DOI: 10.1111/cdev.13674

EMPIRICAL ARTICLE

CHILD DEVELOPMENT

Ability-grouping and problem behavior trajectories in childhood and adolescence: Results from a U.K. population-based sample

Efstathios Papachristou[1] | **Eirini Flouri**[1] | **Heather Joshi**[1] | **Emily Midouhas**[1] | **Glyn Lewis**[2]

[1]UCL Institute of Education, University College London, London, UK

[2]Division of Psychiatry, University College London, London, UK

Correspondence
Efstathios Papachristou, Department of Psychology and Human Development, UCL Institute of Education, 25 Woburn Square, London WC1H 0AA, UK.
Email: Efstathios.papachristou@ucl.ac.uk

Funding information
Economic and Social Research Council, Grant/Award Number: ES/N007921/1

Abstract

Ability-grouping has been studied extensively in relation to children's academic, but not emotional and behavioral outcomes. The sample comprised 7259 U.K. children (50% male) with data on between-class and within-class ability-grouping at age 7. Peer, emotional, hyperactivity, and conduct problems were measured at ages 7, 11, and 14 years. Children in low within-class ability groups showed more hyperactivity and emotional problems across the study period compared to non-grouped children, after adjustments for the different types of ability grouping and confounding. Additionally, children in the middle within-class ability groups showed more, and those in the top within-class groups less, hyperactivity compared to non-grouped children, after adjustment. Children in lower within-class groups should be monitored closely to ensure that their well-being is not compromised.

Ability-grouping pupils within schools, also referred to as attainment-grouping (Taylor et al., 2018) or tracking, has a long history in the United Kingdom and has attracted much research and debate (Ireson & Hallam, 1999). In primary schools in the United Kingdom, two main types of *between-class* ability-grouping are practiced: streaming and setting. Streamed pupils stay in a group of children with the similar ability for all lessons, while set pupils are placed in an ability group only for certain lessons (Ireson & Hallam, 2001). The term "tracking," which is often used in the U.S. literature, refers to practices analogous to setting or streaming (Gamoran & Nystrand, 1994). *Within-class* ability-grouping is a third type of ability-grouping and involves teachers organizing pupils into small groups by their skill levels. The three types of ability-grouping are not mutually exclusive, and children can be streamed, placed in sets, and allocated to within-class groups concurrently (Wilkinson et al., 2016).

Within-school ability-grouping often starts early in the United Kingdom. Evidence from the Millennium Cohort Study (MCS), a U.K. population-based cohort of more than 19,000 children born around 2001, suggests that a significant proportion of U.K. primary school children, as young as seven, are in streamed classes (Hallam & Parsons, 2013b); at age 7, some 16% of the MCS children were streamed, 64% of whom were also set for literacy and 70% for mathematics. Nearly 26% of children were set for both literacy and maths, and 11% were set for only maths (8%) or literacy (3%). Within-class grouping appears to be the most prevalent practice with 79% of MCS pupils in England reported to be in-class grouped at age 7 (Campbell, 2014).

Abbreviations: BFLPE, big-fish-little-pond-effect; CFA, confirmatory factor analyses; ICC, intraclass correlation coefficients; MCS, Millennium Cohort Study; MICE, multivariate imputation by chained equations; MLMs, multilevel linear models; MREC, multi-centre research ethics committee; SDQ, strengths and difficulties questionnaire.

Efstathios Papachristou and Eirini Flouri joint first authors.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2021 The Authors. *Child Development* published by Wiley Periodicals LLC on behalf of Society for Research in Child Development

Ability group allocation in U.K. primary schools is (at least for streaming and setting) unlikely to reflect the current class teacher's decisions. As streaming takes place at the whole-year level, placement is typically determined by some combination of performance in previous years, assessments by previous years' teachers, pre-established placements and/or school-based test performance. This does not mean that children's characteristics, unrelated to "ability," may not be influential. Once streams have been decided upon, the subsets of pupils are usually allocated to one of the year group's assigned class teachers. In contrast, in grouping by ability within-class, a key decision maker is likely to be the class teacher (Campbell, 2017).

Ability-grouping and academic outcomes

The main purpose of grouping students into ability groups is to provide instruction that is tailored to their individual academic ability, thereby enhancing their school performance. Although the practice is likely to have far-reaching implications in terms of the students' healthy behavioral and emotional development, it was not until very recently that studies attempted to examine the impact of ability grouping on mental health outcomes directly (Lipps et al., 2010; Müller & Hofmann, 2016; Müller & Zurbriggen, 2016; Van Houtte & Stevens, 2008). Rather unsurprisingly given the prime basis of implementing ability grouping, the literature to date has mainly focused on examining the associations between ability grouping and academic performance. A large body of studies has also examined the associations between ability grouping and academic self-concept, a strong correlate of academic performance but also of emotional and behavioral problems. Therefore, before reviewing the evidence on the direct relations between ability grouping and mental health outcomes, we briefly review its associations with academic performance and self-concept. It is likely that the same underlying mechanisms can explain the associations between ability grouping with those outcomes and adverse mental health outcomes too.

Studies in the United Kingdom examining the impact of ability grouping on attainment have mixed findings. Some suggest that there is no beneficial effect of ability-grouping (Boaler et al., 2000; Ireson & Hallam, 2001; Kutnick et al., 2005). Others show a marginal benefit for high attainers but a more significant detrimental impact for low attainers (Boaler & Wiliam, 2001). In line with the latter, findings from the MCS suggest that children placed in a top stream appear to be achieving more and to make significantly more academic progress than other children attending schools that do not stream, while children in middle or bottom streams achieve less and make significantly less academic progress (Parsons & Hallam, 2014). However, the extent to which such apparent peer spillover effects represent an artifact of poor methodological approaches, in particular, measurement error at the individual (child) level or lack of sufficient controls for pre-existing differences, has been the subject of much debate (Dicke et al., 2018; Marks, 2015; Marsh et al., 2000; Televantou et al., 2015; Trautwein et al., 2009). The reason why some question the positive peer achievement spillovers within high-ability groups is because pupils placed in higher ability groups have lower academic self-concept compared to those in lower-ability groups (Marsh et al., 2001; Preckel et al., 2010), or the big-fish-little-pond-effect (BFLPE) (Marsh & Hau, 2003; Marsh et al., 2008). According to the BFLPE, a pupil compares her own achievement with the achievement of other members within her group. In high-ability groups, such comparisons are more likely to be unfavorable because children in those groups are more likely to feel insecure about their own achievement, which results in lower academic self-concept (the "big fish" does not feel so big when placed in a pond with other "big fish").

The seemingly contradictory findings that high-achieving environments are advantageous for academic progress but hinder academic self-concept have been extensively discussed by Dicke et al. (2018) and Marsh et al. (2000). Dicke et al. suggest that the positive effect of a high-achieving environment on academic progress is likely a "phantom effect" and that high-achieving environments have negative effects on pupils' academic achievement once appropriate methodological approaches are followed. Marsh et al. argue that pupils in high-performing selective schools might suffer from a poor self-concept owing to comparisons of their ability levels with those of other pupils in their immediate context (a contrast effect, also known as "frame of reference" effect). However, at the same time, the perception of an improved school status can have an advantageous impact on their self-concept ("reflected glory" or "assimilation effect"). Marsh et al. propose that the negative BFLPE effect on pupils' self-concept in selective schools is the net effect of strong contrast effects and the weaker positive "assimilation effects" (Marsh et al., 2000). Ignoring "reflected glory" effects may, arguably, explain at least some of the mixed results about the impact of setting on academic self-concept (Ireson & Hallam, 2005, 2009).

