The Student Grouping Study: investigating the effects of setting and mixed attainment grouping
Statistical Analysis Plan

Evaluator (institution): UCL Institute of Education
Principal investigator(s):
Jeremy Hodgen & Becky Taylor

Template last updated: August 2019

PROJECT TITLE	The Student Grouping Study: investigating the effects of setting and mixed attainment grouping
DEVELOPER (INSTITUTION)	n/a
EVALUATOR (INSTITUTION)	IOE, UCL's Faculty of Education and Society
PRINCIPAL INVESTIGATOR(S)	Professor Jeremy Hodgen Dr Becky Taylor
SAP AUTHOR(S)	Dr Jake Anders Professor Jeremy Hodgen Dr Becky Taylor Dr Nicola Bretscher
STUDY DESIGN	Matched design study in a natural context
STUDY TYPE	School Choices
PUPIL AGE RANGE AND KEY STAGE	11-13, Key Stage 3
NUMBER OF SCHOOLS	112 schools (30 mixed attainment schools, 82 matched setting schools)
NUMBER OF PUPILS	21370
PRIMARY OUTCOME MEASURE AND SOURCE	Attainment in mathematics (GL Assessment Progress Test in mathematics)
SECONDARY OUTCOME MEASURE AND SOURCE	Self-confidence in mathematics (Survey developed by the research team)

SAP version history

VERSION	DATE	REASON FOR REVISION
1.0 [*original*]	24/07/2024	*N/A*

Table of Contents

Introduction

The study uses a matched design in a natural context, to explore the difference in student outcomes of two approaches to grouping students for mathematics: grouping by subject attainment (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.

This SAP should be read in conjunction with the study plan, in which the rationale for the study is described in detail.

The study addresses the following research questions to investigate the effects of setting and mixed attainment grouping on student attainment and attitude:

RQ1. What difference is there, if any, in the attainment of low-attaining students over Years 7 and 8 attending schools that use mixed attainment grouping for mathematics, and low-attaining students over Years 7 and 8 attending a similar group of schools that use setting?

RQ2. What difference is there, if any, in the attainment of *all* students over Years 7 and 8 attending schools that use mixed attainment grouping for mathematics, and *all* students over Years 7 and 8 attending a similar group of schools that use setting? (a) for all students; (b) for students receiving FSM.

RQ3. What difference is there, if any, in the mathematics self-confidence of low-attaining students over Years 7 and 8 attending schools that use mixed attainment grouping for mathematics, and low-attaining students over Years 7 and 8 attending a similar group of schools that use setting? (a) for all low-attaining students; (b) for low-attaining students receiving FSM.

RQ4. What difference is there, if any, in the mathematics self-confidence of *all* students over Years 7 and 8 attending schools that use mixed attainment grouping for mathematics, and *all* students over Years 7 and 8 attending a similar group of schools that use setting? (a) for all students; (b) for students receiving FSM.

RQ5. To what extent do (a) opportunity to learn and (b) teacher quality explain any differential outcomes for different grouping practices, for different set allocations or for students at different attainment levels?

Four further research questions address implementation and process evaluation (IPE) and are described in the study plan.

Note: For the purposes of this study, low attaining students will be defined as the lowest attaining tertile at entry to Year 7. This broadly corresponds to those students who do not achieve at the expected level at KS2 (as per the DfE NPD).[1] High attaining students will be defined as the highest attaining tertile.

Design overview

Trial design, including number of arms	Two-arm, matched design study in a natural context, students clustered in schools

[1] See: https://explore-education-statistics.service.gov.uk/find-statistics/key-stage-2-attainment-national-headlines/2023-24

Unit of matching		School
Stratification variables (if applicable)		Schools recruited into 20 matched groups based on matching variables as described below
Primary outcome	variable	Attainment in mathematics
	measure (instrument, scale, source)	GL Assessment Progress Test in mathematics (PTM13) [Scale range: 71-136][2]
Secondary outcome(s)	variable(s)	Self-confidence in mathematics
	measure(s) (instrument, scale, source)	Survey developed by the research team: Seven-item scale, using five-point Likert scale responses. Scores from 7 to 35.[3]
Baseline for primary outcome	**variable**	Attainment in mathematics
	measure (instrument, scale, source)	Key Stage 2 mathematics raw score [Scale range: 0 – 110]
Baseline for secondary outcome	**variable**	Self-confidence in mathematics
	measure (instrument, scale, source)	Survey developed by the research team (Francis et al., 2017); 7 items, Likert scale 1-5, total score range 7-35.

This study has been designed as an embedded mixed methods evaluation, using a matched design in a natural context to investigate differences in the effects on student outcomes of two different approaches to grouping students in mathematics. This is combined with an investigation of the implementation of the two approaches (as set out in the study plan). This approach to the impact evaluation builds on the results of two previous trials (Best Practice in Setting and Best Practice in Mixed Attainment) conducted by the research team as discussed in the study plan.[4] These evaluations demonstrated that introducing changes to grouping practices is complex and takes time and effort over a sustained period to establish in a school, so is not amenable to evaluation by randomised controlled trial (see, e.g., Anders et al., 2017, on the evaluation of complex whole school interventions).[5]

[2] https://support.gl-assessment.co.uk/media/3035/ptm-technical-information.pdf
[3] See 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.2017.09.001for further information.
[4] 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.
[5] 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. Report to the Education Endowment Foundation.* London: EEF.

