Academic Journal of Psychology and Education (AJPE)
Volume.14, Number 1; January-2023;
ISSN: 2836-3760 | Impact Factor: 7.87
https://zapjournals.com/Journals/index.php/ajpe/index
Published By: Zendo Academic Publishing

THE ACHIEVEMENT GAP: THE IMPACT OF BETWEEN-CLASS ATTAINMENT GROUPING ON PUPIL ATTAINMENT AND EDUCATIONAL EQUITY OVER TIME

[1]Jeremy Hodgen

Article Info

Keywords: between-class attainment grouping, setting, academic outcomes, lower secondary pupils, English, mathematics, prior attainment, attainment gap, equity, social justice, teaching practice, England.

Abstract

This study examines the impact of between-class attainment grouping, also known as setting, on the academic outcomes of lower secondary pupils in England. The research was conducted with 2944 12-13 year-olds from 76 schools across the country, who were allocated to attainment groups in English and mathematics over the first two years of secondary schooling. The study found that pupils who had been placed in the top set demonstrated significantly better performance than those in the middle and bottom sets, in both English and mathematics, after accounting for prior attainment. However, even after adjusting for prior attainment, the research indicates a noticeable gap in performance, particularly in English. The study shows that between-class grouping results in the widening of the attainment gap, specifically for pupils placed in low sets. The research raises concerns related to equity in education and social justice in teaching practice. Its findings also provide contemporary evidence on the effects of between-class grouping in England, which was lacking in the prior literature.

INTRODUCTION

Few topics in education have generated such controversy or longstanding study as grouping by ‘ability’ (‘tracking’).[1] And in spite of what Steenbergen-Hu et al. (2016) characterise as a century of research on this topic, the impact of grouping by prior attainment—and especially, of different *methods* of grouping—remains contested. This can to some extent be explained by the problematic nature of the existing research literature. There is a scarcity of contemporary work focused on pupil-level outcomes; and different types of grouping are often conflated within the meta-analyses and syntheses that have predominated in the field (Francis et al., 2020), making it difficult to draw clear conclusions. Many policymakers and practitioners believe attainment grouping to be effective (see Francis et al., 2017), and practices of ability grouping (tracking)—including between-school grouping, between-class grouping (e.g. setting) and within-class grouping—are prevalent in many systems internationally (Jerrim, 2019; OECD, 2016). In England, grouping by attainment is widespread in both primary and secondary schools, with 37% of 6–7 year olds placed in attainment groups for either literacy or numeracy (Hallam & Parsons, 2012), and more than 70% of secondary schools placing 11 year olds in attainment groups for mathematics (Taylor et al., 2020).

[1] IOE – Faculty of Education and Society, University College London, London, UK

This paper provides timely new empirical evidence on the impact of the contentious practice of grouping by attainment on pupils' attainment outcomes. Specifically, we use data from an experimental study of well-defined practices to examine the different effects on those placed in top and bottom sets. We provide up-to-date evidence about the impact of setting in the United Kingdom. This is important because very few large-scale studies have been carried out in the United Kingdom and because the practice of setting is very different to that of the United States, where the bulk of the research has been conducted. We highlight issues raised both for social justice in teaching practice and for future research.

Contrasting theories about attainment grouping

Proponents of attainment grouping argue that placing pupils in more homogenous classes enables teachers to better tailor the curriculum and pedagogy to those pupils, and, hence, is more efficient and effective for all pupils (e.g. Hallinan, 1994; Rosenbaum, 1999). Many school leaders believe that within-school grouping has benefits for all pupils, including those with low prior attainment. Indeed, a study conducted in England by Macleod et al. (2015) found that more than a third of schools surveyed had 'introduced or improved' setting as a way of raising attainment for disadvantaged pupils.

On the other hand, critics of grouping by attainment point to analyses showing the practice may not be as efficient as hypothesised by its proponents, and may in fact increase educational inequity. For example, Hanushek and Wößmann's (2006) analysis of international comparative tests in mathematics, reading and science suggests that educational systems that adopt early grouping by attainment tend to have a widening gap in attainment over time, thus increasing educational inequity, and may additionally be associated with an overall decrease in mean attainment in comparison to other systems. Similarly, evidence from PISA 2012 suggests a relationship between grouping by attainment within schools and the share of low and top performers in an education system, concluding from their findings that 'more ability grouping within schools is related to a greater number of low performers in mathematics, and fewer top performers' (OECD, 2016, p. 186). Evidence from many observational studies in the United States, Germany and the Netherlands, as well as the United Kingdom, suggests that ability grouping is associated with increased inequity on educational outcomes (e.g. Berends & Donaldson, 2016; Borghans et al., 2020; Capsada-Munsech & Boliver, 2019; Gamoran & Mare, 1989; Matthewes, 2021), although there is some dispute as to whether this widening gap reflects a benefit for those pupils placed in high sets, a disbenefit for those placed in low sets, or both (e.g. Betts & Shkolnik, 2000).

Longstanding research demonstrates that pupils from low socio-economic groups, and from certain minority ethnic (typically Black) backgrounds, are disproportionately likely to be found in low-attainment tracks and groups, whereas White pupils from affluent families are over-represented in high-attainment groups and 'academic' tracks (e.g. Bosworth, 2013; Moller & Stearns, 2012; Muijs & Dunne, 2010; Strand, 2012). Recent research in England bears this out (see Archer et al., 2018; Connolly et al., 2019). Research has also shown that pupils from socially disadvantaged backgrounds are disproportionately misallocated to low-attainment groups (Dunne et al., 2011; Jackson, 1968), compounding existing inequalities at the start of schooling (e.g. Waldfogel & Washbrook, 2010). This over-representation of pupils from disadvantaged backgrounds in low sets (and more affluent pupils in high sets), coupled with ongoing attainment gaps and the hypothesis emerging in numerous studies that attainment grouping practices cause inequitable educational progress, has led attainment grouping to be frequently seen as a matter for social (in)justice in education.

Within this body of work, it is argued that the inequitable outcomes for different attainment groups may be due to several factors such as differences in teacher expectations, teacher quality, curriculum content and opportunity to learn, as well as pupils' self-confidence and motivation (e.g. Francis et al., 2017; Oakes, 1995). Research does indicate that these arguments are to some extent justified. Teachers' expectations do appear to be lower for those pupils placed in lower-attaining groups (Campbell, 2014, 2017; Ireson & Hallam, 2009; Timmermans et al., 2015). There is evidence that lower sets tend to be allocated teachers with less subject-specific expertise or less experience (Francis et al., 2019; Kelly, 2004; Papay & Kraft, 2015). Lower-attaining groups do appear to be taught a reduced curriculum offer (Hallam & Ireson, 2005; Jaremus et al., 2020; Wilkinson et al., 2020), offered fewer opportunities for participation and discussion (Gamoran et al., 1995) or conceptual understanding (Martinková et al., 2020), and have restricted opportunities to progress (Buttaro & Catsambis, 2019), while studies have also found a relationship between pupil self-confidence and attainment grouping (Francis, Craig et al., 2020; Houtte et al., 2012; Ireson & Hallam, 2009; Muijs & Dunne, 2010).

Between-class attainment grouping (‘setting’) and pupil achievement

Given these contrasting theories, attainment grouping practices remain a strong point of interest and contestation within educational practice and research. In this paper, we focus on *setting*, a particular form of between-class grouping, which is prevalent in English secondary schools (Taylor et al., 2020). Setting is where pupils are grouped by subject attainment for teaching in that subject, and is to be distinguished in England from *streaming*, where pupils are grouped by general ability for teaching across a majority of subjects (Ireson & Hallam, 2001). There are also many other different forms of grouping by attainment described in the literature, including between-school grouping (tracking), within-class grouping and acceleration for high-attaining pupils (see Francis, Taylor et al., 2020 for elaboration).

Our focus on between-class attainment grouping, or setting, is for two reasons. Research syntheses suggest that the various different forms of attainment grouping may have statistically significant different sizes, and even directions, of overall effect (Higgins et al., 2018; Steenbergen-Hu et al., 2016), and that particular grouping practices may impact on different groups of pupils in different ways (Rui, 2009). In addition, between-class grouping is widely used in educational systems internationally (Jerrim, 2019).