The relations between *within*-class ability-grouping and academic outcomes have attracted less research, despite the fact that within-class is the most prevalent type of ability grouping and has become standard practice in British primary schools (MacIntyre & Ireson, 2002). In the United Kingdom, within-class ability-grouping was originally seen as a means of raising attainment that avoids the social disadvantages associated with streaming and setting (Harlen, 1997) by promoting greater trust and acceptance among students of different social classes, races, and sexes. It was seen as a way to facilitate social interactions and collaborative learning among students. Specifically, it was anticipated that it would give to the quieter students an opportunity to participate by giving them the option to express their opinions among a small group of classmates that they perceive as equally skilled and knowledgeable. Nonetheless, it seems that within-class ability-grouping is subject to some of the same pitfalls of between-class ability-grouping, described in a later section. For example, it is common for teachers to misallocate children in within-class ability groups according to their perceptions of the child's ability (MacIntyre & Ireson, 2002). The study by MacIntyre and Ireson (2002) additionally showed that teachers influence how far within-class ability-grouping will affect children's self-concept, which can, in turn, influence their achievement and academic self-concept. Hence, within-class ability-grouping might in fact be limiting, rather than facilitating, children's learning, and their emotional and behavioral development.

Ability-grouping and psychological outcomes

Predictions about the effect of ability-grouping on psychological outcomes, such as emotional and behavioral problems, are difficult to make because such outcomes are positively correlated with both academic performance and self-concept (E. J. Lee & Stone, 2012; Moilanen et al., 2010); hence, negative contrast and positive assimilation effects are likely to be operating, and possibly amplified, due to selection in ability groups. It is important to consider selection in research on nonacademic effects of ability-grouping. For example, selection could be behind the evidence from cross-sectional studies suggesting that belonging to lower tracks (both between- and within-school) is associated with higher rates of delinquency and depressive symptomatology (Lipps et al., 2010; Van Houtte & Stevens, 2008). Similarly, a recently published systematic review suggests that being placed in lower ability groups has a negative effect on children's behavior (Henry, 2015); nonetheless, it is suggested that such findings might be an artifact of the difficulties that teachers face in terms of facilitating good quality teaching in groups predominantly comprised of children whom they perceive to display varying forms of negative behavior (Hallam & Parsons, 2013a). In the absence of experimental studies making ability group allocation truly random, longitudinal studies on the psychological effects of ability-grouping are an improvement over cross-sectional ones, but these studies are scant. One of the few exceptions available followed 9059 Year 7 (11–12 years old) students who were placed in three sets (bottom, middle, top) for English and Maths to the end of Year 8 (12–13 years old) (Francis et al., 2020). It was found that the gap in self-confidence between students in the top and bottom sets for mathematics widened significantly over time. Importantly, this finding survived adjustment for prior academic attainment suggesting not only that the effect of being placed in lower sets itself accumulates over time but that it is also independent of academic progress. An additional study followed 734 seventh-grade students in four different streams over a period of 1 year in Switzerland and showed that those in the lower streams did worse in terms of their adjustment, antisocial behavior, and emotional distress, even after controlling for several confounders including ethnicity, sex, socioeconomic status, and parental involvement (Müller & Hofmann, 2016). Again, however, these results might reflect either negative peer-influences of children with problem behavior who are grouped together (Müller & Zurbriggen, 2016), or, simply, a general cognitive and psychopathological vulnerability of pupils in low ability groups.

Another question that does remain unanswered is what happens to emotional and behavioral developmental trajectories of pupils of “fluid” or “uncertain” ability-group status (e.g., those unstreamed but in different sets and within-class ability-groups by subject). In the absence of research on this to date, we can only theorize that for such pupils the role of ability-grouping is going to be complex, as their immediate frame of reference can vary substantively. For example, a pupil can be in a bottom stream but, at the same time, in the top within-class ability group.

Ability-grouping and educational inequalities

An even bigger concern, according to some, is that ability-grouping probably increases educational inequalities (Hanushek & Wößmann, 2006; Taylor et al., 2018). Black children and children from lower socioeconomic backgrounds, single-parent families, and with less educated parents are typically over-represented in lower sets and streams (Hallam & Parsons, 2013a; Hartas, 2017; Moller & Stearns, 2012; Muijs & Dunne, 2010). In turn, such group allocation can have long-term implications. Moller and Stearns (2012) showed that ability-grouping at school is differentially associated with income levels in adulthood, independently of the quantity of education received.

The extant literature has discussed at least three routes via which ability-grouping can contribute to or amplify educational inequalities. First, the impact of labeling on pupils’ self-confidence can act as a self-fulfilling prophecy for the low attainers who, once placed in a low attainment group, show poorer progress compared to those in higher ability groups (Francis et al., 2017). Second, the quality of teaching offered might differ by stream placement or set group, with pupils in lower groups experiencing a poorer quality of teaching and hence showing less academic progress (Kutnick et al., 2005). Finally, ability groups can influence the teacher's judgments of pupils (Campbell, 2017; Hartas, 2017). Campbell (2017), for example, found consistent relations between the assigned stream and subsequent teachers' perceptions of the child's academic ability and attainment which survived adjustments for the child's actual performance.

The present study

In summary, *within-school* ability-grouping appears to be related to children's academic and psychological outcomes, although results are mixed. Research on the role of *within-class* ability-grouping in academic outcomes is scant, and that in psychological outcomes, such as emotional and behavioral problems, non-existent. As far as we know, there is also no research on the impact of either within-school or within-class ability-grouping on trajectories of emotional and behavioral problems, known to vary substantially between children (Flouri et al., 2018). Most importantly, no study to date has made mutual adjustments for the different types of ability grouping in the U.K. educational system to account for the multiple frames of reference available to children who are concurrently streamed, set by subject, and grouped into in-class ability groups. In this study with a large sample of U.K. children from the general population, we explored the role of all different types of ability-grouping (streaming, setting, and in-class ability-grouping) in primary school in the development of children's emotional and behavioral problems (henceforth "problem behavior") across the primary and secondary school years. Predictions about how streaming, setting, and in-class ability-grouping may each impact on problem behavior were difficult to make. Problem behavior is associated with both self-concept and academic performance (Deighton et al., 2018; E. J. Lee & Stone, 2012), which, as discussed, appear to be related not indifferently to ability-grouping. We expected, however, in line with the existing evidence on the role of ability-grouping in psychological outcomes, that ability-grouping would have negative emotional and behavioral effects in the low ability groups. In turn, the congregation of children with behavioral problems would exacerbate these problems in the lower ability groups, as contagion theories would predict (Dishion et al., 1999; Hanish et al., 2005).