A randomised controlled trial would not have been feasible for this study. The introduction of mixed attainment teaching appears to be particularly challenging (see, Taylor et al., 2017)[6] and, hence, takes considerable time to prepare for and establish (and, thus, to demonstrate effects). Hence, an intervention introducing mixed attainment teaching at scale could not currently be achieved within an acceptable budget or timescale. Nevertheless, some schools do operate mixed attainment grouping and this provides the potential for a comparison of schools who have already adopted the practice with equivalent, or matched, schools who use setting (grouping students by attainment in mathematics).

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

Recruitment and matching

Recruitment was conducted for the mixed attainment and setted groups of schools in distinct, sequential phases in order to ensure balance across the two groups. The mixed attainment schools were recruited first, because these schools are relatively rare in England. On the basis of survey evidence (Taylor et al., 2022), we estimated that there are between 120 and 190 eligible schools in England teaching mathematics in mixed attainment groups to Year 7 and Year 8.[8] This phase of recruitment involved all state secondary schools in England. Following this, we conducted a matching exercise to identify a pool of matched schools, potentially eligible for the setted arm of the study. In the case of school-level selection effects, however, Weidmann and Miratrix (2021) provide evidence of the effectiveness of a matching approach.[9] This matched pool consisted of sub-sets of 25 schools, with each sub-set matched to one of the mixed attainment schools. The setted schools were then recruited from this pool of matched and potentially eligible schools. This is described in detail below.

Recruitment of mixed attainment schools

For the restarted trial, schools were recruited to the Mixed Attainment group first. Recruitment started on 11 January 2022. We began by contacting all schools in the Mixed Attainment group recruited in 2019, aside from those schools that had previously indicated that they had either changed their grouping practices or no longer wished to take part in the study. Of these schools, 17 agreed to take part in the study. Following this, it was necessary to recruit further mixed attainment schools to achieve an acceptable level of power. As noted above, we estimated that there are between 120 and 190 eligible schools in England teaching

[6] 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; see also: 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*, 1-20. doi:10.1080/00071005.2018.1424317

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

[8] Taylor, B., Hodgen, J., Tereshchenko, A., & Gutiérrez, G. (2022). Attainment grouping in English secondary schools: A national survey of current practices. *Research Papers in Education, 37*(2), 199-220. https://doi.org/10.1080/02671522.2020.1836517

[9] Weidmann, B. & Miratrix, L., 2021. Lurking inferential monsters? Quantifying selection bias in evaluations of school programs. *Journal of Policy Analysis and Management, 40*(3), 964-986.

mathematics in mixed attainment groups to Year 7 and Year 8. we did not know the identity of all these schools in advance, recruitment took place through simultaneous processes of publicising the study in education media, using existing contacts and networks to advertise the study to schools, and cold-calling schools to ask about grouping practices. As a result, a further 15 schools were identified as teaching mathematics to Year 7 and Year 8 students in mixed attainment classes and agreed to participate in the study by returning a completed Memorandum of Understanding, resulting in a mixed attainment group of 32 schools.

MATCHING

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

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

PSM produced a subset of 25 potential matched schools for each mixed attainment school, a list with a total of 800 schools which was used as the basis for recruitment.

See Appendix 1 for further details on the matching process.

RECRUITMENT OF SETTING SCHOOLS

Schools were then recruited to the Setting group. Recruitment took place between 17 June and 25 November 2022.

All 800 matched schools were approached by email to take part in the study. The hope was to recruit three matched comparison schools for each recruited mixed attainment school. Of the schools that responded to these invitations, 80 were judged eligible and agreed to take part in the study, a response rate of 10% of the total schools approached (which is below the positive response rate of 20% as estimated in the study plan, but above the 7.3% rate achieved in the initial 2019 recruitment exercise). The number of matched schools recruited for each mixed attainment school varied from 2 to 7.

The final set of mixed attainment schools and matched comparison schools as at the 25 November 2022 (the date recruitment closed, aka pseudo-randomisation date) is used as the school-level analysis sample. Given the variable number of recruited matched comparison schools for each mixed attainment school, we will use weighting to reduce the importance for estimation of pupils in schools where we recruited multiple matched comparator schools for each mixed attainment school; the sample will be weighted (to reflect the number of matched comparator schools recruited corresponding to each mixed attainment school and to reflect the number of pupils in each school) such that pupils in each mixed attainment school have the same level of importance as all the schools that have been identified and recruited as its matched comparators.

We were unable to recruit a match for two of the mixed attainment schools. These two schools are therefore excluded from the balance tables and the power calculations for the

proposed primary analysis. Additionally, one of the remaining 30 mixed attainment schools and seven of the 80 setted schools did not submit pupil data by 25 November 2022. Three of the remaining setted schools were a match for the mixed attainment school that did not submit data on time.[10] Hence, our primary analysis will be conducted with 99 schools, 29 mixed attainment and 70 setted. These are in 29 matched groups as set out in Table 1.

Table 1: Matched groups of schools in recruited sample of schools with primary data

Matched Group	No of mixed attainment schools	No of setted schools	Total schools
1	1	5	6
2	1	1	2
3	1	2	3
4	1	2	3
5	1	2	3
6	1	1	2
7	1	3	4
8	1	1	2
9	1	1	2
10	1	1	2
11	1	1	2
12	1	2	3
13	1	2	3
14	1	4	5
15	1	1	2
16	1	2	3
17	1	6	7
18	1	2	3
19	1	3	4
20	1	1	2
21	1	4	5
22	1	5	6
23	1	3	4
24	1	5	6
25	1	2	3
26	1	1	2
27	1	2	3
28	1	3	4
29	1	2	3
Total	**29**	**70**	**99**

IMBALANCE AT RECRUITMENT

Standardised differences, calculated using the full sample standard deviations, are calculated in:

- Full analysis sample (unmatched), ie all schools in England),
- All matched schools, i.e., the sample restricted to mixed attainment and matched comparator schools (all potential matched), and

[10] The five schools, which submitted data but were excluded from the primary analysis, were kept in the study for the investigation of implementation as well as both the pupil-level and sensitivity analyses.

