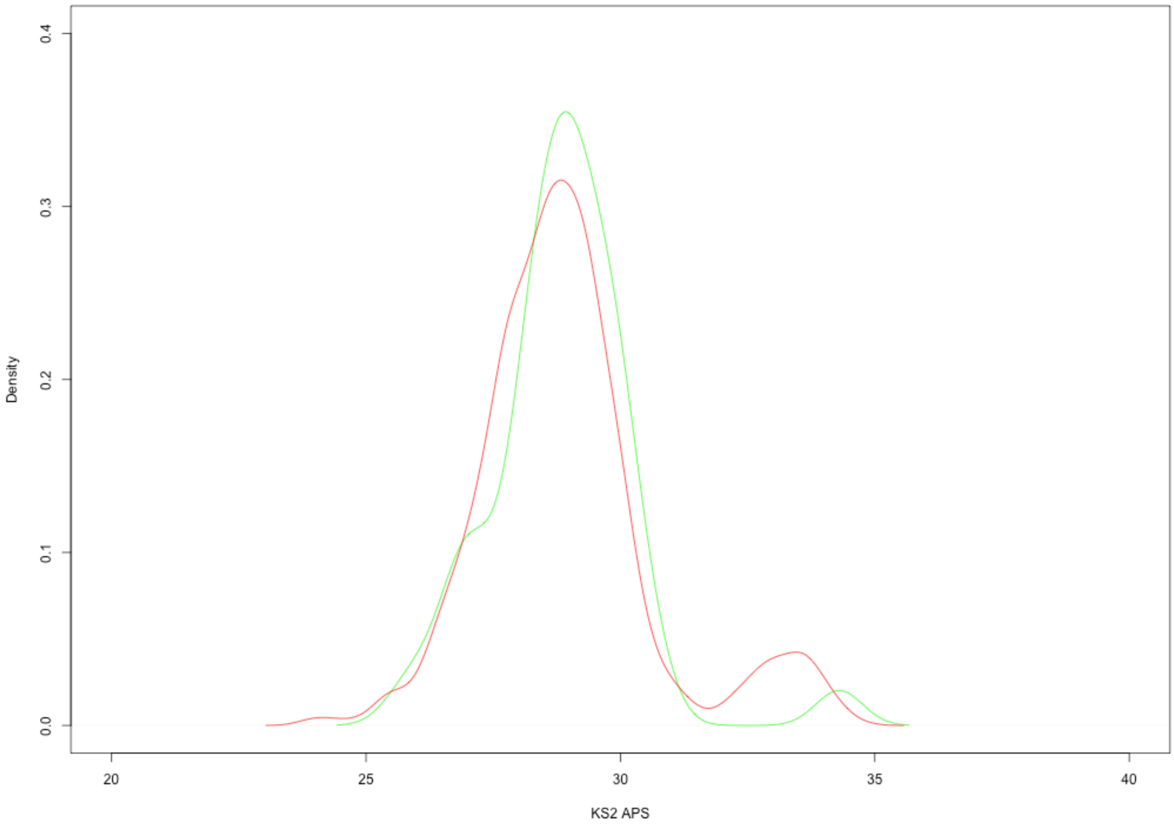
Notes. Kernel density plot of school FSM proportion for treated (green) and comparison (red) schools.

Figure 6. Distribution of average KS2 prior attainment in treatment and potential comparison groups (1:20 matching)



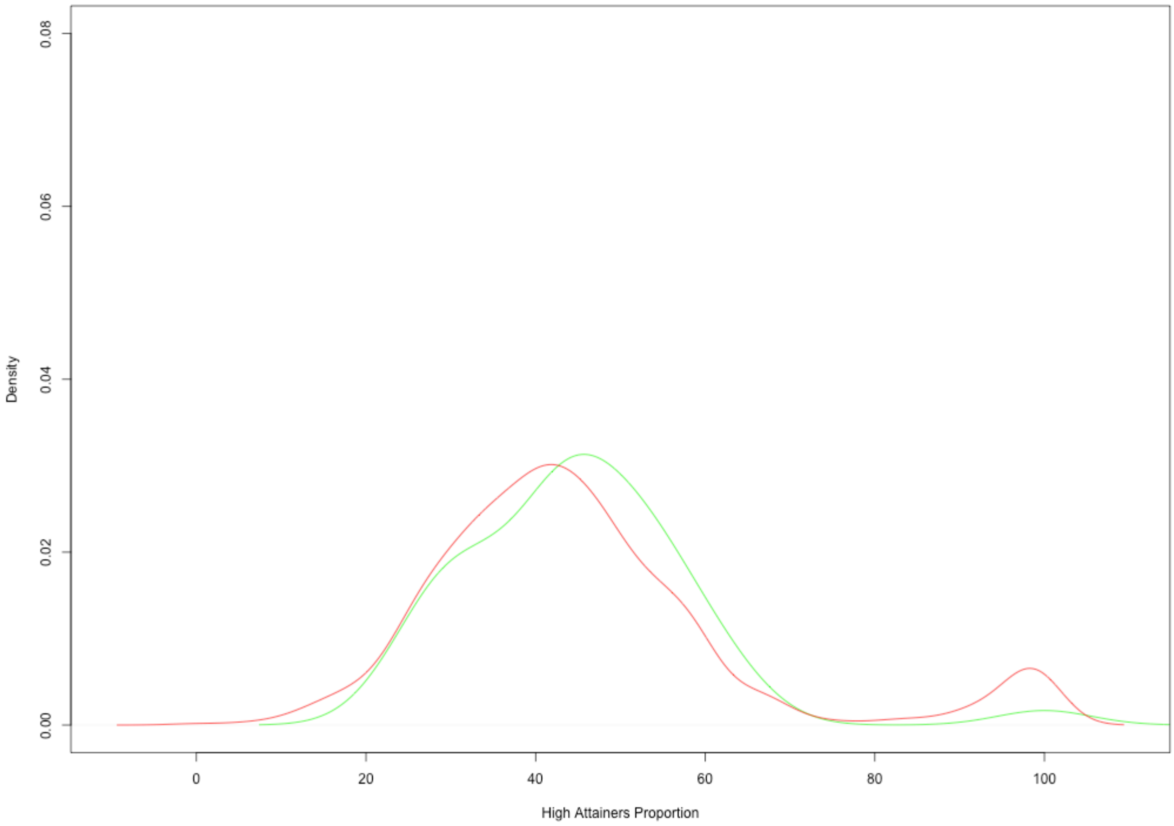
Notes. Kernel density plot of school average KS2 prior attainment for treated (green) and comparison (red) schools.