METHOD

Sample

The data for this study came from the first six sweeps of the U.K. MCS, an ongoing multidisciplinary population-based cohort study following children born between September 1, 2000 and August 31, 2001 (for England and Wales), and between November 24, 2000 and January 11, 2002 (for Scotland and Northern Ireland) (Joshi & Fitzsimons, 2016). MCS is the most recent of the United Kingdom's world-renowned longitudinal birth cohort studies. The children were around 9 months old at Sweep 1, and 3, 5, 7, 11, and 14 years old at Sweeps 2, 3, 4, 5, and 6, respectively. At the six sweeps, the numbers of productive families were 18,522, 15,590, 15,246, 13,857, 13,287, and 11,714, respectively. Ethical approval was gained from NHS Multi-Centre Ethics Committees, and parents (and children after age 11 years) gave informed consent before interviews took place.

Data collected for the MCS come from various sources, including the children themselves, main and second co-resident parents and older siblings. When the children were aged seven information was also collected from their class teachers in all four U.K. countries using a self-completion postal questionnaire. In total, 7235 teachers in 4969 schools were contacted to take part in the survey. Of those, 5364 teachers (74.1%) from 3981 schools (80.1%) completed and returned a questionnaire for 8876 children. The sample in England was 5621 children in 2731 schools, for Wales 1204 children in 434 schools, for Scotland 1099 children in 472 schools, and for Northern Ireland 951 children in 348 schools. Approximately two-thirds of teacher questionnaires had complete data. The proportion of questionnaires with missing data was comparable across countries (39%, 32%, 26%, and 33% in Wales, Northern Ireland, England, and Scotland, respectively). Our analytic sample included children (singletons and first-born twins or triplets) with available information on between- and within-class ability-grouping (n = 7767) provided by teachers. Of those, we excluded from further analyses children who were not at the standard academic stage (Year 2 [Primary 3 for Scotland]) at the time of the age 7 follow-up (N = 450) and also those attending special schools at age 7 (N = 40) or age 11 (N = 37), thus achieving a total analytic sample of 7259 children (82% of children with available information in the teacher survey). The flow chart illustrates in detail the process followed to derive the analytic sample and the associated attrition rate (Figure 1). Ethical approval for the teacher survey was given to by the Northern and Yorkshire Multi-Centre Research Ethics Committee (MREC) of the NHS (Huang & Gatenby, 2010). Further approvals were obtained for carrying out the survey from education authorities: For England, from the Star Chamber in the Department for Children, Schools, and Families; for Wales, from the Schools Workforce Advisory Panel; for Scotland, from the Directors of Education in the Local Educational Authorities. For Northern Ireland, no formal approval was needed.

Problem behavior

Peer problems, emotional symptoms, hyperactivity/inattention, and conduct problems were assessed in MCS at ages 3, 5, 7, 11, and 14, using the parent-reported Strengths and Difficulties Questionnaire (SDQ). The SDQ is a short, psychometrically valid, and widely used

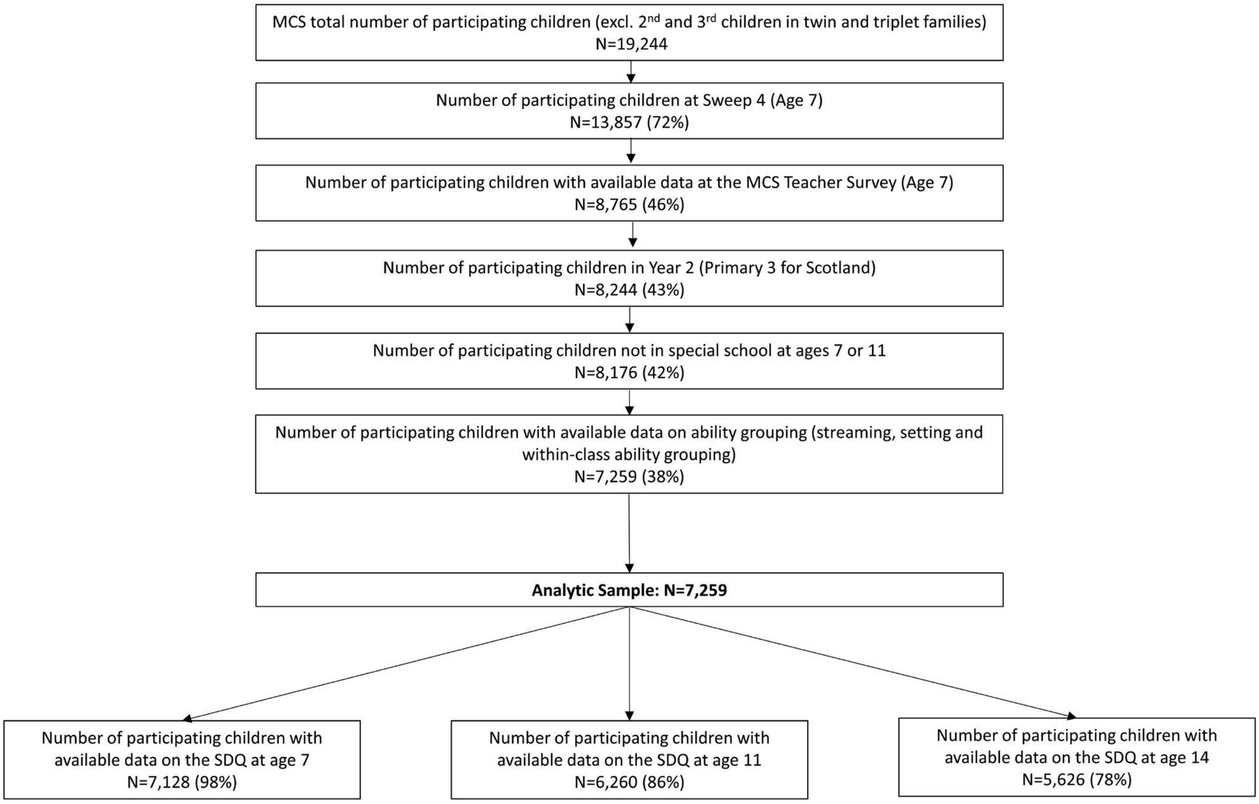
MCS total number of participating children (excl. 2nd and 3rd children in twin and triplet families)
N=19,244
Number of participating children at Sweep 4 (Age 7)
N=13,857 (72%)
Number of participating children with available data at the MCS Teacher Survey (Age 7)
N=8,765 (46%)
Number of participating children in Year 2 (Primary 3 for Scotland)
N=8,244 (43%)
Number of participating children not in special school at ages 7 or 11
N=8,176 (42%)
Number of participating children with available data on ability grouping (streaming, setting and within-class ability grouping)
N=7,259 (38%)
Analytic Sample: N=7,259
Number of participating children with available data on the SDQ at age 7
N=7,128 (98%)
Number of participating children with available data on the SDQ at age 11
N=6,260 (86%)
Number of participating children with available data on the SDQ at age 14
N=5,626 (78%)

FIGURE 1 Flow chart of the study

behavioral screening tool (Goodman, 1997). Each of the four types of problems is measured using five items scored on a 3-point Likert scale (not true, somewhat true, certainly true). In the analytic sample the internal consistency (Cronbach's alpha) of the scales was satisfactory ranging from .58 (peer and conduct problems) to .78 (hyperactivity/inattention) at age 7, .61 (conduct problems) to .79 (hyperactivity/inattention) at age 11 and .62 (peer problems) to .77 (hyperactivity/inattention) at age 14. To ensure adequate fit of this factorial structure of the SDQ in our analytic sample, we ran confirmatory factor analyses (CFA). The results are in the online Supplementary Material.