- Recruited and primary data matched, i.e., the recruited sample restricted to matched comparator schools and mixed attainment schools from which primary data have been collected, excluding schools in strata in which primary data have not been successfully obtained (primary data matched).

Weights are applied to reflect the number of matched comparator schools recruited corresponding to each mixed attainment school and to reflect the number of pupils in each school.

Our aim was to achieve standardized differences of <0.1 on the most recent available school attainment characteristics: KS2 for 2019, 2018 and 2017 and the proportions of low and high attaining pupils. [11] Weighted imbalance is presented in Table 2. See Appendix 2, Table 5, for imbalance weighted only for the number of comparator schools.

Overall, the balance is good, imbalance is <0.1 for all the attainment characteristics, aside from the proportion of high attainers, which is just above the threshold at 0.132. However, the average balance of attainment characteristics is well within the threshold. See Appendices 1 and 2 for further details of the matching process and the achieved balance.

We will report imbalance at pupil-level using NPD data for the entire sample and the sub-sample of students eligible for FSM. See Table 2 for estimates prior to collecting pupil-level data via the NPD.

Table 2: Imbalance at recruited for sample weighted for number of matched comparator schools and the number of pupils in each school

Characteristic	Full analysis sample (Unmatched)	All matched schools	Recruited and primary data matched
KS2 2019	0.051	-0.025	-0.023
KS2 2018	0.051	-0.025	-0.023
KS2 2017	0.037	-0.029	-0.028
Low Prior Attain Prop.	-0.096	0.022	0.006
High Prior Attain Prop.	0.000	-0.084	-0.132
Average (Attainment)	***0.009***	***-0.028***	***-0.040***
Number of pupils on school roll[12]	0.235	-0.129	-0.270
FSM Prop.	-0.113	0.150	0.198
EAL Prop.	0.443	0.629	0.383
Academy status	-0.230	-0.298	-0.326
IDACI1	0.053	-0.118	-0.165
IDACI2	-0.046	-0.057	-0.104
IDACI4	-0.005	0.018	0.156
IDACI5	-0.123	0.075	-0.009
Ofsted Outstanding	0.174	0.024	-0.266
Ofsted Good	0.052	0.016	0.061
Urban	-0.192	-0.105	-0.214
Average	***0.119***	***0.113***	***0.148***

[11] Nguyen, T.-L., Collins, G. S., Spence, J., Daurès, J.-P., Devereaux, P. J., Landais, P., & Le Manach, Y. (2017, 2017/04/28). Double-adjustment in propensity score matching analysis: choosing a threshold for considering residual imbalance. *BMC Medical Research Methodology, 17*(1), 78. https://doi.org/10.1186/s12874-017-0338-0

[12] Weighting have been applied for the number of pupils in each school. Nevertheless, this weighting has a minimal effect on the imbalance in the the size of schools.

Sample size calculations overview

Table 3 shows the MDES calculations at design and for the final recruited sample for both the overall sample and the sub-sample of FSM students. MDES calculations were carried out using the R package PowerUpR. Unfortunately, in standard statistical software including R and Stata, there is no package, routine or function available to estimate the MDES for the 2-level model specified in the primary analysis section below in the context of a matched sample. Hence, we based these calculations on the closest model specified in PowerUpR: a 3-level fixed effects blocked cluster random assignment design (with treatment at level 2) to account for clustering (of students within schools) and blocking (schools within matched groups at level 3). Note that this differs slightly from the approach taken in the original design (see version 1 of the study plan) which was based on a 2-level multi-level model. A 3-level model is a better reflection of the structure of clustering arising from the actual recruitment and matching process and, hence, provides less biased estimate of the power of the trial. The primary analysis model accounts for level 3 via dummy variable for each block. See Appendix 3 for R code.

Table 3: MDES calculations at design

		Study Plan		Recruited Matched Sample	
		OVERALL	FSM / Low Prior Attainment	OVERALL	FSM / Low Prior Attainment
Minimum Detectable Effect Size (MDES)		0.199	0.207	0.241	0.248
Pre-test/ post-test correlations	level 1 (pupil)	0.75	0.75	0.75	0.75
	level 2 (school)	0.38	0.38	0.38	0.38
Intracluster correlations (ICCs)	level 2 (school)	0.15	0.15	0.15	0.15
Alpha		0.05	0.05	0.05	0.05
Power		0.8	0.8	0.8	0.8
One-sided or two-sided?		Two-sided	Two-sided	Two-sided	Two-sided
Average cluster size (pupils in schools)		100	25	190[13]	40
Average cluster size (matched groups of schools)		-	-	3.4	3.4
Number of schools	intervention	40	40	29	29
	comparison	80	80	70	70
	total	120	120	99	99
Number of pupils	intervention	4000	1000	5510	1160
	comparison	8000	2000	13300	2800
	total	12000	3000	18810	3960

[13] Average Year 7 cohort size of recruited sample =190.