The impact of attainment grouping on pupils has been the subject of extensive research, and a large number of literature reviews and meta-analyses synthesise the findings on the topic. These syntheses suggest that between-class attainment grouping has no overall benefit to academic attainment, with a small negative impact for low-attaining pupils and a small positive benefit for high-attaining pupils (Higgins et al., 2018; Rui, 2009; Slavin, 1990). On closer examination, the evidence provided by this extensive evidence base is not as robust or as generalisable as this research base would suggest.

Research specifically examining between-class grouping at secondary level in the United Kingdom is mostly from small-scale studies (Boaler, 1997; Ireson & Hallam, 2001; Wiliam & Bartholomew, 2004). Notable exceptions are studies by Kerckhoff and by Ireson. Kerckhoff (1986) drew on British birth cohort data to analyse the impact of within-school attainment grouping on the achievement of pupils who attended secondary schools in the 1970s, at a point when the educational system was very different to today and particularly so for low-attaining pupils (Hodgen et al., 2022). The findings indicated a widening attainment gap for schools that used attainment grouping compared to those that did not. In the only other large-scale study carried out in the United Kingdom, Ireson and colleagues examined the effects of between-class grouping in comparison to mixed attainment on the achievement of a cohort of pupils from 45 schools who took attainment tests at age 14 in 2000 (Ireson et al., 2002, 2005) and national GCSE examinations at age 16 in 2002 (Ireson et al., 2005), focusing on three subjects: English, mathematics and science. However, their results were mixed and inconclusive. For example, at age 16, they found no effect for setting, although pupils of equivalent prior attainment performed better in all three subjects when placed in higher sets. Ireson et al.’s (2005) study reports data that are now more than 20 years old and, aside from the need to replicate individual studies (Makel & Plucker, 2014), there is a need to provide up-to-date evidence about the effects of setting.

There are a large number of meta-analyses examining international evidence on the topic (e.g. Kulik & Kulik, 1992; Lou et al., 1996; Slavin, 1990). However, in synthesising different sets of studies, these meta-analyses report effects of attainment grouping that vary from $d = -0.45$ (Slavin, 1987) to $d = 0.19$ (Kulik & Kulik, 1984). In an attempt to produce a definitive answer on the issue, Steenbergen-Hu et al. (2016) conducted a secondary meta-analysis in order to review and synthesise the large number of primary meta-analyses on the topic of attainment grouping. They identified no fewer than 11 primary meta-analyses that examined the effects of between-class grouping by attainment and found no statistically significant effect for the practice, either overall or for pupils of high, middle or low attainment. However, these 11 primary meta-analyses were all based on dated original studies; the most recent being published in 1991 and most carried out in the 1960s and 1970s, at a time when statistical methods were much less sophisticated than those currently available, and did not take account of clustering of pupils within classes through approaches such as multilevel modelling (Connolly et al., 2017; Hedges, 2007). Moreover, this was a period when the reporting requirements for experimental studies were relatively weak, since this predated initiatives to pre-register trials and experiments (Styles & Torgerson, 2018). It is likely that, for many studies, attainment grouping was combined with guidance on practice, professional development and/or curriculum adaptation to match different attainment levels. Indeed, it may be that the structural effects of between-class grouping are mediated through teaching quality and opportunity to learn. But the contribution of these elements was not considered in any of the primary meta-analyses through now standard techniques such as moderator analysis or meta-regression. This may be because few of the original studies provide any details on these aspects. Educational practices (and even teaching qualifications) were also very different from the present day. In England, for example, compulsory schooling ended at age 15 until 1972 and many pupils left education without formal qualifications (Gillard, 2018).

The vast majority of these original studies were carried out in the United States and, indeed, the debate around grouping by attainment has largely been framed in terms of the US practices of ‘tracking’ versus ‘detracking’ (see e.g. Loveless, 1999), practices that have been treated as synonymous with the practices of setting versus mixed ‘ability’ teaching in England (Abraham, 2008; Wilkinson & Penney, 2014). In fact, as Domina et al. (2019) show, tracking involves a range of sorting practices that reflect particularities of the American educational system, and tracking, as practised in the United States, is often closer to ‘streaming’ rather than ‘setting’ (Wilkinson & Penney, 2014). Hence, the findings of US studies may not generalise to different educational systems and contexts such as England.

Steenbergen-Hu et al. (2016) also conducted a primary meta-analysis that only included those they selected as the ‘highest quality’ original studies, randomised controlled trials (RCTs) where the full text was available. In contrast to the secondary meta-analysis, this found a positive effect for between-class grouping ($g = 0.15$, 95% CI: 0.01–0.29). However, this result was based on just five dated studies, published between 1962 and 1974, all of which were conducted in the United States. In addition, all five were small-scale interventions, with four of the five studies each conducted in just one school and the fifth in just four schools, and none of the studies used methods that took account of the clustering of pupils within classes (or schools).

The most recent primary meta-analysis (Rui, 2009), which is not included amongst those synthesised by Steenbergen-Hu et al., found attainment grouping had a negative impact on low-attaining pupils, but no effect on middle or high-attainment pupils. Rui's meta-analysis synthesised the results of just 15 studies, all conducted in the United States. Unfortunately, Rui's analysis aggregates the results of both experimental and observational studies, including just four RCTs published between 1972 and 1996. Furthermore, although Higgins et al.’s (2018) secondary analysis suggests that the effects of between-class and within-class grouping are in different directions, Rui does not distinguish between these two forms of grouping, thus conflating their effects.

None of the above take into account additional factors, such as curriculum and quality of teaching, that might influence the impact of attainment grouping. Only very recently have researchers started to carry out quantitative studies of teaching in relation to attainment grouping and pupil outcomes. Magableh and Abdullah (2021) conducted a small-scale experimental study of differentiated instruction in mixed-attainment classes, finding that differentiated teaching resulted in higher outcomes for pupils. Wang et al. (2021) explored the impact of teacher support on outcomes for pupils tracked into three different school bands in Hong Kong. They found that teacher support mediated the higher English and mathematics attainment of pupils in high-band schools and also moderated the English attainment of pupils in low-band schools. However, the context of these two studies is different, focusing on within-class differentiation and between-school tracking, respectively. Furthermore, 'teacher support' differs from 'teaching quality' and perhaps is more analogous to a supportive climate, or high expectations. No quantitative studies yet focus explicitly on the quality of teaching in schools using between-class grouping.

In summary, the limitations highlighted above indicate a need for robust, contemporary studies of specific between-class grouping practices and their outcomes that establish or contest the somewhat fragile conclusions described above. Especially, there is a need to provide up-to-date evidence and to investigate the effect of between-class grouping, or setting, as it is practised in systems like England. Our analysis seeks to do this, exploring the relative attainment outcomes of pupils placed in different attainment sets (between-class attainment groups) over the first 2 years of secondary school in the core subjects of English and mathematics, using a large, robust and representative sample of schools in England.

METHOD

The data discussed in this paper draw on data from a large-scale mixed-methods project 'Best Practice in Setting', funded by the Education Endowment Foundation. Specifically, it analyses data collected during a cluster RCT of the 'Best Practice in Setting' intervention. As already noted, 'setting' is an especially prevalent form of between-class attainment grouping in England (Taylor et al., 2020), comprising tracking by subject. In principle, a pupil might be placed in a high set for several curriculum subjects and in low sets for others, depending on their respective prior attainment in disparate subjects. In practice, setting is sometimes mixed with, or layered upon, other tracking practices such as streaming (see Francis, Taylor et al., 2020 for a discussion). The project sought to address prior gaps in the literature, by exploring: whether practice in setting that remediates some of the problematic practices identified in the literature as affecting those in low groups might improve young people's progress; what comprises good practice in mixed attainment pedagogy; and the experiences and outcomes of pupils subject to attainment and mixed-attainment grouping. It consisted of a 2-year intervention comprising guidance for schools on how to group pupils and allocate teachers to classes, and professional development focusing on high expectations for all pupils and flexible conceptions of 'ability' (Roy et al., 2018). The intervention was tested by a fully powered RCT examining the impact or otherwise of practice in setting pupils for English and mathematics in Year 7 and Year 8 based on research evidence. In addition, there were surveys of 13,462 pupils and 597 teachers, and individual and focus group interviews with 246 pupils and 54 teachers, although results from these data are not reported in this paper.