Ability groups

The age 7 sweep was the first time that MCS asked teachers about between-class (streaming and setting for literacy and maths) and the only time they were asked about within-class ability-grouping. Schools varied in the number of ability groups they had. Across schools that grouped, the number ranged from (a) 2–10 streams; (b) 2–14 literacy sets; (c) 2–10 maths sets; and (d) 2–13 within-class ability groups, with most schools having between two and five ability groups. The grouping variables that we used for our analyses are the teacher's reports of each child's allocation, if applicable, to top, middle, or low ability groups. This information was provided for all three types of ability-grouping that we considered in this study.

Covariates

We controlled for several individual, family, and school characteristics at baseline (age 7, unless otherwise specified). These included *sex*, *age in months*, *season born*, presence (or not) of *long-standing illness*, parent-reported *special educational needs* status, *ethnicity* (White, Indian, Pakistani/Bangladeshi, Black, Mixed, and Other), and *pubertal status* (at age 11). Pubertal status (some signs of puberty vs. no signs) was measured using parental reports of whether there was breast growth, menstruation or hair on body (for females) and, voice change, facial hair, or hair on body (for males). Family characteristics included *maternal education* (university degree or not), *socioeconomic disadvantage* (measured on a 4-item summative index comprising overcrowding [>1.5 people per room excluding bathroom and kitchen], lack of home ownership, receipt of income support, and income poverty [equivalized net family income below 60% of the national median household income] [Malmberg & Flouri, 2011]) and *family structure* (living or not with both biological parents). School characteristics were *in mixed-year class* or not, *number of classes* in the school year, and *school type* (private vs. state). We also considered whether the child *changed school(s)* between ages 7 and 11 years, or secondary school(s) until age 14 years. *Internalizing (emotional and peer) and externalizing (conduct and hyperactivity) problems at age 5* years (using the SDQ) controlled for early problem behavior, likely preceding ability-grouping. Finally, we controlled for time-varying *cognitive ability* (measured as verbal ability in MCS at ages 7–14). In order to make the ability scores comparable across MCS assessments, we transformed the age-adjusted ability scores into a standardized score with a mean of 100 and a standard deviation of 15 at each assessment. More information on the cognitive ability tests are available in the online Supplementary Material.

Statistical analysis

After examining the baseline individual-, family-, and school-level characteristics of the analytic sample, we ran a series of multilevel linear models (MLMs) to examine the associations of between- and within-class ability-grouping at age 7 with trajectories of internalizing (peer and emotional) and externalizing (hyperactivity and conduct) problems at ages 7, 11, and 14 years. We chose to fit MLMs rather than simple regression models to account for the hierarchical nature of our data by having repeated measures (at ages 7, 11, and 14 years) of internalizing and externalizing problems for children, thus by having occasions (level 1) nested in children (level 2). We did not consider a third level (children nested within schools) in the MLMs because the degree of clustering of MCS children within school was not adequate: In the analytic sample, the majority of schools (63%) provided information for a single MCS child, 16% of schools for two and only 10% of schools for more than three (the average MCS pupil count within a single school was 2.1 in the analytic sample). We accounted for missing data in the outcome variables, covariates, and interaction terms between ability group and age by generating 20 datasets with imputed data using chained equations. Multivariate imputation by chained equations (MICE) creates multiple imputations, as opposed to single imputations, thereby accounting for the statistical uncertainty in the imputations. Values were imputed using linear regressions for continuous variables and using multinomial or ordered logit regressions for categorical variables, as appropriate. Rubin's combination rules (Rubin, 1987) were used to consolidate the obtained individual estimates into a single set of multiply imputed estimates. Missing data rates were low. With the exception of the information on whether the children were in the same primary school between the first and second assessments (18%) or in the same secondary school between the second and third assessments (28%) and pubertal status (20%), the proportion of missing data on the remaining covariates was very low ranging from 0% to 6% (Table 1). Regarding the outcomes, the proportion of missing data for all four main outcomes was very low at baseline (age 7; 2%) and

TABLE 1 Sample characteristics of the analytic sample (*N* = 7254) (unweighted data)

Continuous variables	Mean	Standard deviation	% missing data
Emotional problems			
Age 7	1.48	1.72	2
Age 11	1.77	1.95	14
Age 14	1.92	2.08	22
Peer problems			
Age 7	1.15	1.51	2
Age 11	1.28	1.61	14
Age 14	1.65	1.75	26
Hyperactivity			
Age 7	3.25	2.47	2
Age 11	2.98	2.40	14
Age 14	2.83	2.35	22
Conduct problems			
Age 7	1.32	1.49	2
Age 11	1.30	1.52	14
Age 14	1.32	1.59	22
Age (months)	86.72	2.92	0
Socioeconomic disadvantage	0.75	1.07	1
Verbal ability	112.11	18.00	2
Internalizing and externalizing problems (total difficulties score; age 5)	7.01	4.81	6
Categorical variables	N	%	
Streaming			
Not streamed	6038	83	0
Top stream	539	7	0
Middle stream	397	5	0
Bottom stream	285	4	0
Setting for literacy			
Not set	5100	70	0
Top set	928	13	0
Middle set	740	10	0
Bottom set	491	7	0
Setting for maths			
Not set	4812	66	0
Top set	1101	15	0
Middle set	829	11	0
Bottom set	517	7	0
Within-class ability grouping			
Not grouped	1746	24	0
Top group	2238	31	0
Middle group	2257	31	0
Bottom group	1018	14	0

(Continues)

TABLE 1 (Continued)

Continuous variables	Mean	Standard deviation	% missing data
No special educational needs	5276	73	1
Mother has university degree	1352	19	4
No longstanding illness	5929	82	0
Not in mixed-year class	5379	79	6
Number of classes in child's school year			4
1	2168	31	
2	2909	42	
≥3	1919	27	
Ethnicity			0
White	6322	87	
Mixed	181	2	
Indian	154	2	
Pakistani/Bangladeshi	336	5	
Black	183	3	
Other	82	1	
Female	3600	50	0
Signs of puberty at age 11	3986	69	20
Living with both biological parents	5325	73	0
Season born			0
Autumn	2006	28	
Winter	2000	28	
Spring	1702	23	
Summer	1551	21	
Attended the same school at ages 7 and 11	4823	81	18
Attended the same secondary school up to age 14	4909	94	28
Private school (fee-paying school)	267	4	0

increased to 14% and approximately 22% at ages 11 and 14 respectively, as is expected given attrition rates in longitudinal cohort studies.