These calculations assume a pre-post-test correlation of 0.75 (at the student-level). We expect the pre-post test correlation between the KS2 and PTM to be at least 0.5 and, hence, believe a correlation of 0.75 is achievable if we include additional covariates in the model alongside the KS2 score. We assume the school-level correlation to be 50% of the student-level (see, e.g., the KS1 to KS2 clustered correlations in Allen et al., 2018, Table 2, p.6)[14], although we consider this assumption to be conservative for a model with several covariates. The school-level ICC of 0.15 is based on our experience from previous studies (e.g., Evaluation of SMART Spaces Revision Programme[15]). The power calculations for the two subgroups, FSM and low prior attainment, are the same. In each case, we have assumed an average of 40 pupils in each school in the recruited sample.

Analysis

Primary outcome analysis

Using this weighted pupil-level sample, we will estimate the following linear regression model to address Research Question 2 (note that, as discussed in the study plan, we do not use the EEF's standard approach as it is unclear whether this is appropriate within the context of a matching framework):

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

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

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

- Individual KS2 prior attainment
- School average KS2 attainment of intake
- School low prior attainment proportion
- School high prior attainment proportion
- School cohort number of pupils
- Individual FSM
- School FSM proportion
- Individual EAL status
- School EAL proportion
- School academy status
- Individual IDACI
- School composition IDACI quintile

[14] Allen, R., Jerrim, J., Parameshwaran, M., & Thompson, D. (2018). *Properties of commercial tests in the EEF database*. Education Endowment Foundation.

[15] Hodgen, J., Bretscher, N., Hardman, M., Anders, J., & Lawson, H. (2023). *SMART Spaces: Spaced Learning Revision Programme*. Education Endowment Foundation.

- Ofsted grade
- Urban/rural classification

A core part of this project's aim is to explore the potential distributional changes (Research Question 1). We will carry this out using the following model, using the same weighted pupil-level sample as for the primary analysis model described above.

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

where, in addition to the elements already defined in the primary analysis model above, β_2 recovers the average difference in outcome performance associated with being in a mixed attainment school, rather than a comparable setting school, among those who are neither "Low Attainment" nor "High Attainment"; β_3 recovers the difference in outcome performance for higher attainers in setting schools, β_4 recovers the difference in outcome performance for low attainers in setting schools, β_5 recovers the difference in outcome performance among high attainers in mixed attainment schools compared to high attainers in setting schools, β_6 recovers the difference in outcome performance among low attainers in mixed attainment schools compared to low attainers in setting schools. As such, $\beta_5 - \beta_6$ recovers the difference in differences between high attainer and low attainer groups associated with being in a mixed attainment school, rather than a comparable setting school. The effect for high attainers is included because our previous work indicates that the 'attainment gap' between low and high attainers is more pronounced than that between low and 'average', or middle, attainers.[16]

Secondary outcome analysis

The secondary outcome analysis will use the same approach to modelling as in the main analysis to address Research Questions 3 and 4 comparing the effects on student self-confidence in mathematics. There are two separate self-confidence measures, general self-confidence and mathematics self-confidence. Each is a 7-item scale, with each item scored on a 5-point Likert scale and the scale score being the total score for the 7 items. The scales can be found in Appendix 5.

Subgroup analyses

We will conduct a sub-group analysis for FSM students. In line with EEF's guidance (EEF, 2022) for sub-group analysis for RCTs, we will first run separate sub-group analyses identical to the primary analysis model on the sample defined as falling in the FSM sub-group and also for the students with low prior attainment who also fall into the FSM group. As a robustness check, we will estimate a model on the full sample adding a covariate for our sub-group of interest and an interaction between this covariate and the treatment indicator. We will additionally examine the model dependence of the effect size estimate by presenting the range of estimates across different models.

Additional analyses

1. Mediation analysis

[16] Hodgen, J., Taylor, B., Francis, B., Craig, N., Bretscher, N., Tereshchenko, A., Connolly, P., & Mazenod, A. (2023). The achievement gap: The impact of between-class attainment grouping on pupil attainment and educational equity over time. *British Educational Research Journal*, *49*(2), 209-230. https://doi.org/10.1002/berj.3838

In order to investigate RQ5, we will conduct an exploratory analysis focused on two potential mediators, opportunity to learn and teacher quality as identified in the logic model outlined in the study plan. We have developed a bespoke survey instrument for completion by students to measure opportunity to learn as described in Appendix 6. In order to measure teacher quality, we will use two dimensions of the teacher quality student survey developed by Evidence-Based Education[17]: *understanding the content*, and *activating hard thinking*. See study plan for further details on the instrument.

To explore the role of opportunity to learn and teacher quality in mediating the overall changes in outcomes that we observe (Research Question 5), we will estimate mediation models (under assumptions of sequential ignorability and no interaction between treatment and mediator) using the approach proposed by Imai et al. (2010) to carry out appropriate inference.[18] This involves estimation of the following linked models:

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

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

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

where y is the outcome variable for pupil i in school j, mediated by OTL and $Quality$ measured in school j; α are model-specific intercept terms, $\boldsymbol{X_{ij}}$ is a vector of pupil- and school-level control variable characteristics (aiming to further reduce bias on top of that reduced through the matching approach and to increase the precision of our other estimates), and ε are individual-level error terms. From these models the estimate of β_7 recovers the average difference in opportunity to learn associated with being in a mixed attainment school rather than a comparable setting school, while β_8 recovers the same but for teacher quality; $\beta_7\gamma_1$ recovers the average difference in our outcome of interest mediated through opportunity to learn and $\beta_8\gamma_2$ recovers the same for the difference mediated through teacher quality ("mediated effects"), finally β_9 recovers the average difference that is not mediated by through opportunity to learn or teacher quality ("direct effects") (see Imai et al., 2010, pp.313-314). Underlying this approach is the assumption of independence between treatment assignment and the potential outcomes and mediators (OTL and $Quality$). Although treatment allocation is not at random, we consider this a reasonable assumption given Weidmann and Miratrix's (2021) evidence indicating the effectiveness of a matching approach.[19] We will use the R package mediation[20] to implement this approach and conduct a sensitivity analysis. See Appendix 4 for pseudo code to illustrate this proposed analysis. We will conduct a sub-group analysis for students with low prior attainment using an analogous modelling process.