The intervention and research were undertaken in 126 secondary schools in England (divided into intervention or control groups), and involved instigating work with and monitoring pupil cohorts from the beginning of Year 7 (11–12 years old) to the end of Year 8 (12–13 years old), the first 2 years of English secondary schooling. The study focused on their experiences and outcomes in English and mathematics, which were selected as the foci because: (a) they are two subjects given longstanding priority in the national curriculum and within-school performance indicators; and (b) they represent diversity in content and pedagogy.

The trial was conducted by an independent evaluation team who were responsible for the trial design, school recruitment, randomisation, pre-specification and registration of the trial (Roy & Styles, 2017), as well as the administration and marking of the primary outcome attainment tests. The intervention was developed and delivered by the programme delivery team, including the authors of this paper, who were also involved in the school recruitment, quantitative and qualitative data collection, and supporting relationships with schools throughout the trial. The study was approved by the Research Ethics Committee of King's College London and Queen's University Belfast and, later, UCL Institute of Education.[2]

This paper analyses the differential impact of setting on attainment outcomes for pupils placed in different set levels across the 2 years of the intervention. To be clear, where the RCT compared attainment outcomes between the intervention and schools maintaining 'business as usual' setting, this paper explores *the impact of setting* per se on the outcomes of pupils in different attainment groups.

The sample

To be eligible for the trial, schools had to use subject-based between-class grouping by attainment (not streaming) and to have at least three set levels for each subject (top, middle and bottom). Schools were recruited to the 'Best Practice in Setting' trial through a mixture of volunteer and direct 'cold call' approach sampling, then randomised to the intervention and control groups of the RCT. Volunteer-sampled schools were recruited through a traditional and social media campaign by the authors. Direct approach-sampled schools were identified through a stratified random sample then approached by the independent evaluation team (see Roy et al., 2018). Of these 126 schools, 121 took part in the mathematics trial and 79 took part in the English trial. However, there was considerable dropout of participant schools during the duration of the 2-year trial, and a significant portion of schools did not deliver the final outcome tests in English and mathematics. Hence, the achieved sample consisted of 73 schools in mathematics and 35 schools in English. Since some schools took part in both subjects for the trial, there was a total of 76 schools in the achieved sample.

The overall characteristics of the pupils and schools in the mathematics and English samples are summarised in Table 1. Demographic data in relation to gender, household background, free school meals entitlement, ethnicity and set allocation are provided for the 2236 pupils in the mathematics trial and 919 pupils in the English trial.[3] The samples are reasonably reflective of the national population. In particular, it can be seen that the sample is well balanced in terms of gender and also broadly representative of the national population in relation to ethnicity [where it is reported that, nationally, 76% are White, 10% Asian, 6% Black and 5% mixed; see DfE (2015, p. 15)]. The sample is also broadly representative in relation to the proportion of disadvantaged pupils, with 30.6% of the present sample having been eligible for free school meals (FSM) at some point, compared to the nationally reported figure of 32% (DfE, 2015, p. 14).[4]

It is also noteworthy that there is a large amount of missing data on ethnicity and household socio-economic status (SES). This is because a large proportion of pupils chose not to provide these data: in the mathematics trial, 35% and 42% did not provide data on ethnicity or SES, respectively, and, in the English trial, 38% and 45% did not provide data on ethnicity or SES, respectively. The two samples in each subject, with and without attrition, were broadly similar. See Table S1 for further details. We address the issue of missing data further in the analysis section, below.

The overall characteristics of the schools are broadly reflective of the national population of state-funded, non-selective schools (see Connolly et al., 2019). The proportions of OFSTED grades across schools are generally representative of the national picture (in 2015) of 22% outstanding, 56% good, 17% requires improvement and 5% inadequate (OFSTED, 2016, p. 133), although we note that the English sample is slightly skewed towards poorer performing schools.

TABLE 1 Sample characteristics

	Mathematics		**English**	
Schools	*N*	**%**	*N*	**%**
OFSTED grade				
Outstanding	18	25	5	14
Good	39	53	19	54
Requires improvement	15	21	10	29
Inadequate	1	1	1	3
Proportion eligible for free school meals (FSM)	Mean	(SD)	Mean	(SD)
	28.3%	(16.1%)	27.5%	(13.3%)
Total schools	**73**		**35**	

Students	*N*	**%**	*N*	**%**
Gender				
Boy	1168	52.7	497	54.3
Girl	1048	47.3	419	45.7
Missing	20		3	
Household socio-economic background (SES)				
Higher	614	47.6	221	43.9
Intermediate	473	36.6	198	39.4
Lower	204	15.8	84	16.7
Missing	945		416	
Ever eligible for free school meals (FSM)				
No	1533	68.7	649	70.9
Yes	699	31.3	267	29.1
Missing	4		3	
Ethnicity				
White	1115	76.7	481	84.4
Black	115	7.9	31	5.4
Asian	104	7.2	26	4.6
Other	119	8.2	32	5.6
Missing	783		349	

Set allocation				
Top	733	34.9	282	33.5
Middle	1073	51.1	427	50.8
Bottom	293	14.0	132	15.7
Missing	137		78	
Total students	**2236**		**919**	

This sample of schools was recruited for the purpose of the trial and, as such, had expressed some interest in adopting 'best practice' in attainment grouping. Hence, whilst not fully representative, the sample may be considered a 'telling case' in that these schools might be expected, if anything, to be more interested than other schools in increasing equity across attainment groups (Mitchell, 1984).

For the purposes of this analysis and for comparability with previous analyses (Connolly et al., 2019; Francis, Craig et al., 2020), we have combined the intervention and control group schools, which is justified because no significant effect was found for the intervention for either subject (see Roy et al., 2018).

Instruments

Outcome measures

At post-test, attainment was measured using the paper versions of the Progress in English (PTE13) and Progress in Mathematics (PTM13) tests, which are standardised tests produced and validated by GL Assessment (2015a, 2015b). The independent evaluation team conducted the post-tests. They drew a random sample of 30 pupils in each school participating in the mathematics trial to complete the outcome test in mathematics and a random sample of 30 pupils in each school participating in the English trial to complete the outcome test in English.

Pre-test measures

Pupils' Key Stage 2 (KS2) national assessment results for mathematics and English (DfE, 2015) were used for pre-test measures of attainment, and were collected at the beginning of the school year in September 2015 through the National Pupil Database, as the pupils began Year 7. Full decimalised KS2 'fine points' scores (rather than simply levels) were used. Outcome attainment was measured at the end of the *following* academic year as pupils completed Year 8, after two intervening years of schooling, in June 2017.

Household socio-economic status

Household socio-economic status data were collected via questions on a pupil survey concerning parental/carer occupation, with categorisation according to the highest-status occupation between parents. Following this analysis (and given longstanding difficulties in judging the nature and content of some occupations), the tiered occupations were further categorised into three categories, higher, intermediate and lower, corresponding to the ONS three-class model (ONS, n.d.).

Set level

Schools in our sample varied in relation to the number of set levels they applied, from two to ten, with most falling between three and five (intervention schools in the setting trial had been specifically asked to cap the set level number at four). For the purposes of this current analysis, pupils were coded into three groups for English and mathematics, respectively, in each school: those in the very top set; those in the middle set(s); and those in the very bottom set. Thus, for a school with four sets, the top set was coded '1', the middle two sets coded '2' and the bottom set coded '3'. Similarly, for a school with five sets, the top set was coded '1', the middle three coded '2' and the bottom set coded '3'. The breakdowns of the sample by these three categories for English and mathematics are also shown in Table 1.

Analysis

The data were analysed in Stata 17.0 (StataCorp, 2021) by fitting a series of three multilevel models in each subject, mathematics and English, with pupils (level 1) clustered within individual subject sets (level 2) and then within schools (level 3). In each model, dummy variables representing the three categories of set level (top, middle and bottom) were included, along with other covariates representing pre-test attainment (KS2 in mathematics and English, respectively), gender, allocation to the intervention and total number of sets within the school. The principal model for each subject, M1, also included household occupation (SES) and ethnicity as covariates. However, as already noted, there was a large amount of missing data in these two variables. To investigate the effect of these missing data, we used two approaches. First, we ran two further models in each subject, M2 and M3, to assess the sensitivity of the results of the primary model, M1. Model M2 excluded household occupation (SES) and ethnicity as covariates and was based on the entire sample of pupils. Model M3 also excluded household occupation (SES) and ethnicity as covariates, but was based on the samples of pupils with complete data (i.e. the same dataset as for M1). Second, we used multiple imputation to impute the missing data for household occupation (SES) and ethnicity, then re-ran the principal model on the imputed dataset to compare this with the complete case analysis, M1. Our assumption is that robust, practically significant effects would not be sensitive to changes in the modelling.