MLMs were run separately for each of the four SDQ scales. In the first set of MLMs (Model A), we examined the crude mutually adjusted associations between all types of ability-grouping and the four problem scores. In the next set (Model B), we adjusted for the covariates. In the third (Model C) we further adjusted for internalizing and externalizing problems at age 5. For those outcomes that retained a significant association with ability-grouping in the fully adjusted models, we performed two additional MLMs. For the first, we added to Model C an interaction term between ability group and age in order to examine the effect of ability-grouping on the trajectories of problems. One interaction term with age was created for each of top, middle, and low ability groups and they were all included in the model simultaneously. These interaction terms capture the effects of ability grouping on the slope (the rate of change) of the trajectories of problems and show whether the effect of ability-grouping on internalizing and externalizing problems varies with age (DeLucia & Pitts, 2006). For the second, we added to Model C an additional time-varying covariate, verbal ability at ages 7, 11, and 14 years. The reference group of each type of ability-grouping in all MLMs was children who were not grouped; that is, all regression coefficients represent comparisons of children in low, middle, and high groups against those not grouped for each of the different types of ability grouping. Age was grand-mean centered (127.57 months, i.e., approximately 10.63 years) in all analyses to aid estimation and interpretability of the models. Therefore, the effects of the predictor variables on the intercept of the trajectories reflect mean differences at approximately age 11 years (mainly just before the transition to secondary school). In order to allow for changes in problems across time to vary between children, and for the relation between verbal ability and problem behavior to differ between children, we specified random slopes on age and verbal ability. All MLMs were run taking into account the stratified sample design of MCS by including the MCS strata (England-disadvantaged, Wales-advantaged, Wales-disadvantaged, Scotland-advantaged, Scotland-disadvantaged, Northern Ireland-advantaged, and Northern Ireland-disadvantaged) as dummy variables in the fixed part of the models (England-advantaged was the reference category and hence not included). (Note: We also ran all MLMs on a subsample of children who went to school in England only [55% of the analytic sample], after excluding those from the remaining three countries which were under-represented in the original sample. These results were almost identical to the full sample analysis and, hence, are not presented here). We also used study-specific weights to account for the disproportionate attrition of participants in MCS. These weights correct for bias which may be introduced through disproportionate losses to the sample, through non-response at the first survey and attrition at subsequent waves. In fact, it has been shown that attrition in MCS is not random as children from ethnic minorities (Pakistani, Bangladeshi, and Black groups) and more deprived backgrounds are less likely to participate in future follow-ups (Mostafa, & Ploubidis, 2017). These weights compound the sampling weights with a factor reflecting each productive family's chance of having been lost to the survey. Thus families with characteristics resembling those of many drop-outs are given a bigger attrition weight than those who do not (Hansen et al., 2010). To allow for multiple testing we considered significant values with $p \le .01$. Analyses were run using Stata/SE 15.1 (StataCorp, 2011) and MLwiN 3.02 (Charlton et al., 2018).

RESULTS

The baseline (age 7) characteristics of the 7259 (50% male) children with complete data on between- and within-class ability-grouping at age 7 are summarized in Table 1. Overall, the children in the analytic sample were predominantly White (87%), with no special educational needs (73%), no longstanding illness (82%), and most lived with both biological parents (73%). The majority attended state schools (96%) and were not in a mixed-year class (79%). Few parents reported a change in their child's school between the first and second assessment (10%) or since starting secondary school (6%), suggesting limited between-school mobility. Compared to the rest of the children who participated in MCS sweep 4 (age 7) but were not in the analytic sample ($n = 6598$), those in the analytic sample were not different regarding their sex distribution, but they were more likely to be White, to come from less deprived backgrounds, to have more educated mothers and to have lower internalizing and externalizing problems scores and higher verbal ability scores (all p-values <.01).

Figure S1 in the Supplementary Material illustrates the prevalence of each type of ability-grouping, and their combinations, in the analytic sample. Within-class ability-grouping was more prevalent in the analytic sample (76.0% grouped) compared to streaming (16.8%) and setting (29.7% for literacy and 33.7% for maths). Among the streamed children, the majority were placed in the top stream (7.4% of the total) followed by those in the middle (5.5%) and lower streams (3.9%). A similar pattern was observed for children set for literacy and maths with the majority of children in the sample being in the top set (12.8% and 15.2% for literacy and maths, respectively) followed by those in the middle (10.2% and 11.4% for literacy and maths, respectively) and bottom sets (6.8% and 7.1% for literacy and maths, respectively). For those 76.0% of children placed in within-class ability groups, 30.8% were in the top, 31.1% in the middle, and 14.0% in the bottom group. Most children were put in at least two types of ability-grouping. In fact, only a small minority of children in the sample were placed in one type of grouping ($n = 985$; 13.6%). Of those streamed ($n = 1221$), 1051 (86.1%) were also put in within-class ability groups and 912 (74.7%) were set for maths or literacy. In addition, of the 2675 who were set for maths or literacy, 1968 (73.6%) were also placed in within-class ability groups.

Tables S1–S4 in the supplementary material summarize results of tests comparing mean problem behaviors and sociodemographic characteristics across ability groups. We found that those in lower ability groups were characterized by higher levels of emotional and behavioral problems, had lower mean cognitive ability scores, had less educated mothers and came from poorer families compared to children in higher ability groups and ungrouped children.

Externalizing problems

The intraclass correlation coefficients (ICC) for hyperactivity and conduct problems were .64 and .55, respectively, suggesting that a considerable proportion of variance in both types of externalizing problems can be explained by both within-person change over time but also between-children differences. (Note: ICC is a measure of the proportion of variance in the outcome variable that is explained by the grouping structure of the hierarchical model and is not conceptually related to “school classes” in this context). Table 2 summarizes the results of multilevel models examining the relation of within-class and between-class ability-grouping with externalizing problems (hyperactivity and conduct problems) assessed at ages 7, 11, and 14 years. Before adjusting for covariates (Model A), within-class ability-grouping was significantly associated with both hyperactivity and conduct problems. Specifically, children in the top in-class ability group had fewer hyperactivity and conduct problems at around age 11 compared to non-grouped children, whereas those in the bottom or middle in-class ability group had more. After adjustments for individual and school characteristics (Model B) these associations retained their significance for hyperactivity problems, while only being in the bottom within-class ability group remained significantly associated with increased levels of conduct problems. After further adjustment for externalizing and internalizing problems at age 5 (Model C), the associations between within-class ability-grouping and hyperactivity became weaker, albeit remaining significant. The results of the fully adjusted models also showed that children with no special educational needs and children with university-educated mothers had fewer hyperactivity and conduct problems. Being female, being Black (compared to being White), living with both biological parents and having fewer internalizing and externalizing problems at age 5 were also independently associated with fewer externalizing problems. Finally, being Pakistani/Bangladeshi, staying in the same school between ages 7 and 11 years, and coming from a less disadvantaged socioeconomic background were all significantly associated with fewer conduct problems only.