2. Quantile analysis

As a robustness check on the analysis of distributional change, we will use quantile regression to explore distributional changes in performance associated with being in a school with setting compared to being in a mixed attainment school. We will specify models analogous to our

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

[18] Imai, K., Keele, L., & Tingley, D. (2010) A General Approach to Causal Mediation Analysis. Psychological Methods, 15, 4, pp. 309-334.

[19] Weidmann, B. & Miratrix, L., 2021. Lurking inferential monsters? Quantifying selection bias in evaluations of school programs. *Journal of Policy Analysis and Management, 40*(3), 964-986.

[20] Tingley D, Yamamoto T, Hirose K, Keele L, Imai K (2014). "mediation: R Package for Causal Mediation Analysis." *Journal of Statistical Software*, **59**(5), 1–38. http://www.jstatsoft.org/v59/i05/.

primary analysis for the following points of the outcome distribution: the 25th, 50th (median), and 75th percentile.

3. Sensitivity analysis

We will conduct a sensitivity analysis to examine whether schools administering outcome testing independently influence outcomes.

We will also carry out a replication of the primary analysis using a sample constructed from an additional pupil-level matching exercise on all data gathered from mixed attainment and setting schools, including pupils from the two mixed attainment schools for which matched setting schools were not able to be recruited. This will be carried out using the same matching variables considered for inclusion in the school-level matching exercise, along with their pupil-level counterparts. This approach is not feasible as our primary approach because pupil-level matching was not possible before the recruitment of schools, which required the matching exercise to have been completed; furthermore, it is appropriate to match at the level of the treatment variation (i.e. at the school level in this case). Nevertheless, we think this an appropriate robustness check.

We will explore the sensitivity of our primary analysis results using the spirit of the approach discussed by Rosenbaum:

> In a matched randomized experiment, each subject in a matched set has the same chance of being assigned to treatment or control because randomization has ensured that this is so. Without randomization, two people who look similar may differ in their chances of receiving treatment because they differ in terms of an unmeasured covariate not controlled by matching for measured covariates. The sensitivity analysis assumes that one subject in a matched set may be $\Gamma \geq 1$ times more likely than another to receive treatment because they differ in terms of unobserved covariates. If $\Gamma = 1$, then subjects who look the same are the same: matched subjects have equal chances of treatment, as in a randomized experiment. For $\Gamma = 1$, the sensitivity analysis reports a single answer, for instance a single p-value testing the null hypothesis of no treatment effect, and that single answer is the p-value that would be appropriate in a matched randomized experiment. For $\Gamma>1$, there is no longer a single p-value, but rather an interval of possible p-values. The sensitivity analysis asks: How large must Γ be before the interval is so long that it is inconclusive, perhaps both accepting and rejecting the null hypothesis of no effect at the 0.05 level? The interval of possible p-values would be inconclusive in this sense if it extended from below 0.05 to above 0.05. *Rosenbaum, P. (2015) Observational Studies 1, 1-17*

There is an R package developed by Rosenbaum for this purpose for the setting where there are matched sets with variable numbers of controls, which is the case in our analysis (sensitivitymv). Under the assumption of finding a difference that is considered statistically significant at a conventional level, this approach and tool can be used to estimate the value of Γ that would be required to increase the p-value of our findings such that it would no longer be considered significant at conventional levels.

However, there is a potential issue with use of this approach in this setting, which is that all available software (of which we are aware) that implement this approach does so under the assumption that data are independent, which is not the case for us as they will be clustered within schools. We posit that the approach continues to be valid if we inflate the p-value

obtained from this tool when Γ>1 to that found from our main analysis, and then re-use this inflation factor on the p-values obtained from the tool while we increase Γ.

Imbalance at baseline

Given the likely importance of inequality in grouping, it is particularly important to consider balance in measures of centrality, spread and skewness (1st, 2nd and 3rd order moments). We will test balances between groups on a fixed set of variables for means, medians, standard deviations, and skewness at both school- and student-level, and for the sub-group of pupils eligible for FSM. As in Table 2, we will report the comparison of means for the analysed sample with actual student-level data in a similar way to that required for a standard EEF RCT trial with standardised 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 (mixed attainment) and matched comparison (setted) groups to give an overall impression of the different distributions. As discussed above under "Recruitment and matching", w e will compare our treated sample with the following samples and discuss any differences, parrticulrly with all English schools in order to consider whether there are any factors that may be particularly associated with schools that adopt mixed attainment grouping:

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

Missing data

We will describe and summarise the extent of any missing data in the primary and secondary outcomes, and in the model associated with the analysis. Where possible, reasons for any missing data will also be described to enable a judgment of whether the data is missing at random.