The models were then used to estimate the adjusted mean attainment scores for pupils in the three set levels, controlling for these covariates. Practically, this was done by adding in a series of values to the model. These values consisted of either: the relevant values of the dummy variables for the set levels (i.e. either ‘0’ or ‘1’); or the mean scores for each of the other covariates included in the model; or ‘1’ for the constant. The standard deviations for each of the mean scores estimated were calculated using the standard error of the associated null model multiplied by the square root of the sample size to account for the clustered nature of the data, and the size of each subsample represented the total number in each category for whom there were full data (and thus whose data were included in the model).

Standardised effect sizes were calculated using Hedges' *g*. To account for the effects of clustering, 95% confidence intervals were calculated using the standard errors of the regression coefficient and transformed into an effect size to produce the upper and lower bounds of the effect size from the model.

RESULTS

A summary of the results for the main models, M1, M2 and M3, showing the effects on pupil attainment after experiencing setting for two school years, from the hierarchical regression models, is shown in Table 2, for English, and Table 3, for mathematics. A summary of the results of the imputation models is provided in TableS1. The findings show that after 2 years, there was a statistically significant increase in the attainment level for pupils in the top set when compared to the middle set(s) in both subjects, and this effect was robust across all three models and also for M1 on the imputed dataset. However, the effect is much larger for English than for mathematics, where the effect of prior attainment at KS2 is comparatively very much larger. The finding of lower attainment for pupils placed in the bottom set for English when compared to those in the middle set was

TABLE 2 Summary of three multilevel models used to compare post-test attainment by set level for English. Statistically significant results ($p < 0.05$) indicated in bold

Independent variables in the model	**M1**	**M2**	**M3**
Number of observations			
Pupils	476	841	476
	Coefficient (SE)	Coefficient (SE)	Coefficient (SE)

Pre-test score			
KS2 English fine score	**6.907 (0.587)**	**7.302 (0.391)**	**6.954 (0.594)**
Set allocation			
Top	**6.346 (1.041)**	**6.828 (0.872)**	**6.241 (1.060)**
Middle (reference category)			
Bottom	**−4.399 (1.564)**	−2.132 (1.141)	**−4.919 (1.587)**
Household socio-economic background (SES)			
Higher	**1.873 (0.571)**		
Intermediate	**1.432 (0.541)**		
Lower (Ref Cat)			
Ethnicity			
White	0.257 (0.638)		
Asian	**1.756 (0.761)**		
Black	0.055 (0.645)		
Other (Ref Cat)			
Gender			
Male	−0.580 (0.400)	**−0.819 (0.313)**	−0.407 (0.402)
Female (Ref Cat)			
School level			
No. of sets in school	1.552 (1.539)	0.769 (0.977)	1.564 (1.648)
Allocation to intervention	0.141 (1.792)	0.091 (1.493)	0.528 (1.913)
Constant	**30.647 (1.270)**	**28.718 (1.025)**	**30.384 (1.341)**
Variance			
School level	3.765 (0.667)	3.498 (0.560)	4.120 (0.709)
Set level	2.176 (0.806)	2.635 (0.595)	2.264 (0.808)
Pupil level	7.986 (0.301)	8.375 (0.235)	8.079 (0.305)
−2LL	1697.748	3035.758	1705.496

not of a consistent size across the models and was statistically significant in only two of the models, M1 and M3, but not for the third model, M2, based on the entire dataset including those pupils with missing SES and ethnicity data, nor for M1 on the imputed dataset. Hence, whilst the attainment of those placed in the bottom set for English is lower than those in the middle set, this effect was not robust across all models and the significant results for models M1 and M3 may have been subject to bias due to missing data. The attainment of pupils in the bottom set for mathematics was lower after 2 years compared to the middle set, although this effect was relatively small and not statistically significant in any of the models, and the imputation analysis showed an effect very close to zero. Hence, despite some negative trends, we found no evidence to indicate that the attainment of those in the bottom set decreased significantly relative to similar pupils placed in the middle set.

TABLE 3 Summary of three multilevel models used to compare post-test attainment by set level for mathematics. Statistically significant results ($p < 0.05$) indicated in bold

Independent variables in the model	**M1**	**M2**	**M3**
Number of observations			
Pupils	1237	2084	1237
	Coefficient (SE)	Coefficient (SE)	Coefficient (SE)
Pre-test score			
KS2 Mathematics fine score	**11.794 (0.327)**	**11.254 (0.244)**	**11.818 (0.325)**
Set allocation			
Top	**2.475 (0.641)**	**3.595 (0.528)**	**2.541 (0.642)**
Middle (reference category)			
Bottom	−1.1813 (0.862)	−0.693 (0.663)	−1.163 (0.863)
Family occupation (SES)			
Higher	**0.865 (0.324)**		
Intermediate	0.564 (0.317)		
Lower (Ref Cat)			
Ethnicity			
White	−0.524 (0.340)		
Asian	0.050 (0.309)		
Black	0.031 (0.287)		
Other (Ref Cat)			
Gender			
Male	−0.401 (0.227)	**−0.369 (0.181)**	−0.358 (0.228)
Female (Ref Cat)			
School level			
No. of sets in school	0.282 (0.589)	0.475 (0.531)	0.334 (0.599)
Allocation to intervention	0.125 (1.024)	0.132 (0.911)	0.173 (1.041)

Constant	30.269 (0.724)	29.610 (0.638)	30.228 (0.735)
Variance			
School level	3.263 (0.412)	3.242 (0.355)	3.337 (0.415)
Set level	1.873 (0.430)	2.211 (0.289)	1.872 (0.431)
Pupil level	7.416 (0.168)	7.569 (0.129)	7.448 (0.168)
−2LL	4306.291	7295.342	4312.143

The effect sizes for attainment of pupils in the top and bottom sets compared to the middle set for both subjects and for all three models are summarised in Tables 4 and 5 and illustrated graphically in Figures 1 and 2.

It can be seen from Table 4 and Figure 1 that, in mathematics, the relative increase for pupils placed in the top set compared to those in the middle set after controlling for prior attainment is consistent across the three models at $g = 0.1$. Table 4 and Figure 2 show that, in English, the relative increase for pupils placed in the top set compared to those in the middle set is also consistent across the models, but is almost three times as large at around $g = 0.27$.

In summary, when controlling for prior attainment, pupils in the top set performed significantly better than pupils in the middle and bottom sets in both English and mathematics, and

TABLE 4 Summary of effects on attainment by the top set for all four models (including imputation) in both subjects: mathematics and English. Statistically significant indicated in bold
results ($p <$

0.05)		Top set			Middle set (reference category)					
Subject	**Model**	) *N* **pupils schools**	**Adj. mean gain (SD adj. for clustering)**		) *N* **pupils schools**	**Adj. mean gain (SD adj. for clustering)**		**ES (*g*)**	**% confidence interval**	***p***
Maths	M1 (complete case) (	456 (61)	32.744	(24.455) (	643 (58)	30.269	(27.101)	**0.095**	**(0.047, 0.143)**	**<0.001**
	M2	718 (68)	33.205	(29.016)	1073 (66)	29.610	(28.965)	**0.124**	**(0.088, 0.160)**	**<0.001**
	M3	456 (61)	32.769	(24.455)	643 (58)	30.228	(27.101)	**95 0.098**	**(0.049, 0.146)**	**<0.001**
	M1 (imputed)	716 (68)	33.185	(28.899)	1071 (66)	29.630	(28.930)	**0.123**	**(0.087, 0.159)**	**<0.001**
English	M1 (complete case)	186 (28)	36.994	(23.067)	228 (27)	30.647	(21.159)	**0.288**	**(0.195, 0.380)**	**<0.001**
	M2	282 (32)	35.546	(25.174)	427 (30)	28.718	(24.345)	**0.276**	**(0.207, 0.346)**	**0.002**
	M3	186 (28)	36.624	(23.067)	228 (27)	30.384	(21.159)	**0.283**	**(0.189, 0.377)**	**<0.001**
	M1 (imputed)	293 (32)	35.357	(25.128)	425 (30)	28.618	25.128	**0.273**	**(0.087, 0.159)**	**<0.001**