In light of the significant association of being placed in the bottom within-class ability group and hyperactivity, we ran two additional MLMs (Table S5 in the Supplementary Material). For the first one, we added the interactions of within-class ability groups and age as covariates in Model C (Model D). The results showed that the interaction between being in the bottom group and age was negative and statistically, albeit marginally, significant ($b = -0.003$, $SE = 0.001$, $p = .01$), suggesting that hyperactivity in children in the bottom within-class ability group is reduced at a higher rate compared to that for non-grouped children, as they grow older. Nonetheless, the main effects of all three within-class ability groups remained comparable in magnitude to the ones obtained

TABLE 2 Crude and adjusted unstandardized regression coefficients of multilevel models examining the relationship of within- and between-class ability grouping with externalizing problem trajectories at ages 7–14

	Conduct problems			**Hyperactivity**		
	Model A	**Model B**	**Model C**	**Model A**	**Model B**	**Model C**
Fixed effects	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**
Streaming						
Not streamed	Ref	Ref	Ref	Ref	Ref	Ref
Top stream	0.05 (0.06)	0.06 (0.05)	0.02 (0.05)	0.01 (0.09)	0.01 (0.09)	−0.07 (0.07)
Middle stream	0.04 (0.07)	0.06 (0.07)	0.05 (0.06)	0.01 (0.11)	0.08 (0.11)	0.05 (0.09)
Bottom stream	0.10 (0.11)	0.02 (0.10)	0.02 (0.09)	0.12 (0.16)	−0.00 (0.15)	−0.01 (0.12)
Setting for literacy						
Not set	Ref	Ref	Ref	Ref	Ref	Ref
Top set	−0.07 (0.06)	−0.08 (0.06)	−0.10 (0.05)	−0.08 (0.09)	−0.07 (0.09)	−0.09 (0.08)
Middle set	0.16 (0.07)	0.11 (0.07)	0.08 (0.06)	0.19 (0.10)	0.11 (0.10)	0.05 (0.09)
Bottom set	0.19 (0.09)	−0.01 (0.09)	−0.08 (0.07)	0.54 (0.14)*	0.14 (0.13)	0.02 (0.11)
Setting for maths						
Not set	Ref	Ref	Ref	Ref	Ref	Ref
Top set	−0.05 (0.06)	0.00 (0.06)	0.08 (0.05)	−0.19 (0.09)	−0.15 (0.09)	−0.02 (0.07)
Middle set	−0.04 (0.06)	0.02 (0.06)	0.04 (0.05)	−0.06 (0.10)	0.05 (0.10)	0.09 (0.08)
Bottom set	0.09 (0.09)	0.11 (0.08)	0.08 (0.07)	0.14 (0.13)	0.20 (0.13)	0.15 (0.11)
Within-class ability grouping						
Not grouped	Ref	Ref	Ref	Ref	Ref	Ref
Top group	−0.15 (0.04)*	−0.05 (0.04)	−0.04 (0.03)	−0.47 (0.06)*	−0.26 (0.06)*	−0.24 (0.05)*
Middle group	0.11 (0.04)*	0.06 (0.04)	−0.01 (0.03)	0.42 (0.07)*	0.35 (0.06)*	0.22 (0.05)*
Bottom group	0.64 (0.06)*	0.29 (0.06)*	0.11 (0.05)	1.51 (0.09)*	0.82 (0.09)*	0.51 (0.08)*
Age in months (centered at age 11 years)	—	0.00 (0.00)	0.00 (0.00)	—	−0.00 (0.00)*	−0.00 (0.00)*
No special educational needs	—	−0.40 (0.04)*	−0.21 (0.04)*	—	−0.88 (0.06)*	−0.55 (0.05)*
Mother has university degree	—	−0.19 (0.03)*	−0.09 (0.03)*	—	−0.41 (0.06)*	−0.22 (0.05)*
Socioeconomic disadvantage	—	0.20 (0.02)*	0.11 (0.02)*	—	0.18 (0.03)*	0.03 (0.02)
No longstanding illness	—	−0.16 (0.04)*	−0.05 (0.03)	—	−0.30 (0.06)*	−0.10 (0.05)
Not in mixed-year class	—	−0.03 (0.03)	−0.03 (0.03)	—	−0.03 (0.06)	−0.02 (0.05)
Number of classes in child's school year	—	−0.03 (0.02)	−0.03 (0.02)	—	−0.02 (0.03)	−0.00 (0.03)
Ethnicity						
White		Ref	Ref		Ref	
Mixed	–	−0.09 (0.09)	−0.13 (0.08)		−0.04 (0.14)	−0.09 (0.12)
Indian		−0.03 (0.09)	−0.13 (0.08)	—	−0.07 (0.15)	−0.24 (0.13)
Pakistani/Bangladeshi		−0.04 (0.08)	−0.21 (0.07)*		0.03 (0.12)	−0.26 (0.11)
Black		−0.40 (0.09)*	−0.34 (0.09)*		−0.54 (0.15)*	−0.45 (0.13)*
Other		0.06 (0.13)	−0.13 (0.13)		−0.14 (0.18)	−0.46 (0.18)*
Female	—	−0.20 (0.04)*	−0.12 (0.04)*	—	−0.70 (0.06)*	−0.56 (0.05)*
Signs of puberty at age 11		0.08 (0.05)	0.08 (0.05)		0.00 (0.07)	−0.00 (0.06)
Living with both biological parents	—	−0.21 (0.03)*	−0.14 (0.03)*	—	−0.30 (0.06)*	−0.18 (0.05)*
Season born						
Autumn		Ref	Ref		Ref	Ref
Winter	—	−0.06 (0.04)	−0.06 (0.03)	—	−0.07 (0.06)	−0.08 (0.05)

(Continues)

TABLE 2 (Continued)

	Conduct problems			Hyperactivity		
	Model A	Model B	Model C	Model A	Model B	Model C
Fixed effects	Coeff. (SE)	Coeff. (SE)	Coeff. (SE)	Coeff. (SE)	Coeff. (SE)	Coeff. (SE)
Spring		0.01 (0.04)	0.01 (0.03)		−0.05 (0.06)	−0.04 (0.05)
Summer		0.03 (0.04)	0.03 (0.04)	—	0.00 (0.06)	0.00 (0.05)
Attended the same school at ages 7 and 11	—	−0.14 (0.04)*	−0.11 (0.03)*	—	−0.09 (0.06)	−0.04 (0.05)
Attended the same secondary school up to age 14	—	−0.21 (0.09)	−0.13 (0.08)	—	−0.28 (0.09)*	−0.15 (0.09)
Private school (fee-paying school)	—	−0.18 (0.07)*	−0.08 (0.06)	—	−0.15 (0.11)	0.04 (0.10)
Internalizing and externalizing problems at age 5	—	—	0.13 (0.00)*	—	—	0.22 (0.00)*
Constant	1.10 (0.04)*	2.13 (0.11)*	0.94 (0.11)*	2.67 (0.06)*	4.65 (0.16)*	2.59 (0.14)*
Random effects						
Level 2 (child) intercept variance (SE)	1.00 (0.03)*	0.86 (0.03)*	0.56 (0.02)*	2.70 (0.07)*	2.30 (0.06)*	1.37 (0.05)*
Slope variance (SE)	—	0.00 (0.00)*	0.00 (0.00)*	—	0.00 (0.00)*	0.00 (0.00)*
Covariance (SE)	—	0.00 (0.00)	0.00 (0.00)*	—	−0.00 (0.00)*	−0.00 (0.00)
Level 1 (occasion) intercept variance (SE)	1.24 (0.02)*	1.14 (0.04)*	1.14 (0.04)*	2.58 (0.04)*	2.33 (0.07)*	2.34 (0.07)*

All models were adjusted for the stratified design of MCS (regression coefficients of the strata are not shown in the table for parsimony).
Variables measure baseline (age 7) characteristics unless otherwise specified.
*$p \leq .01$.

in Model C and statistically significant even after including the interactions in the model suggesting that children in the bottom and middle within-class ability groups were more hyperactive at around age 11, and those in the top group less, compared to the non-grouped children, regardless of the slope of the trajectory of hyperactivity. Figure 2 illustrates the predicted hyperactivity scores over time in this model, stratified by within-class ability group. A final MLM was run using verbal ability at ages 7, 11, and 14 years as an additional covariate to Model C (Model E; results summarized in Table S5 in the Supplementary Material). As expected, verbal ability was negatively associated with hyperactivity ($b = -0.01$, $SE = 0.001$, $p < .001$). In this model too, all three within-class ability groups retained their significant main effects and were not significantly different in their magnitude compared to the ones found in Model C.