If more than 10% of school-level or student-level data in the model is missing (based on the finalised matched sample), we will implement the following missing data strategy. The strategy will be followed separately for each instance of model and variable for which the threshold is exceeded. We will first explore whether there is evidence that the missing data is missing at random (MAR), since this is a pre-requisite for missing data imputation modelling to produce meaningful results. To do this we will create an indicator variable for each variable in the impact model specifying whether the data is missing or not. We will then use logistic regression to test whether this missing status can be predicted from the variables used for imbalance testing (listed above). Where predictability is confirmed we will proceed to use these same variables to estimate a Multiple Imputation (MI) model using a fully conditional specification, implemented using Stata MI to create 20 imputed data sets; we believe this is an appropriate number of imputed datasets given the anticipated level of potential missing data as a result of the administrative data source we are employing.[21] We

[21] Graham, J.W., Olchowski, A.E. & Gilreath, T.D. How Many Imputations are Really Needed? Some Practical Clarifications of Multiple Imputation Theory. Prev Sci 8, 206–213 (2007). https://doi.org/10.1007/s11121-007-0070-9

will re-estimate the treatment effect using each dataset and take the average and estimate standard error using Rubin's combination rules.[22]

Analysis using the dataset produced through either of these missing data strategies will be used as a sensitivity analysis i.e. we will base confirmation of the effectiveness of the treatment on complete case analysis only but assess the sensitivity of the estimate to missingness using the estimates from the multiply imputed dataset. If the complete case analysis model implies effectiveness but the imputed dataset analysis model does not (or changes the direction of the estimated effect) we must assume that the missing data is missing not at random to such an extent as to invalidate our conclusion of effectiveness, which we would state in the reporting of the evaluation.

Compliance

We will explore compliance using a a CACE/instrumental variables analysis.

Compliance will be defined according to whether the schools meet the eligibility criteria throughout the period of research, i.e., during the academic years 2022/23 and 2023/24 We are aware that a small number of schools changed their grouping practices prior to the beginning of the academic year, 2022/23. Hence, compliance will be judged on a school's actual practices, and will not be dependent on whether a school was recruited as a mixed attainment or a setted school. These are:

- *For mixed attainment schools*, they teach mathematics to Year 7 and Year 8 students in mixed attainment classes. Mixed attainment classes are defined as those in which the range of attainment in each class broadly reflects the full range of attainment in the year group for that subject. Mixed attainment schools may additionally have, or introduce, a 'nurture group', in which the very lowest attaining students are taught separately. Hence, we propose to assess compliance by examining whether mixed attainment schools have, in effect, one or more high attaining 'sets
- ', where the average attainment of the class is above the 75th percentile for the school as a whole.
- *For setting schools*, they teach mathematics to Year 7 and Year 8 students in three or more attainment sets. Attainment sets are defined as classes in which students are grouped by their attainment in a subject and taught together for that subject. Since this study focuses on the effects of grouping by subject attainment compared to mixed attainment grouping, schools that introduce streaming will be considered non-compliant. Streaming is defined as the allocation of students to groups for teaching in some or all subjects, based on a notion of general ability.

School compliance will be assessed on the basis of a simple question in the Head of Maths surveys at the end of each academic year.

Intra-cluster correlations (ICCs)

In order to estimate the intra-cluster correlation (ICC) of the outcome measures at school-level we will employ an empty variance components model, as follows:

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

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

where $\boldsymbol{Y_{ij}}$ is the relevant outcome for pupil $\boldsymbol{i}$ who is in school $\boldsymbol{j}$, $\boldsymbol{\eta_j}$ is a school-level random effect, and $\boldsymbol{\varepsilon_{ij}}$ is a pupil-level error term. The school-level random effect is assumed to be normally distributed and uncorrelated with the pupil-level errors.

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

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

Effect size calculation

Hedges' g effect size will be calculated as follows:

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

where our conditional estimate of $\boldsymbol{\overline{x_1} - \overline{x_2}}$ is recovered from $\boldsymbol{\beta_1}$in the primary ITT analysis model;

$\widehat{s*}$ is estimated from the analysis sample as follows:

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

where n_1 is the sample size in the control group, n_2 is the sample size in the treatment group, s_1 is the standard deviation of the control group, and s_2 is the standard deviation of the treatment group (all estimates of standard deviation used are unconditional, in line with the EEF's analysis guidance to maximise comparability with other trials);

and j and λ are calculated as follows:

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

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

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

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

Confidence intervals for the effect size will be calculated by substituting the upper and lower bound confidence intervals from the estimate of $\boldsymbol{\beta_1}$ into the above equation (other aspects remaining fixed) and using these as upper and lower confidence intervals for the effect size.

Appendix 1: Matching for Student Grouping Study

This document details the matching process for the Student Grouping Study. First, we import the full list of schools, where NPD KS4 data are available in 2019, that we have constructed and merged together elsewhere, including contextual information about schools and school-level averages about the pupils in these schools. The dataset also includes an identification of whether the school is one of the recruited mixed attainment schools; other known mixed attainment schools (notably those in earlier phases of this project) have been removed from the sample to prevent them being identified as potential comparators.