Figure 7. Distribution of average KS2 prior attainment in treatment and potential comparison groups (1:25 matching)



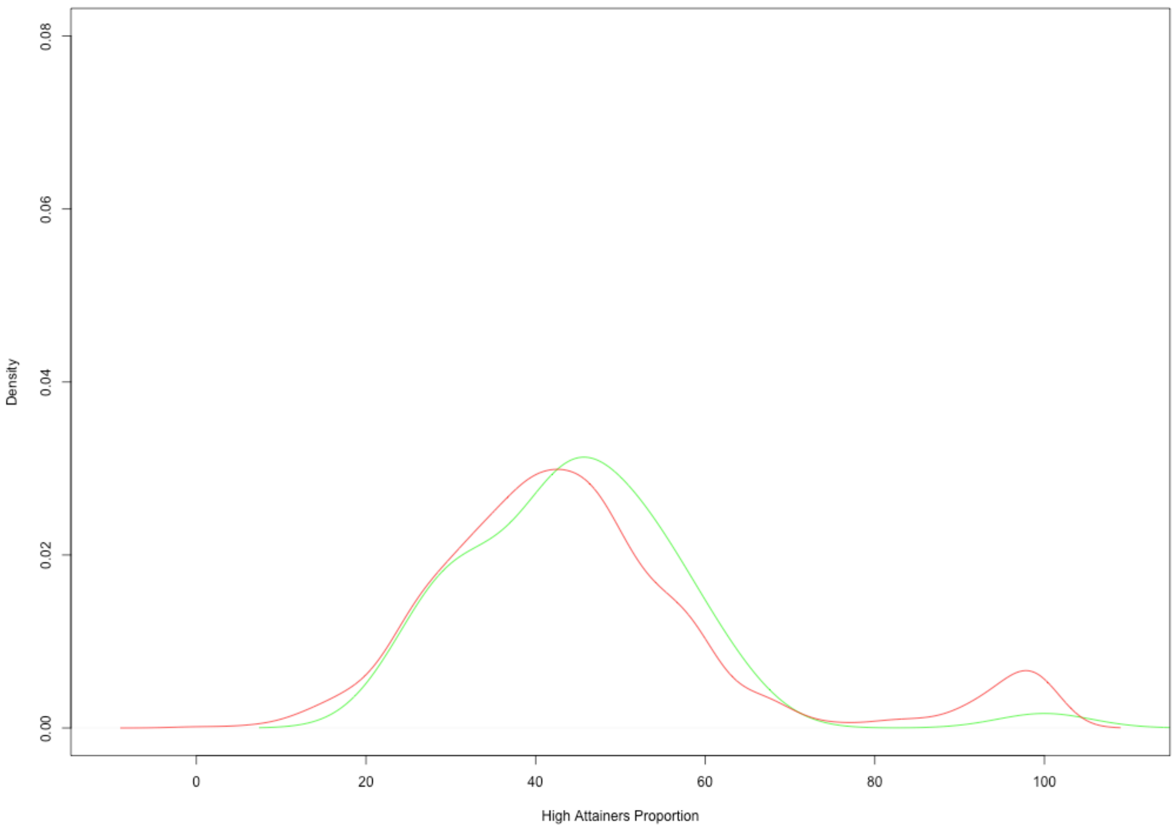
Notes. Kernel density plot of school average KS2 prior attainment for treated (green) and comparison (red) schools.