TABLE 5 Summary of effects on attainment by the bottom set for all four models (including imputation) in both subjects: mathematics and English. Statistically significant results ($p <$ 0.05) indicated in bold

Subject		**Bottom set**			**Middle set (reference category)**					
	Model	*N* **pupils (schools)**	**Adj. mean gain (SD adj. for clustering)**		*N* **pupils (schools)**	**Adj. mean gain (SD adj. for clustering)**		**ES (*g*)**	**95 % confidence interval**	**Sig.**
Maths	M1 (complete case)	138 (57)	29.088	(15.533)	643 (58)	30.269	(27.101)	−0.046	(−0.113, 0.02)	0.171
	M2	293 (62)	28.917	(17.252)	1073 (66)	29.610	(28.965)	−0.026	(−0.074, 0.023)	0.296
	M3	138 (57)	29.187	(15.533)	643 (58)	30.228	(27.101)	−0.041	(−0.107, 0.026)	0.178
	M1 (imputed)	716 (68)	29.560	(17.254)	1071 (66)	29.630	(28.930)	−0.003	(−0.051, 0.046)	0.289
English	M1 (complete case)	62 (21)	26.248	(10.209)	228 (27)	30.647	(21.159)	**−0.227**	**(−0.384, −0.069)**	**<0.001**
	M2	132 (29)	26.586	(16.006)	427 (30)	28.718	(24.345)	−0.094	(−0.193, 0.005)	0.062
	M3	62 (21)	25.465	(10.209)	228 (27)	30.384	(21.159)	**−0.252**	**(−0.411, −0.093)**	**0.002**
	M1 (imputed)	132 (32)	26.603	(15.947)	1071 (30)	28.618	25.128	−0.089	(−0.188, 0.010)	0.077

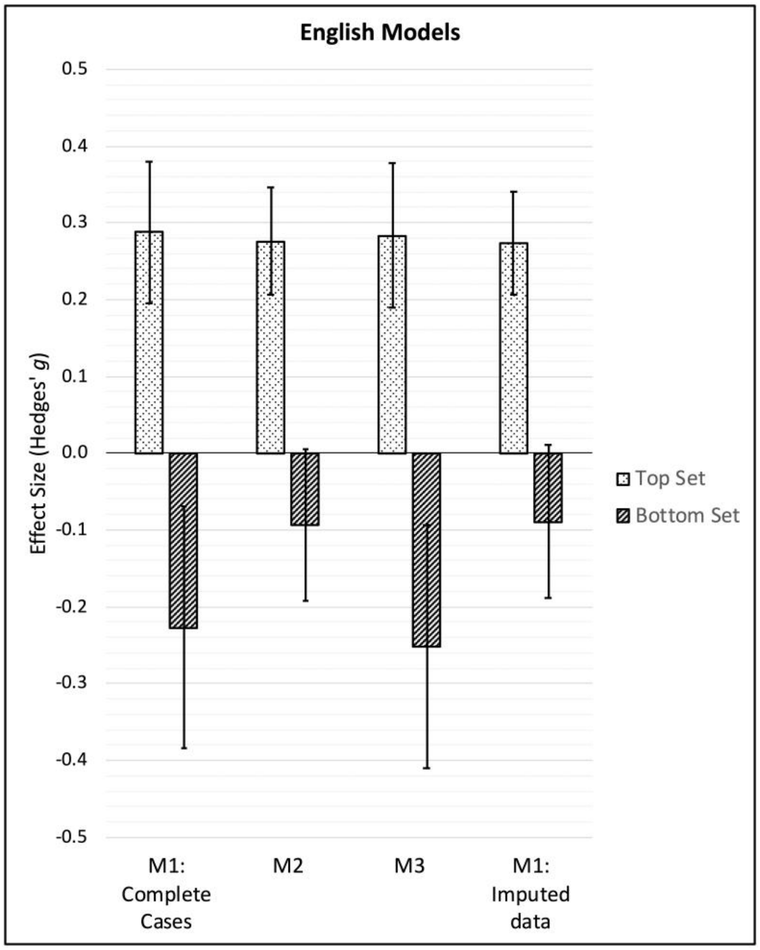

FIGURE 1 Post-test mean gains (with 95% confidence intervals) in attainment by set level (compared to middle set) for all four models in mathematics.

these effects were larger for English than for mathematics. However, our data suggest that pupils placed in the bottom set for English performed slightly worse than pupils placed in the middle set, although this trend was not statistically significant. In other words, our models indicate a widening gap in attainment, but provide more evidence of a relative benefit for pupils placed in top sets compared to all other pupils, rather than a relative disbenefit for those in bottom sets. In addition, our models suggest the effect is larger for English than mathematics.

DISCUSSION

Our study provides up-to-date evidence from a large-scale study in England to show that setting, between-class grouping by subject, is associated with positive impacts on pupils placed in high sets in comparison to those placed in middle and low sets, *after* controlling for prior attainment. This finding is broadly in line with Ireson et al.'s (2005) now dated results from the early part of this century. In other words, in our study, a pupil who was allocated to a high set tended to make larger gains than a pupil of similar prior attainment who was

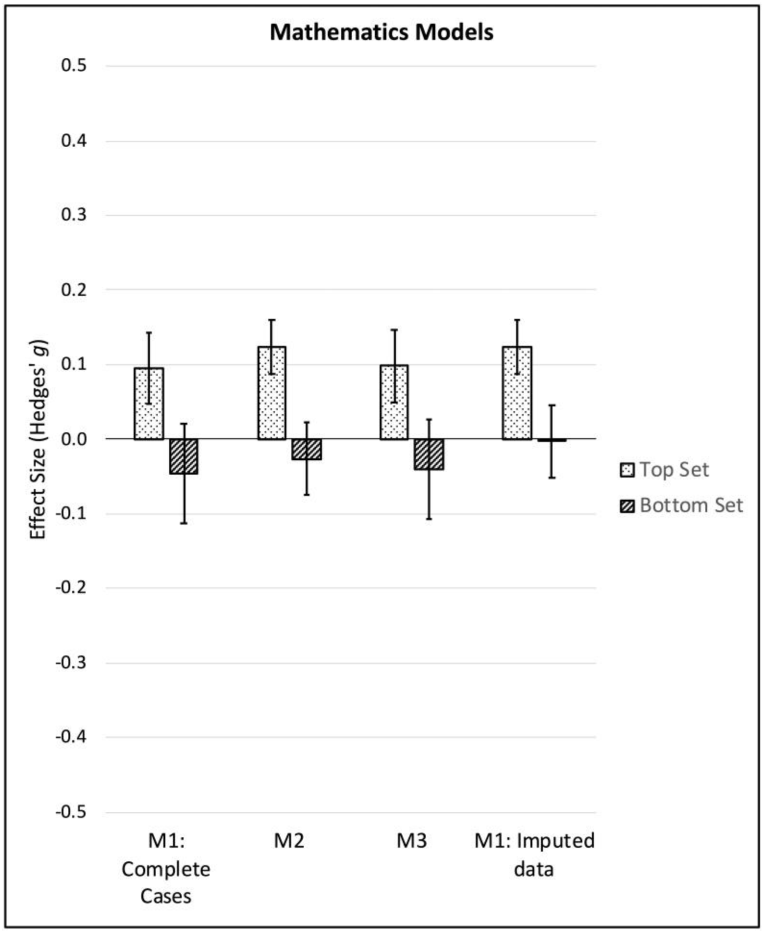

FIGURE 2 Post-test mean gains (with 95% confidence intervals) in attainment by set level (compared to middle set) for all four models in English.

placed in a middle or low set. Our study provides stronger and more robust evidence for placement in a top set as a key factor in increasing pupil attainment. Additionally, in contrast to Ireson et al., who found similar effect sizes across subjects, we found a much larger effect for English in comparison to mathematics.