These are the variables that must be present for schools to be included in our sample:

- URN
- Mixed attainment identifier
- Academy status
- Number of pupils in KS4 cohort in 2018
- Average KS2 performance of 2016, 2017 and 2018 GCSE cohorts (observations are carried backwards where necessary)
- Proportion of 2018 GCSE cohort identified as low, average and high prior attainers
- Proportion of FSM in school
- Proportion of EAL in school
- Single sex school
- Most recent Ofsted judgement
- Region
- Urban/rural identifier
- IDACI quintile

Matching is carried out using the MatchIt software. The propensity score is estimated using the following logistic regression model fit using school-level data:

$$MixedAttain = FSMProp + KS2_{2019} + KS2_{2018} + KS2_{2017} + PriorLo + PriorHi + PupilNo. + IDACI' + Ofsted' + \varepsilon$$

Appendix 2: Balance

Standardised differences, calculated using the full sample standard deviations, are presented in Tables 4 and 5 for:

- the full analysis sample (unmatched),
- the sample restricted to mixed attainment and matched comparator schools (all matched),
- the sample further restricted to recruited matched comparators schools and mixed attainment schools for whom at least one appropriate match has been recruited (recruited matched), and
- the recruited sample restricted to matched comparator schools and mixed attainment schools from which primary data have been collected, excluding schools in strata in which primary data have not been successfully obtained (primary data matched).

Table 4: Student-Weighted Imbalance (Note: this is an expanded version of Table 2)

	Unmatched	All Matched	Recruited Matched	Primary Data Matched
KS2 2019	0.051	-0.025	-0.018	-0.023
KS2 2018	0.051	-0.025	-0.018	-0.023
KS2 2017	0.037	-0.029	-0.013	-0.028
Low Prior Attain Prop.	-0.096	0.022	0.001	0.006
High Prior Attain Prop.	0.000	-0.084	-0.109	-0.132
Average (Attainment)	***0.009***	***-0.028***	***-0.031***	***-0.040***
Number of pupils on school roll	0.235	-0.129	-0.222	-0.270
FSM Prop.	-0.113	0.150	0.199	0.198
EAL Prop.	0.443	0.629	0.352	0.383
Academy status	-0.230	-0.298	-0.379	-0.326
IDACI1	0.053	-0.118	-0.240	-0.165
IDACI2	-0.046	-0.057	0.018	-0.104
IDACI4	-0.005	0.018	0.140	0.156
IDACI5	-0.123	0.075	-0.022	-0.009
Ofsted Outstanding	0.174	0.024	-0.220	-0.266
Ofsted Good	0.052	0.016	0.075	0.061
Urban	-0.192	-0.105	-0.227	-0.214
Average	***0.119***	***0.113***	***0.141***	***0.148***

Table 5: Imbalance at school-level, weighted for number of comparator schools, but not for number of students in each school

	Unmatched	All Matched	Recruited Matched	Primary Data Matched
KS2 2019	0.409	-0.020	-0.026	-0.032
KS2 2018	0.394	-0.021	-0.012	-0.030
KS2 2017	0.390	-0.025	-0.046	-0.057
Low Prior Attain Prop.	-0.485	0.013	0.003	0.009
High Prior Attain Prop.	0.309	-0.057	-0.096	-0.123
Average (attainment)	***0.203***	***-0.022***	***-0.035***	***-0.047***
Number of pupils on school roll	0.565	-0.069	-0.191	-0.249
FSM Prop.	-0.468	0.054	0.134	0.129
EAL Prop.	0.453	0.595	0.354	0.398
Academy status	0.011	-0.263	-0.288	-0.239
IDACI1	0.154	-0.085	-0.223	-0.174
IDACI2	0.069	-0.027	0.041	-0.051
IDACI4	0.064	0.000	0.102	0.129
IDACI5	-0.107	0.027	-0.072	-0.065
Ofsted Outstanding	0.181	-0.055	-0.290	-0.330
Ofsted Good	-0.037	0.050	0.079	0.072
Urban	-0.065	-0.046	-0.152	-0.153
Average	***0.260***	***0.088***	***0.132***	***0.140***

Appendix 3: Code for power calculations

```
# Power calculations for Student Grouping Study using PowerUpR
# Calculated for SAP: 30/04/2023

library(PowerUpR)

# Recruited matched sample as per SAP
# 70 setted schools and 29 mixed attainment schools
# in 29 groups i.e. 3.4 (setted) schools per block (each mixed attainment school)

# Calculations assume 3-level MLM with treatment at Level 2
# Average pupil number per school = 190 based on actual recruitment
# Assumptions: ICC (schools) = 0.15; pre/post-test correlations 0.75 at student-level & 0.38 at school-level
# Standard assumptions for power and alpha levels

mdes.rec <- mdes.bcra3f2(power=.80, alpha=.05, two.tailed=TRUE, rho2=0.15, p=.25, g2=1, r21=0.56, r22=0.14, n=190, J=3.41, K=29)

# FSM power calculations on the same basis and assuming average 40 FSM students per school

mdes.recFSM <- mdes.bcra3f2(power=.80, alpha=.05, two.tailed=TRUE, rho2=0.15, p=.25, g2=1, r21=0.56, r22=0.14, n=40, J=3.41, K=29)
```

Appendix 4: Pseudo-Code for mediation analysis

```
library(mediation)

model.otl <- lm(otl ~ mixedAttain + Xs, data = data)
model.quality <- lm(quality ~ mixedAttain + Xs, data = data)
model.y <- lm(y ~ mixedAttain + otl + quality + Xs, data = data)
mediate.otl <- mediate(model.otl, model.y, sims = 1000, treat = "mixedAttain", mediator =
"otl")
mediate.quality <- mediate(model.quality, model.y, sims = 1000, treat = "mixedAttain",
mediator = "quality")
```

Appendix 5: Self-confidence scales

Self-Confidence in Mathematics

Items marked with an asterisk (*) are reverse-scored.