Figure 8. Distribution of proportion of pupils identified as high attainment on intake in treatment and potential comparison groups (1:20 matching)



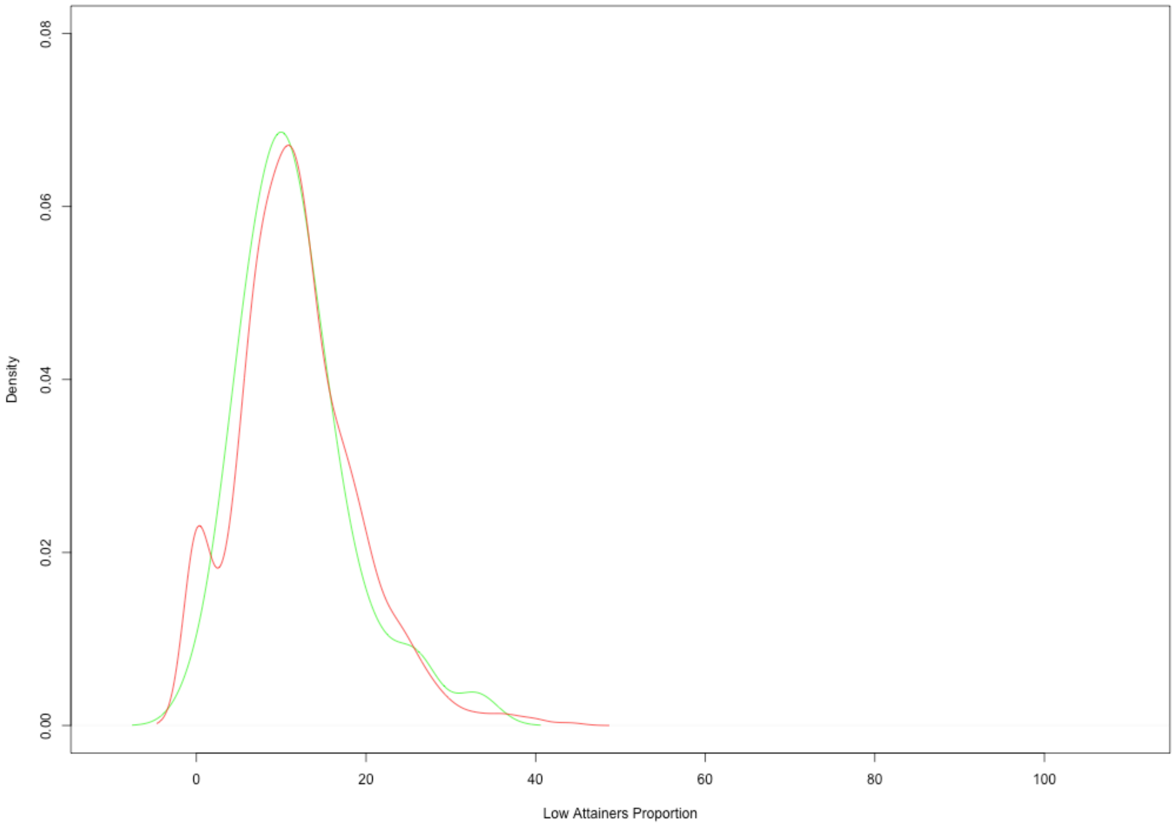
Notes. Kernel density plot of proportion of pupils identified as high attainment on intake by DfE for treated (green) and comparison (red) schools.

Figure 9. Distribution of proportion of pupils identified as high attainment on intake in treatment and potential comparison groups (1:25 matching)



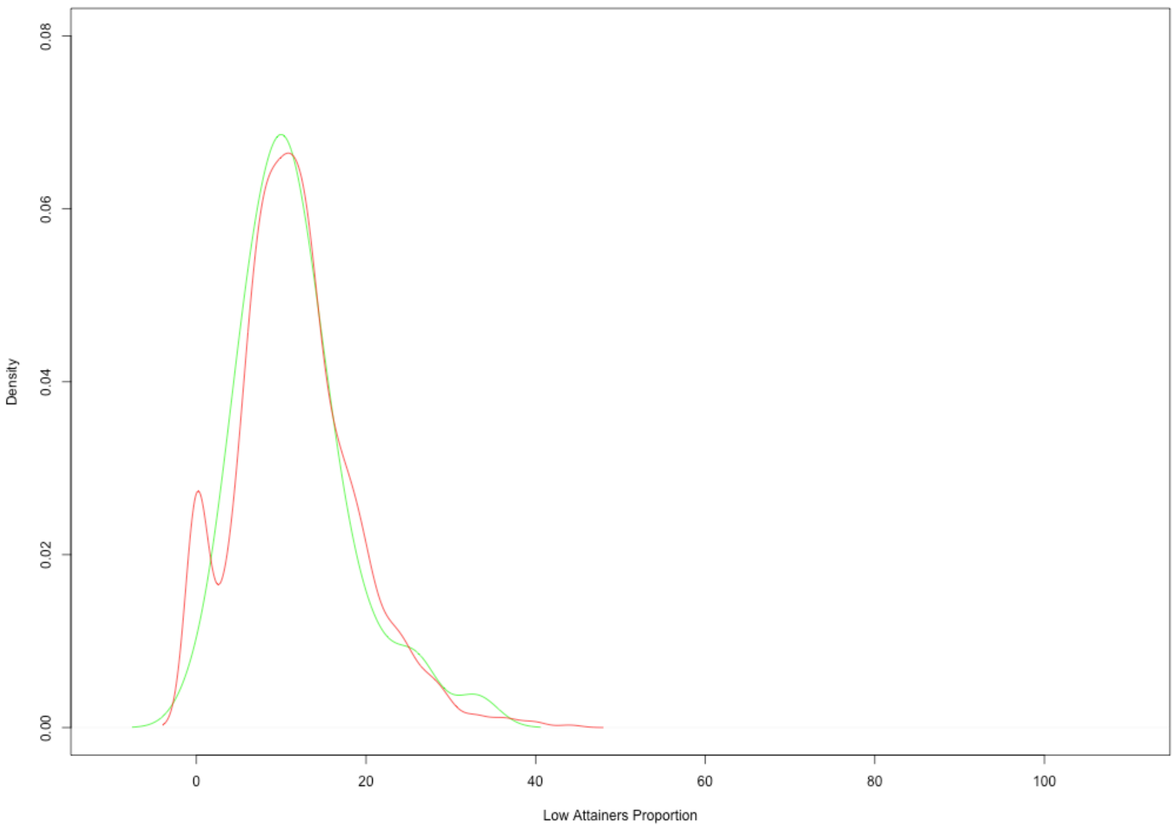
Notes. Kernel density plot of proportion of pupils identified as high attainment on intake by DfE for treated (green) and comparison (red) schools.