Before examining the implications of these findings, there are two important points to make. First, our results do not indicate that setting benefits high-attaining pupils. Rather, they show that setting benefits those pupils who are *placed in higher sets*. There is a great deal of evidence highlighting how pupils are misallocated to high and low sets, and this results in the over-representation of pupils from Black and minority ethnic backgrounds in lower sets (e.g. Connolly et al., 2019) and pupils from socially disadvantaged backgrounds in lower sets (e.g. Kutnick et al., 2005). Hence, in benefitting pupils allocated to top sets, this disadvantages those pupils misallocated to middle or low sets. Second, our results indicate a relative advantage for pupils placed in top sets, but they do not show that these pupils performed better than they would otherwise have done in a class of mixed attainment.

These findings are of concern from educational and social justice perspectives. They illustrate a growing attainment gap, and divergence between top-set pupils in comparison with pupils in middle and bottom sets. This self-fulfilling prophecy (Merton, 1948) affecting attainment and pupil self-confidence (Francis, Craig et al., 2020; Francis, Taylor et al., 2020) may be due to a Pygmalion effect (Rosenthal & Jacobson, 1992), specifically for those pupils assigned to top sets, who receive more teacher encouragement and higher expectations (cf. Wang et al., 2021). Alternatively, it may be that pupils in top sets are offered a richer curriculum with much greater opportunity to learn (Burris et al., 2006). Or it may be that top sets are allocated better qualified and more experienced teachers (Francis et al., 2019).

This widening gap is of concern to educationalists, as failing to promote the educational thriving and effective learning for pupils that all educational professionals intend. It is also of concern to policymakers. The United Kingdom is famously dogged by a 'long tail' of underachievement (Marshall, 2013), and our findings provide a clear potential explanation, given the prevalence of within-school tracking in our system (Taylor et al., 2020). Moreover, our findings also highlight that, in spite of the envisaged equality of entitlement to high-quality educational provision facilitated by comprehensive state education, provision is inequitable, with some pupils advantaged and others disadvantaged.

But our findings also have implications for interventions directed at addressing disadvantage in education. For pupils placed in top sets, the effect sizes that we found are of the order, and for English larger, than are identified in most educational trials (see e.g. Cheung & Slavin, 2016). In addition, the effect sizes for low set placement in English, whilst not statistically significant or consistent across all three models, were nevertheless negative and at least of the order of those identified in most educational trials. In mathematics, the effect sizes for low set placement were small, but nevertheless negative. As we have highlighted, socially disadvantaged pupils (and those from certain minority ethnic groups) are over-represented in, and often misallocated to, lower sets (Connolly et al., 2019). And yet, as we noted earlier, many schools use attainment grouping as one element of a strategy to address educational disadvantage (Macleod et al., 2015). Our results suggest that, especially in English, this may be at best counter-productive and that, despite the best efforts of schools, the effects of attainment grouping may counteract the effects of genuinely beneficial interventions.

The findings of greater significance for setting in the case of English for pupil outcomes (positive and negative) also suggest that: (a) there may be different impacts of setting for different curriculum subject areas, demanding further research in this area; and (b) schools concerned with equity should review setting in English. Interestingly, setting is somewhat less prevalent in English compared to mathematics, in England.

There are three limitations with our study. First, there was no control group in which a different form of grouping practice, such as mixed attainment, was used. Hence, we cannot be certain whether the effects on attainment are either caused or exacerbated by setting, nor can we say whether setting resulted in higher attainment for those placed in top sets than would otherwise have been the case. Nevertheless, our findings are in line with much of the previous literature in that they do strongly suggest that attainment grouping is associated with a widening attainment gap, which is due to a relative, but not necessarily an absolute, advantage for those pupils placed in top sets. Second, there was significant attrition of schools from the study, and the remaining schools could be atypical and committed to good practice and equity, given (a) their original voluntary participation in a study focused on best practice in setting and (b) their dedicated completion of the 2-year period of study. Nevertheless, they reflect a national sample, and any atypicality as 'conscientious schools' might be postulated to have mitigated the trends identified, rather than exacerbating them. Third, no measures of teaching quality or opportunity to learn were applied and, hence, we cannot say whether the observed effects are due to setting per se or are a result of the effect of setting on teaching quality or opportunity to access curriculum content.

Finally, our results highlight important issues for further research into the effects of setting in different subjects. There is also an urgent need for more robust research into the effects of setting as compared to mixed-attainment grouping and to investigate the relationships between setting, teacher quality, opportunity to learn and attainment. Despite 100 years of research into the effects of ability grouping, the evidence is still inconclusive. It is clear that research in this area is technically, methodologically and practically difficult. Previous studies highlight some of these difficulties. Ideally, one would carry out an RCT comparing pupils from schools randomly assigned into groups with setting or mixed-attainment classes. This is simply not feasible at scale, because the effort—and time—needed to effect a change in attainment grouping across a school is considerable (Taylor et al., 2019). However, naturalistic studies are also problematic. For example, in Betts and Shkolnik's (2000) comparison of schools with and without a school policy to group pupils by 'ability', classes amongst the no-grouping schools were no less stratified than in those schools with a grouping policy. In other words, despite the official policy, informal grouping by attainment was used. In a new study with which some of the authors are engaged (Hodgen et al., 2019), we seek to approach this methodological challenge by comparing carefully selected and robustly matched samples of schools already using different forms of grouping and examining the effects of teacher quality and opportunity to learn. We hope that this study will help answer this important outstanding question.

ACKNOWLEDGEMENTS

We gratefully acknowledge the following for their assistance throughout the project, without which it would not have been possible: the evaluation team at the National Foundation for Educational Research, particularly Ben Styles and Palak Roy; the Department for Education (DfE) National Pupil Database (NPD) team; the EEF staff, including Anneka Dawson and Emily Yeomans; and, finally, the schools, teachers and pupils who took part in the research. We are also grateful to Jake Anders for advice and comments. The analysis for this paper was conducted whilst Antonina Tereshchenko was at UCL and Paul Connolly was at Lancaster University.

FUNDING INFORMATION

This research was funded by the Education Endowment Foundation (EEF).

CONFLICT OF INTEREST

The authors have no conflict of interest to disclose. Note that, since this research was conducted, Becky Francis has become Chief Executive of the Education Endowment Foundation.

DATA AVAILABILITY STATEMENT

The data are not publicly available due to privacy or ethical restrictions. This is due to recent changes in how National Pupil Database extracts can be shared.

ETHICS STATEMENT

The study was approved by the Research Ethics Committee of King's College London and Queen's University Belfast and, later, UCL Institute of Education (now IOE – Faculty of Education and Society, University College London).

ORCID

Jeremy Hodgen https://orcid.org/0000-0002-9196-4088
Becky Taylor https://orcid.org/0000-0002-7257-4463
Becky Francis https://orcid.org/0000-0002-9966-0003
Nicola Bretscher https://orcid.org/0000-0002-1226-4025
Antonina Tereshchenko https://orcid.org/0000-0002-4443-3188
Paul Connolly https://orcid.org/0000-0001-9176-9592
Anna Mazenod https://orcid.org/0000-0003-0175-1634

ENDNOTES

[1] The term 'ability grouping' is frequently applied in the United Kingdom. We avoid this terminology, which suggests a perception of 'ability' as fixed. We refer instead to 'attainment grouping'.

[2] The findings of the cluster RCT have previously been reported in the evaluation report (Roy et al., 2018), showing no significant difference between the intervention and control groups on the outcome measures of pupil attainment and self-confidence.

[3] 211 students were included in both the English and the mathematics samples.

[4] Unfortunately, a national breakdown of students' household socio-economic background using the ONS (n.d.) three-class model is not available.

REFERENCES

Abraham, J. (2008). Pupils' perceptions of setting and beyond—a response to Hallam and Ireson. *British Educational Research Journal*, *34*(6), 855–863. https://doi.org/10.1080/01411920802044511

Archer, L., Francis, B., Miller, S., Taylor, B., Tereshchenko, A., Mazenod, A., Pepper, D., & Travers, M.-C. (2018). The symbolic violence of setting: A Bourdieusian analysis of mixed methods data on secondary students' views about setting. *British Educational Research Journal*, *44*(1), 119–140. https://doi.org/10.1002/berj.3321

Berends, M., & Donaldson, M. (2016). Does the organization of instruction differ in charter schools? Ability grouping and students' mathematics gains. *Teachers College Record*, *118*(11), 1–38.