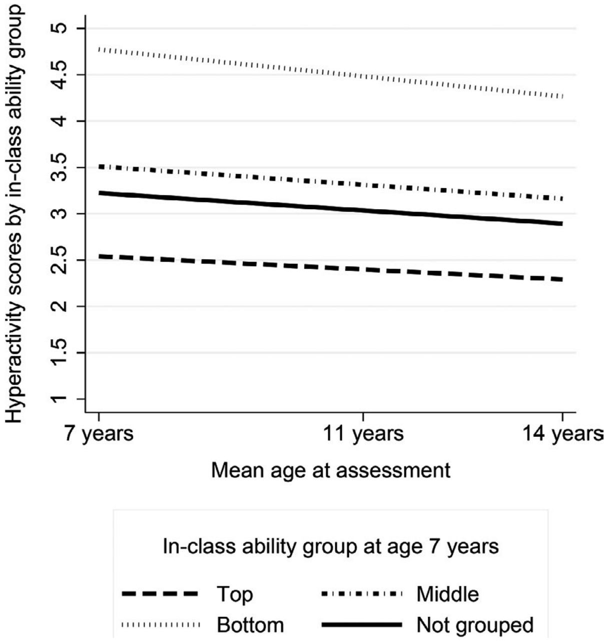

FIGURE 2 Predicted hyperactivity scores at ages 7, 11 and 14 years by within-class ability group at age 7 years

Internalizing problems

The ICC for both emotional and peer problems was .47, suggesting that a considerable proportion of variance in both types of internalizing problems can be explained by both within-person change over time but also between-children differences. Table 3 summarizes the results of MLMs examining the relation of between- and within-class ability-grouping with internalizing (peer and emotional) problems. Model A (prior to adjustments for covariates) showed that those in the bottom set for literacy had higher levels of peer problems at around age 11 compared to the non-set children. Additionally, children in the middle and bottom within-class ability groups had on average higher levels of both types of internalizing problems. The association between being in the bottom within-class ability group and emotional problems survived adjustments for the covariates (Model B) including internalizing and externalizing problems at age 5 (Model C). The regression coefficients for the covariates in Model C suggest that having special educational needs or a longstanding illness were both independently associated with higher levels of emotional and peer problems, as were attending a state school, changing secondary school by age 14 and having higher levels of internalizing and externalizing problems at age 5. Girls and spring- or winter-born (compared to autumn-born) children showed increased levels of emotional problems only. Being Indian, was also associated with fewer emotional problems compared to being White. Finally, children from a socioeconomically disadvantaged background and those not living with both biological parents scored higher on peer problems only.

Using the same rationale followed for hyperactivity in the previous section, we fitted two additional MLMs to explore further the significant association between being placed in the bottom within-class ability group and emotional problems. First, we re-ran Model C while also including the interactions between within-class ability groups and age as covariates (Model D; results summarized in Table S5 in the Supplementary Material). None of the interactions were significant while the main effect of being in the bottom within-class ability group on emotional problems retained its significance level ($b = 0.26$, $SE = 0.06$, $p < .001$), suggesting that the level, but not the rate of change, of emotional problems differs for children placed in the bottom within-class ability group compared to non-grouped children. Moreover, the main effects of ability-grouping were similar in magnitude compared to the ones obtained in Model C. Figure 3 illustrates the predicted scores of emotional problems over time from this model stratified by within-class ability group. In the second MLM we re-ran Model C while further adjusting for children's verbal ability scores at ages 7 to 14 (Model E; results summarized in Table S5 in the Supplementary Material). Verbal ability was not significantly associated with emotional problems while the main effect of being in the bottom within-class ability group remained significant and comparable in magnitude to the one obtained for Model C.

Supplementary analysis

We also sought to examine the associations between the individual types of ability-grouping and internalizing and externalizing problems. We, therefore, ran additional MLMs (adjusted for covariates as in Model C) for each of streaming, setting, and within-class ability-grouping separately. The results of this supplementary analysis are presented in the Supplementary Material (Section 6). Overall, the results showed that children in the bottom within-class ability groups had higher levels of emotional problems at around age 11, whereas those in the bottom within-class ability group had additionally increased levels of peer problems. Being placed in the middle and bottom streams, middle and bottom within-class ability groups, or middle and bottom sets for maths or literacy was associated with higher levels of hyperactivity. In contrast, children in the top stream, top within-class group, top set for literacy, or top maths set were less hyperactive at around age 11 years. Finally, being placed in the bottom within-class ability group or middle set for literacy was associated with increased levels of conduct problems.

In a further sensitivity analysis, we aimed to test whether the disadvantageous effect of being placed in a bottom within-class ability group on emotional problems and hyperactivity is an artifact of the poorer average academic performance of children in these groups compared to the ungrouped children who served as the reference group in the MLMs presented above. Academic performance was measured using the children's average key stage 1 (KS1) scores, collected during the January 2008 census and obtained from the National Pupil Database. Information on academic performance was available for children attending state schools in England only ($n = 3829$). We restricted the sample of these analyses to those children with KS1 scores in the lower tertile of the distribution ($n = 1538$ of whom 529 [34%] were placed in a low within-class ability group). We ran two MLMs, one for each of emotional problems and hyperactivity, adjusted for sex and other types of grouping. The results showed that children in low within-class ability groups had significantly increased levels of emotional problems ($b = 0.50$, $SE = 0.13$, $p < .001$) and hyperactivity ($b = 0.52$, $SE = 0.16$, $p = .001$) at around age 11 years compared to their ungrouped academically low performing counterparts. Nonetheless, these findings must be interpreted with caution as these analyses were somewhat underpowered and apply to students in state schools in England only.