Figure 10. 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 11. 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 2. 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 3. 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.

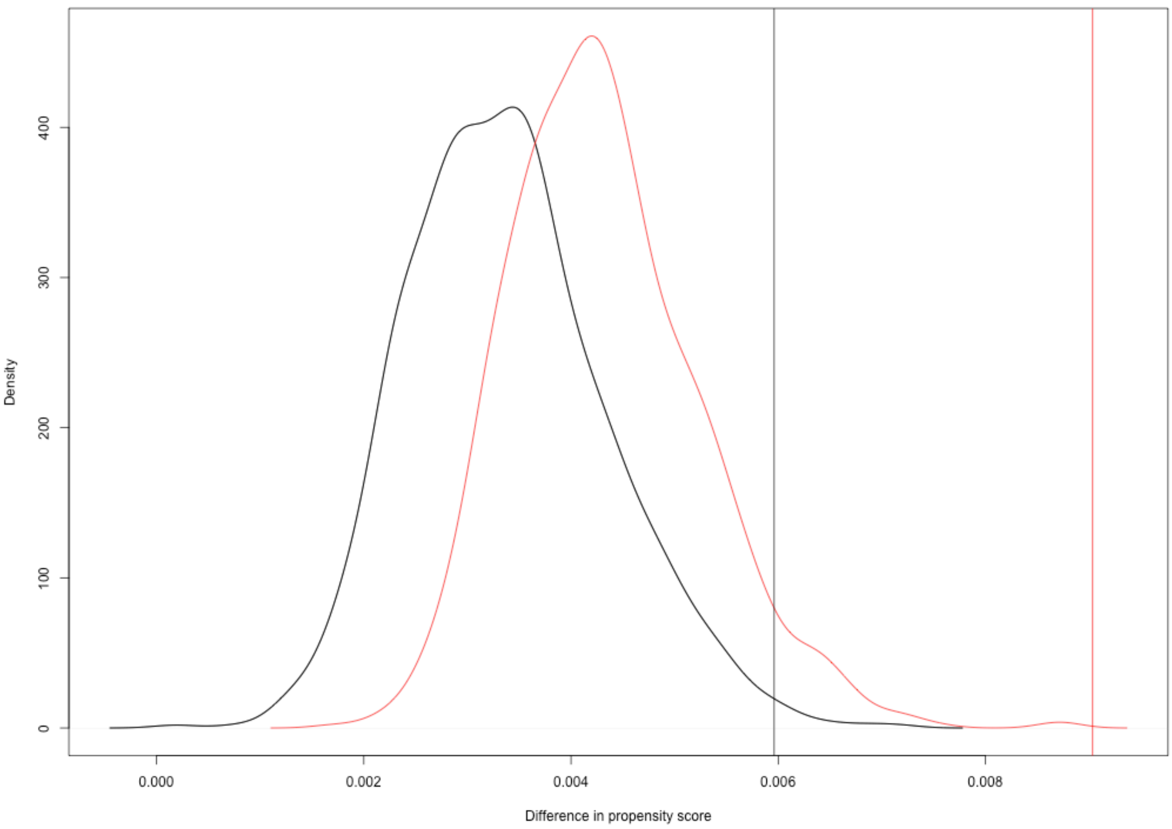
Figure 12 plots the distribution of the difference in propensity score between treatment and simulated recruited schools by whether 1:20 or 1:25 matching was carried out in the initial matching process. Figures 13 to 18 repeat this but for standardised difference measures of imbalance in key characteristics. 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.