Betts, J. R., & Shkolnik, J. L. (2000). Key difficulties in identifying the effects of ability grouping on student achievement. *Economics of Education Review*, *19*(1), 21–26. https://doi.org/10.1016/S0272-7757(99)00022-9

Boaler, J. (1997). Setting, social class and the survival of the quickest. *British Educational Research Journal*, *23*(5), 575–595. https://doi.org/10.1080/0141192970230503

Borghans, B. L., Diris, R., Smits, W., & de Vries, J. (2020). Should we sort it out later? The effect of tracking age on long-run outcomes. *Economics of Education Review*, *75*, 101973. https://doi.org/10.1016/j.econedurev.2020.101973

Bosworth, R. (2013). What sort of school sorts students? *International Journal of Quantitative Research in Education*, *1*(1), 20–38. https://doi.org/10.1504/ijqre.2013.055639

Burris, C. C., Heubert, J. P., & Levin, H. M. (2006). Accelerating mathematics achievement using heterogeneous grouping. *American Educational Research Journal*, *43*(1), 105–136. https://doi.org/10.3102/000283120430011

Buttaro, A. J., & Catsambis, S. (2019). Ability grouping in the early grades: Long-term consequences for educational equity in the United States. *Teachers College Record*, *121*(2), 1–50.

Campbell, T. (2014). Stratified at seven: In-class ability grouping and the relative age effect. *British Educational Research Journal*, *40*(5), 749–771. https://doi.org/10.1002/berj.3127

Campbell, T. (2017). The relationship between stream placement and teachers' judgements of pupils: Evidence from the Millennium Cohort Study. *London Review of Education*, *15*(3), 505–522. https://doi.org/10.18546/ LRE.15.3.12

Capsada-Munsech, Q., & Boliver, V. (2019). *Educational tracking and sorting in England: Education system, reforms, trends and empirical evidence for the 1970 Birth Cohort Study (BCS70)*. DIAL.

Cheung, A. C. K., & Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher*, *45*(5), 283–292. https://doi.org/10.3102/0013189X16656615

Connolly, P., Biggart, A., Miller, S., O'Hare, L., & Thurston, A. (2017). *Using randomised controlled trials in education*. Sage.

Connolly, P., Taylor, B., Francis, B., Archer, L., Hodgen, J., Mazenod, A., & Tereshchenko, A. (2019). The misallocation of students to academic sets in maths: A study of secondary schools in England. *British Educational Research Journal*, *45*(4), 873–897. https://doi.org/10.1002/berj.3530

DfE. (2015). *National curriculum assessments at Key Stage 2 in England, 2015 (revised): Statistical first release*. Department for Education.

Domina, T., McEachin, A., Hanselman, P., Agarwal, P., Hwang, N., & Lewis, R. W. (2019). Beyond tracking and detracking: The dimensions of organizational differentiation in schools. *Sociology of Education*, *92*(3), 293–322. https://doi.org/10.1177/0038040719851879

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. https://doi.org/10.1080/09585176. 2011.627206

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. https://doi.org/10.1080/0305764X.2015.1093095

Francis, B., Craig, N., Hodgen, J., Taylor, B., Tereshchenko, A., Connolly, P., & Archer, L. (2020). The impact of tracking by attainment on pupil self-confidence over time: Demonstrating the accumulative impact of self-fulfilling prophecy. *British Journal of Sociology of Education*, *41*(5), 626–642. https://doi.org/10.1080/01425692.2020 .1763162

Francis, B., Hodgen, J., Craig, N., Taylor, B., Archer, L., Mazenod, A., Tereshchenko, A., & Connolly, P. (2019). Teacher 'quality' and attainment grouping: The role of within-school teacher deployment in social and educational inequality. *Teaching and Teacher Education*, *77*, 183–192. https://doi.org/10.1016/j.tate.2018.10.001

Francis, B., Taylor, B., & Tereshchenko, A. (2020). *Reassessing 'ability' grouping: Improving practice for equity and attainment*. Routledge.

Gamoran, A., & Mare, R. D. (1989). Secondary school tracking and educational inequality: Compensation, reinforcement, or neutrality? *American Journal of Sociology*, *94*, 1146–1183. https://doi.org/10.1086/229114

Gamoran, A., Nystrand, M., Berends, M., & LePore, P. C. (1995). An organizational analysis of the effects of ability grouping. *American Educational Research Journal*, *32*(4), 687–715. https://doi.org/10.3102/00028312032004687

Gillard, D. (2018). *Education in England: A brief history*. Retrieved 8 January 2021 from www.educationengland.org.uk/history

GL Assessment. (2015a). *Progress test in mathematics*. Retrieved 8 January 2021 from https://www.gl-assessment. co.uk/media/1346/ptm-technical-information.pdf

GL Assessment. (2015b). *Progress test in English*. Retrieved 8 January 2021 from https://www.gl-assessment. co.uk/media/1366/pte-technical-information.pdf

Hallam, S., & Ireson, J. (2005). Secondary school teachers' pedagogic practices when teaching mixed and structured ability classes. *Research Papers in Education*, *20*(1), 3–24. https://doi.org/10.1080/0267152052000341318

Hallam, S., & Parsons, S. (2012). Prevalence of streaming in UK primary schools: Evidence from the Millennium Cohort Study. *British Educational Research Journal*, *39*, 514–544. https://doi.org/10.1080/01411926.2012.659721

Hallinan, M. T. (1994). School differences in tracking effects on achievement. *Social Forces*, *72*(3), 799–820. https:// doi.org/10.2307/2579781

Hanushek, E. A., & Wößmann, L. (2006). Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries. *The Economic Journal*, *116*(510), C63–C76. https://doi. org/10.1111/j.1468-0297.2006.01076.x

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics*, *32*(4), 341–370. https://doi.org/10.3102/1076998606298043

Higgins, S., Major, L. E., Coleman, R., Katsipataki, M., Henderson, P., Mason, D., Aguilera, A. B. V., & Kay, J. (2018). *The Sutton Trust–Education Endowment Foundation teaching and learning toolkit*. Education Endowment Foundation.

Hodgen, J., Foster, C., & Brown, M. (2022). Low attainment in mathematics: An analysis of 60 years of policy discourse in England. *The Curriculum Journal*, *33*(1), 5–24. https://doi.org/10.1002/curj.128

Hodgen, J., Taylor, B., Anders, J., Tereshchenko, A., & Francis, B. (2019). *The Student Grouping Study: Investigating the effects of setting and mixed attainment grouping*. Education Endowment Foundation.

Houtte, M. V., Demanet, J., & Stevens, P. A. (2012). Self-esteem of academic and vocational students: Does within-school tracking sharpen the difference? *Acta Sociologica*, *55*(1), 73–89. https://doi. org/10.1177/0001699311431595

Ireson, J., & Hallam, S. (2001). *Ability grouping in education*. Paul Chapman.

Ireson, J., & Hallam, S. (2009). Academic self-concepts in adolescence: Relations with achievement and ability grouping in schools. *Learning and Instruction*, *19*(3), 201–213. https://doi.org/10.1016/j.learninstruc.2008.04.001

Ireson, J., Hallam, S., Hack, S., Clark, H., & Plewis, I. (2002). Ability grouping in English secondary schools: Effects on attainment in English, mathematics and science. *Educational Research and Evaluation*, *8*(3), 299–318. https://doi.org/10.1076/edre.8.3.299.3854

Ireson, J., Hallam, S., & Hurley, C. (2005). What are the effects of ability grouping on GCSE attainment? *British Educational Research Journal*, *31*(4), 443–458. https://doi.org/10.1080/01411920500148663 Jackson, P. W. (1968). *Life in classrooms*. Holt, Rinehart & Winston.