- Work in maths is easy for me
- I am not very good at maths*
- Maths is one of my best subjects
- I hate maths*
- I do well at maths
- I get good marks in maths
- I learn things quickly in maths lessons

General Self-Confidence in Learning

- I learn quickly
- Most things I do, I do well
- I am proud of my achievements at school
- I can do things as well as most people
- If I really try I can do almost anything I want to
- I am confident in my abilities
  I am generally high achieving in my studies

Both scales are scored on a 5-point Likert scale

Appendix 6: Opportunity to Learn Instrument

Development and Validation of the Opportunity to Learn (OTL) Instrument

Our aim was to develop and validate an instrument to be completed by Y8 students to assess OTL operationalised in terms of learning time in class on the range of topics representing the mathematics expected to be covered in Y8 lessons.[23] In addition, we wanted an instrument that could be completed relatively quickly by students (thus minimising any additional burdens to schools, and avoiding dangers of test fatigue for students) and that could be delivered online (using a multiple-choice format similar to that used in the PISA OTL survey). Ideally, we wanted an instrument producing a unidimensional OTL scale consisting of between 20 and 25 items.

There were three stages to the validation process: validation interviews with Y8 students and mathematics education experts, followed by two 'pilot' rounds of validation.

The initial survey consisted of 34 items with a Likert scale. The survey was designed (and administered) using Research Electronic Data Capture (REDCap) tools hosted at UCL.[24]2

Cognitive interviews were conducted using 'think aloud' protocols with 8 students.[25] These enabled us to assess whether students understood the format of the items and the instruction to consider time spent on the topic (not to solve the item) as well as whether the items adequately represented the intended topics. Additionally, three mathematics education experts were interviewed to assess the mathematics coverage and whether the items would be easily understood by all Y8 students. As a result of the Stage 1 interviews, 10 items were deleted and 12 items added, resulting in a total of 36 items.

In the initial pilot survey (Autumn 2020), 187 Y8 students from one school completed the online survey. Surveys were completed on-site during mathematics lessons in the school's ICT room with the students' mathematics teacher present for the lesson. On average, students took 17 minutes to complete the survey (S.D. = 3 minutes). 17 students did not complete the whole survey. This first pilot indicated good internal consistency (Cronbach α = .86). Rasch[26] statistics were generally good, although factor analysis suggested the possibility of two factors. The Rasch modelling identified a number of potentially problematic items. As a result, adjustments were made to the wording of eight items, but no items were dropped at this stage.

In the second pilot survey (Winter 2021), 295 Y8 students from 4 schools completed the survey, again administered using REDCap. However, due to the pandemic and school closures, the surveys were completed by students at home. 44 students did not complete the whole survey. Additionally, 31 students completed the survey in less than 5 minutes (which,

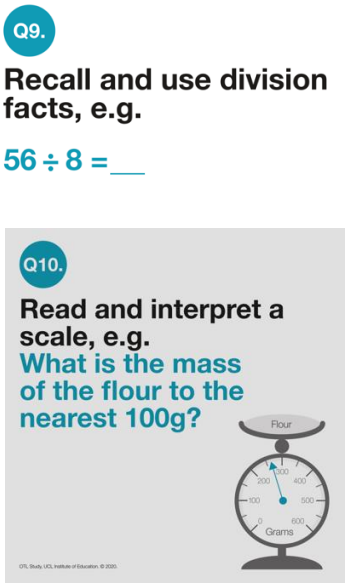
[23] Our expectation is that some classes, particularly lower sets, may cover some aspects of the KS2 curriculum, so the topics covered were slightly broader than the Y8/KS3 national curriculum for mathematics.
[24] REDCap (Research Electronic Data Capture) is a secure, web-based software platform designed to support data capture for research studies, providing 1) an intuitive interface for validated data capture; 2) audit trails for tracking data manipulation and export procedures; 3) automated export procedures for seamless data downloads to common statistical packages; and 4) procedures for data integration and interoperability with external sources (Harris et al., 2009, 2019).
[25] The original intention had been to interview 25 students from 3 schools. However, school closures in Spring 2020, meant that it was only possible to interview students from 1 school.
[26] Rasch modelling is a form of Item Response Theory (IRT) and is commonly used to validate single trait (or uni-dimensional) instruments and tests.

on the basis of the first pilot timings, was judged to be too fast to have considered OTL for all items) and 5 students took longer than 20 minutes (suggesting that they attempted to solve the items rather than consider OTL.) Hence, responses from 215 students were analysed.

Analysis consisted of factor analysis and Rasch modelling (using the partial credit model).[27] As a result of this analysis, several items were dropped (including problematic items identified in the first pilot) leaving a final survey of 22 items. Rasch statistics for these items are attached in Appendix 1 and are considered satisfactory (e.g., infit and outfit values of all items fall into the acceptable range between 0.4 – 1.6[28]). Internal consistency is good (Cronbach α = 0.85). However, factor analysis indicated two factors 'explaining' 25% and 10% of the variance, respectively. Whilst not ideal, we judge this to be satisfactory for the purposes of the analyses that will be conducted using this measure.

As a final validation, the instrument will be validated early in 2022 with a sample of at least 300 Y8 students from 3-4 schools under the conditions in which the instrument will be delivered (i.e., in students mathematics classes) and with the 22 items presented in random order. This has not been possible due to the impact of the pandemic on schools.
I

Example Items from the OTL Survey



How often have you encountered these types of problems in your mathematics lessons?

- ◯ Frequently
- ◯ Sometimes
- ◯ Rarely
- ◯ Never

[27] The partial credit model is a rating scale model and, thus, appropriate to modelling a set of Likert scale items. PCM allows each item to have its own structure.

[28] Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch Model: Fundamental measurement in the human sciences* (2nd ed.). Lawrence Erlbaum.