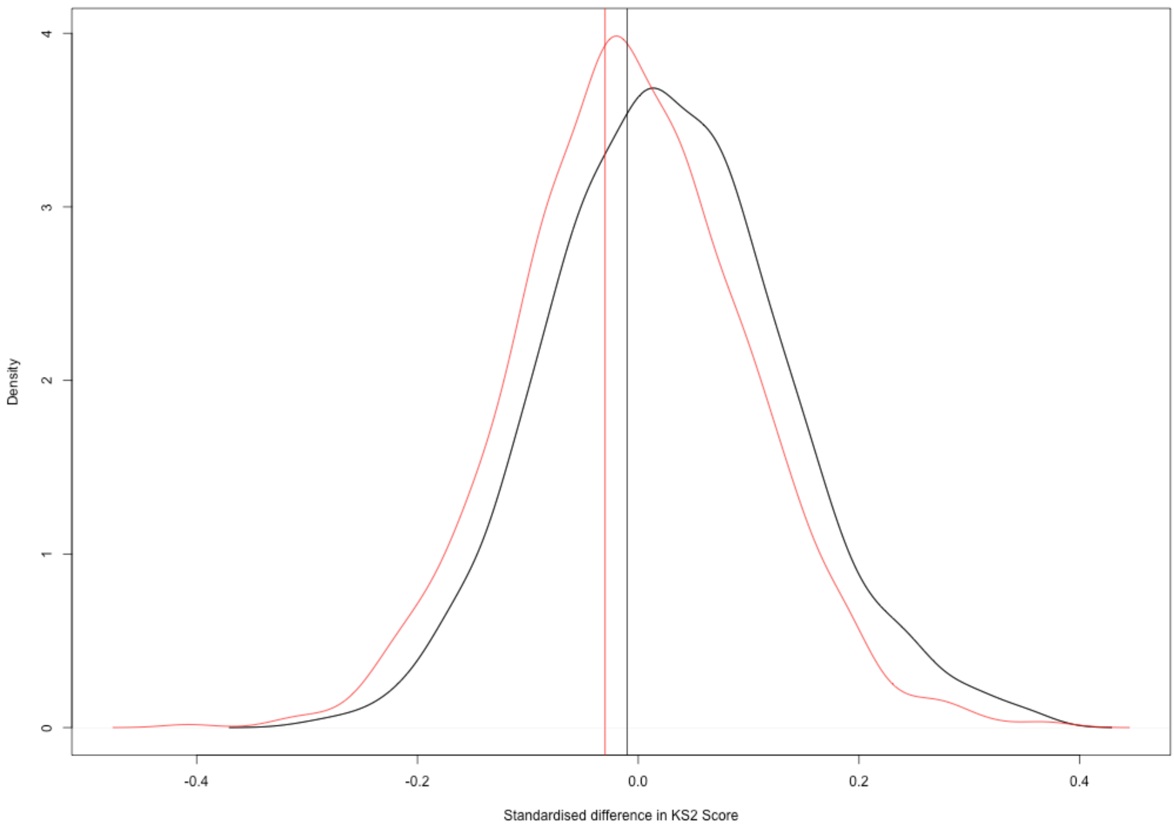
The exercise has also emphasised the particular importance of estimating propensity scores using a parsimonious model as part of this project, given the small number of mixed attainment schools available. This makes prioritisation of characteristics on which we need to achieve a good match to have confidence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

Figure 12. Simulated density of imbalance in propensity score measures after response: comparing 1:20 (black) to 1:25 (red) matching



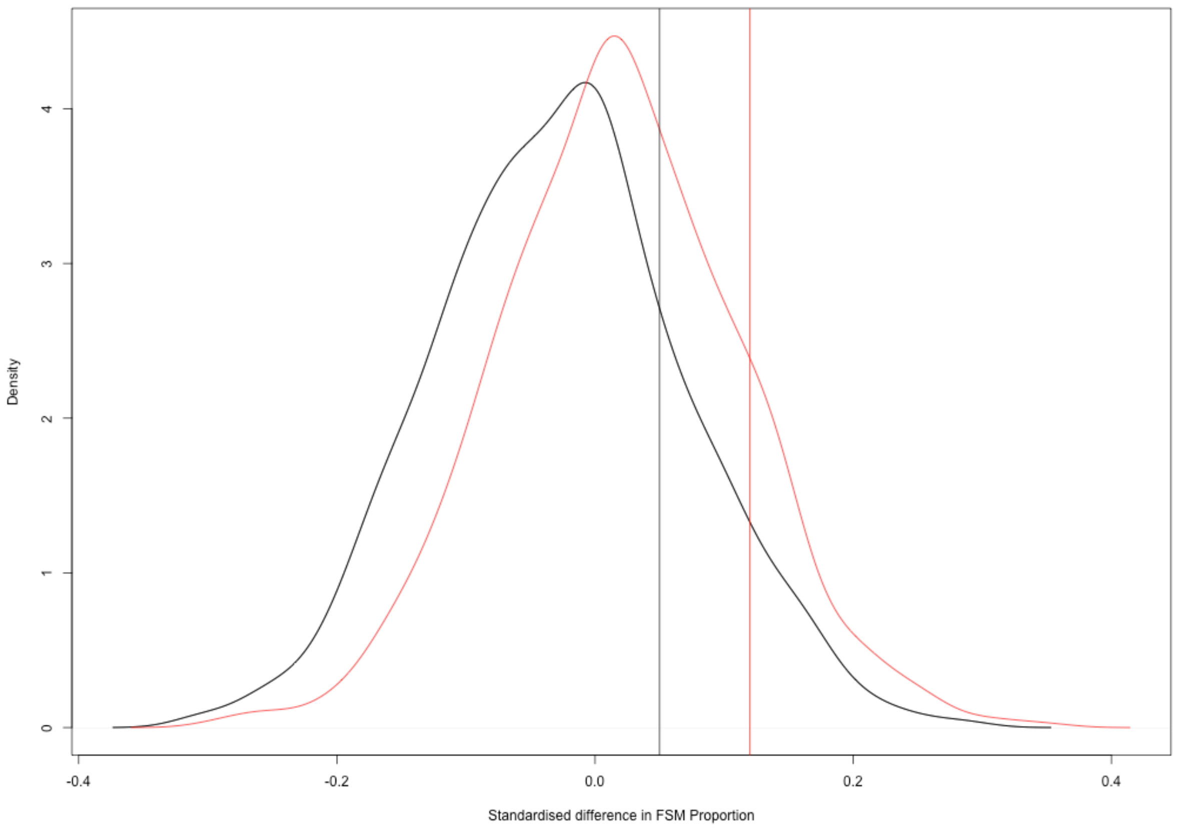
Notes. Density plots showing difference in propensity score between treated and simulated responders among the matched comparison sample. Simulations based on 1:20 matched sample plotted in black; simulations based on 1:25 matched sample plotted in red. Simple imbalance from full sample of potential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