Jaremus, F., Gore, J., Fray, L., & Prieto-Rodriguez, E. (2020). Grouped out of STEM degrees: The overlooked mathematics 'glass ceiling' in NSW secondary schools. *International Journal of Inclusive Education*, *26*(11), 1141–1157. https://doi.org/10.1080/13603116.2020.1776778

Jerrim, J. (2019). *England's schools segregate by ability more than almost every other country in the world*. Retrieved 8 January 2021 from https://ffteducationdatalab.org.uk/2019/09/englands-schools-segregate -by-ability-more-than-almost-every-other-country-in-the-world/

Kelly, S. (2004). Are teachers tracked? On what basis and with what consequences. *Social Psychology of Education*, *7*(1), 55–72. https://doi.org/10.1023/B:SPOE.0000010673.78910.f1

Kerckhoff, A. C. (1986). Effects of ability grouping in British secondary schools. *American Sociological Review*, *51*(6), 842–858. https://doi.org/10.2307/2095371

Kulik, C. C., & Kulik, J. A. (1984). *Effects of ability grouping on elementary school pupils: A meta-analysis*. Paper presented at the Annual Meeting of the American Psychological Association, Toronto, Ont., Canada.

Kulik, J. A., & Kulik, C. (1992). Meta-analytic findings on grouping programs. *Gifted Child Quarterly*, *36*(2), 73–77. https://doi.org/10.1177/001698629203600204

Kutnick, P., Sebba, J., Blatchford, P., Galton, M., Thorp, J., MacIntyre, H., & Berdondini, L. (2005). *The effects of ability grouping: A literature review*. DfES Research Report RR688. Department for Education and Skills (DfES).

Lou, Y., Abrami, P. C., Spence, J. C., Poulsen, C., Chambers, B., & d'Apollonia, S. (1996). Within-class grouping: A meta-analysis. *Review of Educational Research*, *66*(4), 423–458. https://doi.org/10.3102/00346543066004423 Loveless, T. (1999). *The tracking wars: State reform meets school policy*. Brookings Institute.

Macleod, S., Sharp, C., Bernardinelli, D., Skipp, A., & Higgins, S. (2015). *Supporting the attainment of disadvantaged pupils: Articulating success and good practice*. Department for Education.

Magableh, I. S. I., & Abdullah, A. (2021). The impact of differentiated instruction on students' reading comprehension attainment in mixed-ability classrooms. *Interchange*, *52*(2), 255–272. https://doi.org/10.1007/ s10780-021-09427-3

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher*, *43*(6), 304–316. https://doi.org/10.3102/0013189x14545513

Marshall, P. (Ed.). (2013). *The tail: How England's schools fail one child in five and what can be done*. Profile Books.

Martinková, P., Hladká, A., & Potužníková, E. (2020). Is academic tracking related to gains in learning competence? Using propensity score matching and differential item change functioning analysis for better understanding of tracking implications. *Learning and Instruction*, *66*, 101286. https://doi.org/10.1016/j.learninstruc.2019.101286

Matthewes, S. H. (2021). Better together? Heterogeneous effects of tracking on student achievement. *The Economic Journal*, *131*(635), 1269–1307. https://doi.org/10.1093/ej/ueaa106

Merton, R. (1948). The self-fulfilling prophecy. *The Antioch Review*, *8*(2), 193–210.

Mitchell, J. C. (1984). Typicality and the case study. In R. F. Ellen (Ed.), *Ethnographic research: A guide to general conduct* (pp. 238–241). Academic Press.

Moller, S., & Stearns, E. (2012). Tracking success: High school curricula and labor market outcomes by race and gender. *Urban Education*, *47*(6), 1025–1054. https://doi.org/10.1177/0042085912454440

Muijs, D., & Dunne, M. (2010). Setting by ability, or is it? A quantitative study of determinants of set placement in English secondary schools. *Educational Research*, *52*(4), 391–407. https://doi.org/10.1080/00131881.2010.5 24750

Oakes, J. (1995). Two cities' tracking and within-school segregation. *Teachers College Record*, *96*(4), 681–690.

OECD. (2016). *PISA 2015 results (volume II): Policies and practices for successful schools*. OECD Publishing.

OFSTED. (2016). *The annual report of Her Majesty's Chief Inspector of Education, Children's Services and Skills 2015/16*. Her Majesty's Stationery Office.

ONS. (n.d.). *The National Statistics Socio-economic Classification (NS-SEC)*. Retrieved 11 February 2022 from https://www.ons.gov.uk/methodology/classificationsandstandards/otherclassifications/thenationalstatisticsso cioeconomicclassificationnssecrebasedonsoc2010#classes-and-collapses

Papay, J. P., & Kraft, M. A. (2015). Productivity returns to experience in the teacher labor market: Methodological challenges and new evidence on long-term career improvement. *Journal of Public Economics*, *130*, 105–119. https://doi.org/10.1016/j.jpubeco.2015.02.008

Rosenbaum, J. E. (1999). If tracking is bad, is detracking better? A study of a detracked high school. *American Educator*, *23*, 24–47.

Rosenthal, R., & Jacobson, L. (1992). *Pygmalion in the classroom: Teacher expectation and pupils' intellectual development*. Irvington Publishers.

Roy, P., & Styles, B. (2017). *Statistical analysis plan for best practice in setting*. Education Endowment Foundation.

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*. 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. https://doi.org/10.1111/j.1756-5391. 2009.01032.x

Slavin, R. E. (1987). Achievement effects of ability grouping in elementary schools: A best evidence synthesis. *Review of Educational Research*, *57*, 293–336. https://doi.org/10.3102/00346543057003293

Slavin, R. E. (1990). Achievement effects of ability grouping in secondary schools: A best evidence synthesis. *Review of Educational Research*, *60*, 471–499. https://doi.org/10.3102/00346543060003471 StataCorp. (2021). *Stata: Release 17. Statistical Software*. StataCorp LLC.

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. *Review of Educational Research*, *86*(4), 849–899. https://doi.org/10.3102/0034654316675417

Strand, S. (2012). The White British–Black Caribbean achievement gap: Tests, tiers and teacher expectations. *British Educational Research Journal*, *38*(1), 75–101. https://doi.org/10.1080/01411926.2010.526702

Styles, B., & Torgerson, C. (2018). Randomised controlled trials (RCTs) in education research—methodological debates, questions, challenges. *Educational Research*, *60*(3), 255–264. https://doi.org/10.1080/00131881.2 018.1500194

Taylor, B., Francis, B., Craig, N., Archer, L., Hodgen, J., Mazenod, A., Tereshchenko, A., & Pepper, D. (2019). Why is it difficult for schools to establish equitable practices in allocating students to attainment 'sets'? *British Journal of Educational Studies*, *67*(1), 5–24. https://doi.org/10.1080/00071005.2018.1424317

Taylor, B., Hodgen, J., Tereshchenko, A., & Gutiérrez, G. (2020). Attainment grouping in English secondary schools: A national survey of current practices. *Research Papers in Education*, *37*, 199–220. https://doi.org/10.1080/0 2671522.2020.1836517

Timmermans, A. C., Kuyper, H., & van der Werf, G. (2015). Accurate, inaccurate, or biased teacher expectations: Do Dutch teachers differ in their expectations at the end of primary education? *British Journal of Educational Psychology*, *85*(4), 459–478. https://doi.org/10.1111/bjep.12087

Waldfogel, J., & Washbrook, E. V. (2010). *Low income and early cognitive development in the UK: A report for the Sutton Trust*. Sutton Trust.

Wang, H., King, R. B., & McInerney, D. M. (2021). Ability grouping and student performance: A longitudinal investigation of teacher support as a mediator and moderator. *Research Papers in Education*. https://doi.org/10.108 0/02671522.2021.1961293

Wiliam, D., & Bartholomew, H. (2004). It's not which school but which set you're in that matters: The influence of ability grouping practices on student progress in mathematics. *British Educational Research Journal*, *30*(2), 279–293. https://doi.org/10.1080/0141192042000195245

Wilkinson, S. D., & Penney, D. (2014). The effects of setting on classroom teaching and student learning in mainstream mathematics, English and science lessons: A critical review of the literature in England. *Educational Review*, *66*(4), 411–427. https://doi.org/10.1080/00131911.2013.787971

Wilkinson, S. D., Penney, D., Allin, L., & Potrac, P. (2020). The enactment of setting policy in secondary school physical education. *Sport, Education and Society*, *26*, 619–633. https://doi.org/10.1080/13573322.2020.1784869

SUPPORTING INFORMATIONAdditional supporting information can be found online in the Supporting Information section at the end of this article.