Figure 13. Simulated density of standardised imbalance in KS2 score after response: comparing 1:20 (black) to 1:25 (red) matching



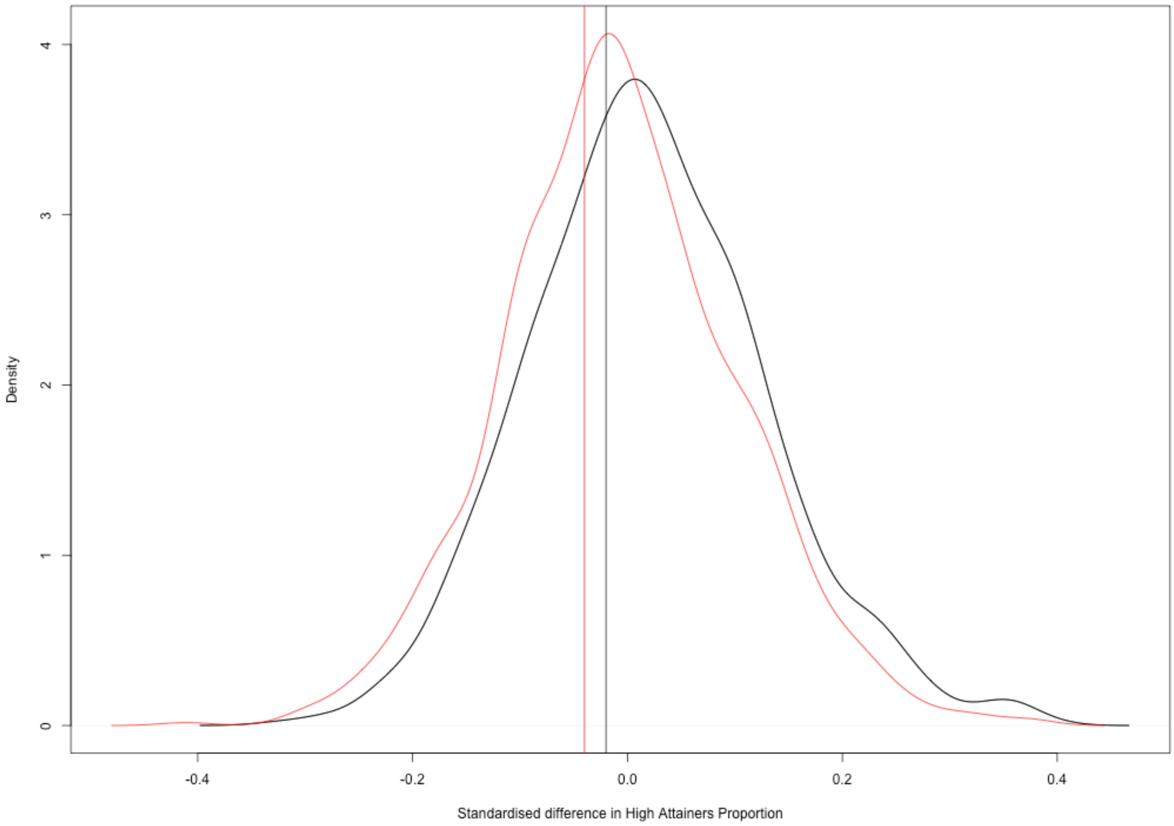
Notes. Density plots showing standardised difference in KS2 score between treated and simulated responders among the matched comparison sample. Simulations based on 1:20 matched sample plotted in black; simulations based on 1:25 matched sample plotted in red. Simple imbalance from full sample of potential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

Figure 14. Simulated density of standardised imbalance in FSM proportion after response: comparing 1:20 (black) to 1:25 (red) matching



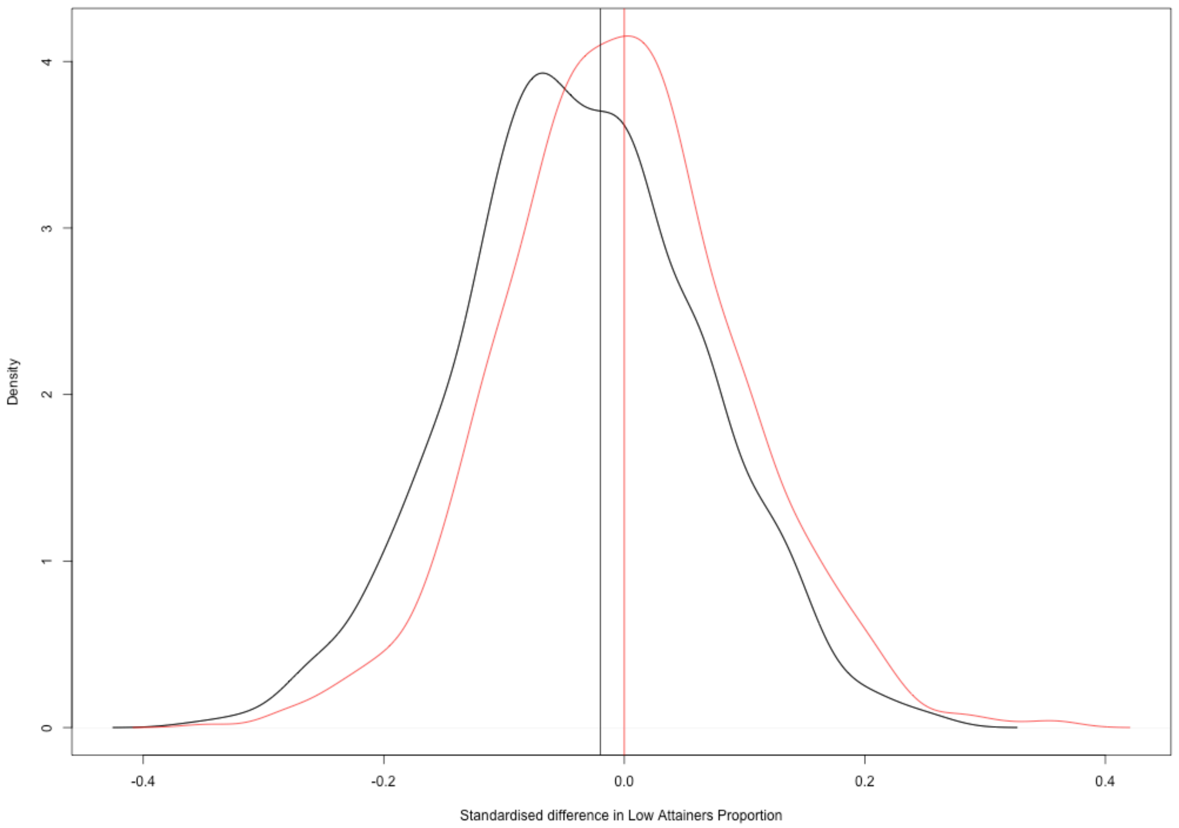
Notes. Density plots showing standardised difference in FSM proportion between treated and simulated responders among the matched comparison sample. Simulations based on 1:20 matched sample plotted in black; simulations based on 1:25 matched sample plotted in red. Simple imbalance from full sample of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

Figure 15. Simulated density of standardised imbalance in proportion of high attainers after response: comparing 1:20 (black) to 1:25 (red) matching



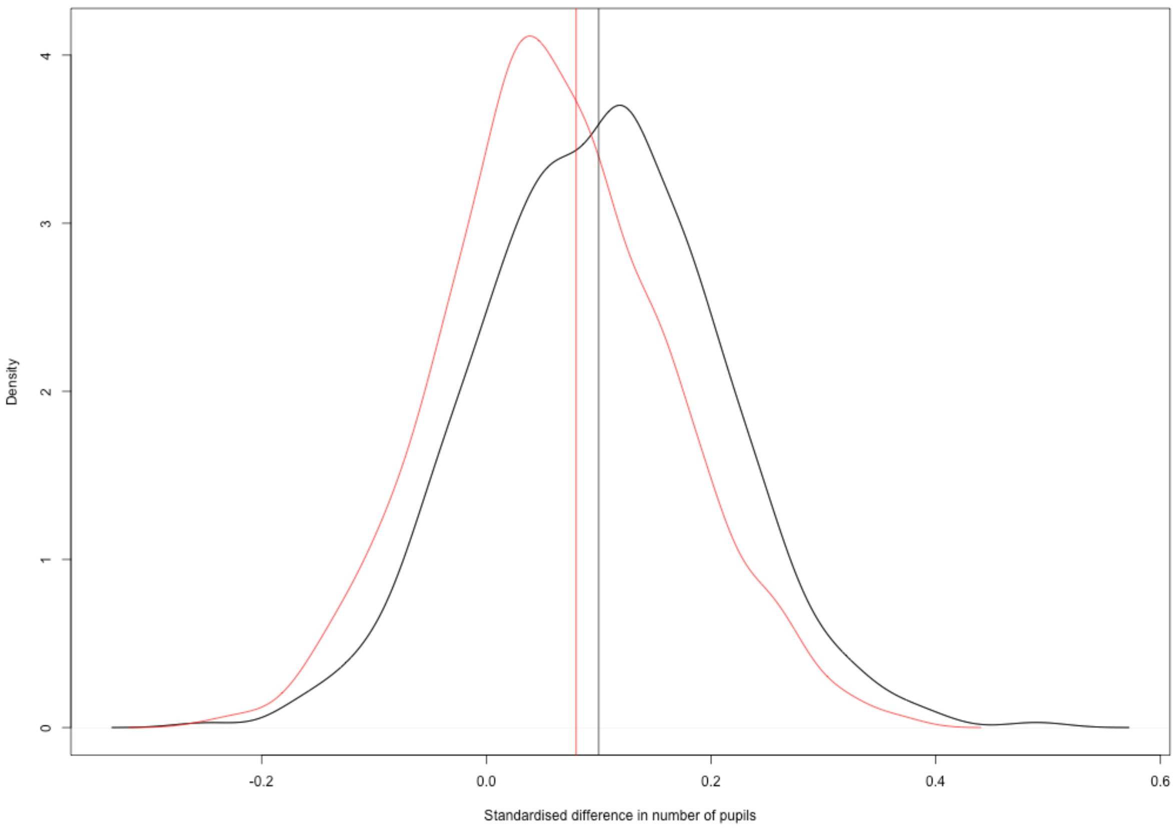
Notes. Density plots showing standardised difference in proportion of high attainers between treated and simulated responders among the matched comparison sample. Simulations based on 1:20 matched sample plotted in black; simulations based on 1:25 matched sample plotted in red. Simple imbalance from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

Figure 16. Simulated density of standardised imbalance in proportion of low attainers after response: comparing 1:20 (black) to 1:25 (red) matching



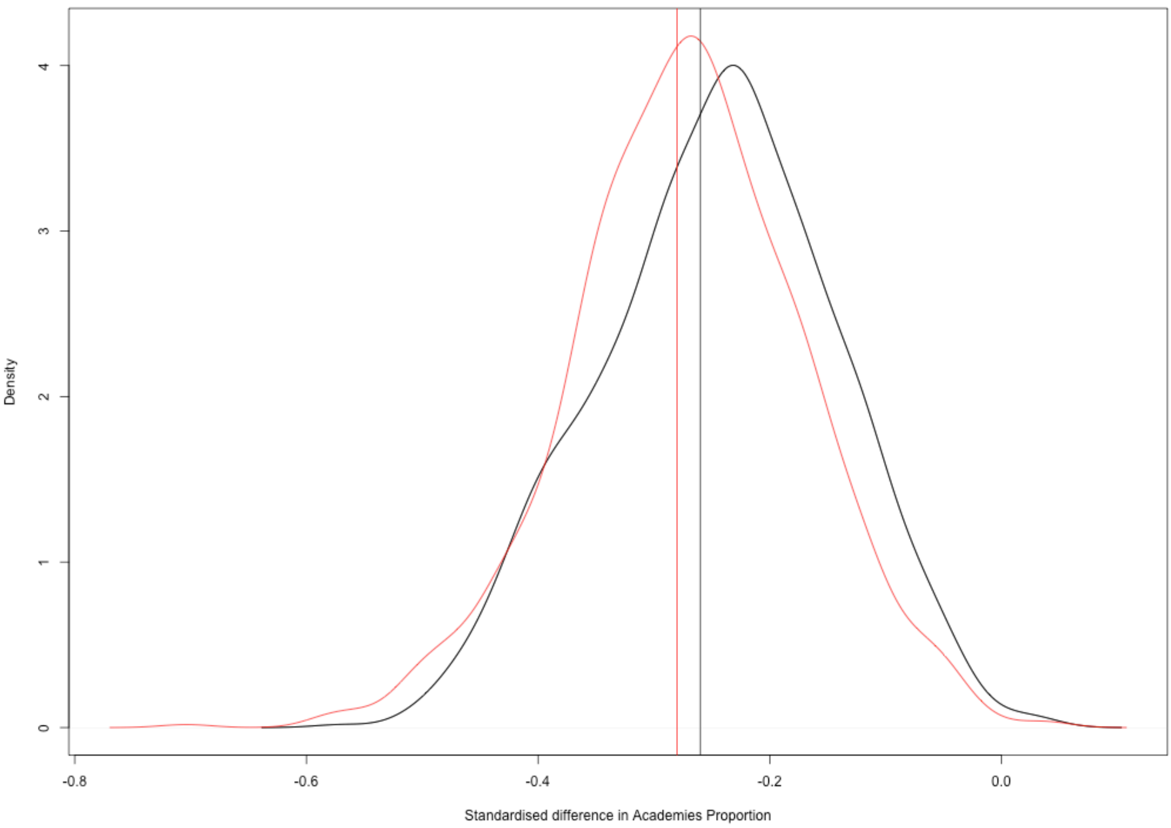
Notes. Density plots showing standardised difference in proportion of low attainers between treated and simulated responders among the matched comparison sample. Simulations based on 1:20 matched sample plotted in black; simulations based on 1:25 matched sample plotted in red. Simple imbalance from full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

Figure 17. 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 18. 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 (SAP) 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.