DISCUSSION

The results of this study suggest that U.K. children who are placed in the bottom within-class ability group in primary school show increased levels of emotional problems and hyperactivity, but not conduct or peer problems, into their secondary school years. This pattern of findings persisted after adjustments for several individual and family characteristics known to be associated with ability-group allocation and emotional and behavioral problems, as well as after controls for school characteristics. Moreover, it survived adjustments for

TABLE 3 Crude and adjusted unstandardized regression coefficients of multilevel models examining the relationship of within- and between-class ability grouping with internalizing problem trajectories at ages 7–14

	Peer problems			Emotional problems		
	Model A	Model B	Model C	Model A	Model B	Model C
Fixed effects	Coeff. (SE)	Coeff. (SE)	Coeff. (SE)	Coeff. (SE)	Coeff. (SE)	Coeff. (SE)
Streaming						
Not streamed	Ref	Ref	Ref	Ref	Ref	Ref
Top stream	−0.02 (0.06)	−0.01 (0.05)	−0.05 (0.05)	−0.04 (0.07)	0.00 (0.07)	−0.05 (0.06)
Middle stream	0.06 (0.08)	0.08 (0.07)	0.07 (0.07)	0.02 (0.09)	0.01 (0.08)	−0.00 (0.07)
Bottom stream	0.10 (0.10)	0.03 (0.10)	0.03 (0.09)	0.06 (0.12)	0.01 (0.12)	0.01 (0.10)
Setting for literacy						
Not set	Ref	Ref	Ref	Ref	Ref	Ref
Top set	−0.03 (0.06)	−0.04 (0.06)	−0.05 (0.05)	−0.02 (0.07)	−0.05 (0.07)	−0.07 (0.06)
Middle set	0.01 (0.07)	−0.02 (0.06)	−0.05 (0.06)	0.08 (0.08)	0.06 (0.08)	0.02 (0.07)
Bottom set	0.26 (0.09)*	0.07 (0.09)	0.01 (0.08)	0.09 (0.11)	−0.02 (0.10)	−0.10 (0.09)
Setting for maths						
Not set	Ref	Ref	Ref	Ref	Ref	Ref
Top set	−0.07 (0.06)	−0.02 (0.05)	0.04 (0.05)	−0.04 (0.07)	0.04 (0.07)	0.12 (0.06)
Middle set	−0.04 (0.06)	−0.00 (0.06)	0.01 (0.06)	0.09 (0.08)	0.10 (0.08)	0.12 (0.07)
Bottom set	0.01 (0.09)	0.01 (0.08)	−0.02 (0.07)	0.20 (0.10)	0.14 (0.10)	0.10 (0.09)
Within-class ability grouping						
Not grouped	Ref	Ref	Ref	Ref	Ref	Ref
Top group	−0.09 (0.04)	0.01 (0.04)	0.02 (0.03)	−0.08 (0.05)	−0.01 (0.05)	0.00 (0.04)
Middle group	0.12 (0.04)*	0.08 (0.04)	0.02 (0.04)	0.17 (0.05)*	0.14 (0.05)*	0.06 (0.04)
Bottom group	0.63 (0.06)*	0.25 (0.06)*	0.09 (0.06)	0.76 (0.07)*	0.47 (0.07)*	0.26 (0.06)*
Age in months (centered at age 11 years)	—	0.01 (0.00)*	0.01 (0.00)*	—	0.01 (0.00)*	0.01 (0.00)*
No special educational needs	—	−0.54 (0.04)*	−0.37 (0.04)*	—	−0.45 (0.05)*	−0.22 (0.04)*
Mother has university degree	—	−0.14 (0.04)*	−0.04 (0.03)	—	−0.14 (0.04)*	−0.00 (0.04)
Socioeconomic disadvantage	—	0.14 (0.02)*	0.06 (0.02)*	—	0.11 (0.02)*	0.01 (0.02)
No longstanding illness	—	−0.24 (0.04)*	−0.14 (0.04)*	—	−0.43 (0.05)*	−0.30 (0.04)*
Not in mixed-year class	—	−0.02 (0.04)	−0.02 (0.03)	—	0.03 (0.04)	0.03 (0.04)
Number of classes in child's school year	—	−0.03 (0.02)	−0.02 (0.02)	—	−0.02 (0.02)	−0.01 (0.02)
Ethnicity						
White		Ref	Ref		Ref	Ref
Mixed	—	−0.09 (0.08)	−0.11 (0.08)	—	−0.04 (0.11)	−0.08 (0.10)
Indian		0.23 (0.10)	0.15 (0.09)		−0.16 (0.12)	−0.28 (0.11)*
Pakistani/Bangladeshi		0.28 (0.08)*	0.13 (0.08)		0.19 (0.10)	−0.01 (0.09)
Black		−0.03 (0.10)	0.02 (0.09)		−0.23 (0.12)	−0.16 (0.11)
Other		0.31 (0.13)	0.14 (0.11)		0.12 (0.17)	−0.10 (0.15)
Female	—	−0.12 (0.04)*	−0.05 (0.04)	—	0.32 (0.05)*	0.41 (0.04)*
Signs of puberty at age 11		0.10 (0.05)	0.10 (0.05)		0.08 (0.06)	0.07 (0.05)
Living with both biological parents	—	−0.19 (0.04)*	−0.13 (0.04)*	—	−0.17 (0.05)*	−0.09 (0.04)
Season born						

(Continues)

TABLE 3 (Continued)

	Peer problems			Emotional problems		
	Model A	Model B	Model C	Model A	Model B	Model C
Fixed effects	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**
Autumn		Ref	Ref		Ref	Ref
Winter	—	0.04 (0.04)	0.04 (0.03)	—	0.15 (0.05)*	0.15 (0.04)*
Spring		0.05 (0.04)	0.06 (0.04)		0.11 (0.05)	0.11 (0.04)*
Summer		0.01 (0.04)	0.01 (0.04)		0.07 (0.05)	0.07 (0.04)
Attended the same school at ages 7 and 11	—	−0.11 (0.04)	−0.08 (0.04)	—	−0.02 (0.05)	0.02 (0.04)
Attended the same secondary school up to age 14	—	−0.34 (0.09)*	−0.27 (0.09)*	—	−0.39 (0.09)*	−0.29 (0.09)*
Private school (fee-paying school)	—	−0.31 (0.06)*	−0.22 (0.06)*	—	−0.35 (0.08)*	−0.23 (0.07)*
Internalizing and externalizing problems at age 5	—	—	0.11 (0.00)*	—	—	0.15 (0.00)*
Constant	1.11 (0.04)*	2.30 (0.3)*	1.24 (0.12)*	1.44 (0.05)*	2.40 (0.13)*	0.98 (0.13)*
Random effects						
Level 2 (child) intercept variance (SE)	0.95 (0.04)*	0.82 (0.03)*	0.58 (0.03)*	1.37 (0.05)*	1.19 (0.04)*	0.77 (0.03)*
Slope variance (SE)	—	0.00 (0.00)	0.00 (0.00)*	—	0.00 (0.00)	0.00 (0.00)
Covariance (SE)	—	0.00 (0.00)*	0.00 (0.00)*	—	0.00 (0.00)*	0.01 (0.00)*
Level 1 (occasion) intercept variance (SE)	1.59 (0.03)*	1.44 (0.04)*	1.44 (0.04)*	2.20 (0.04)*	2.09 (0.06)*	2.09 (0.06)*

All models were adjusted for the stratified design of MCS (regression coefficients of the strata are not shown in the table for parsimony).
Variables measure baseline (age 7) characteristics unless otherwise specified.
*$p \leq .01$.

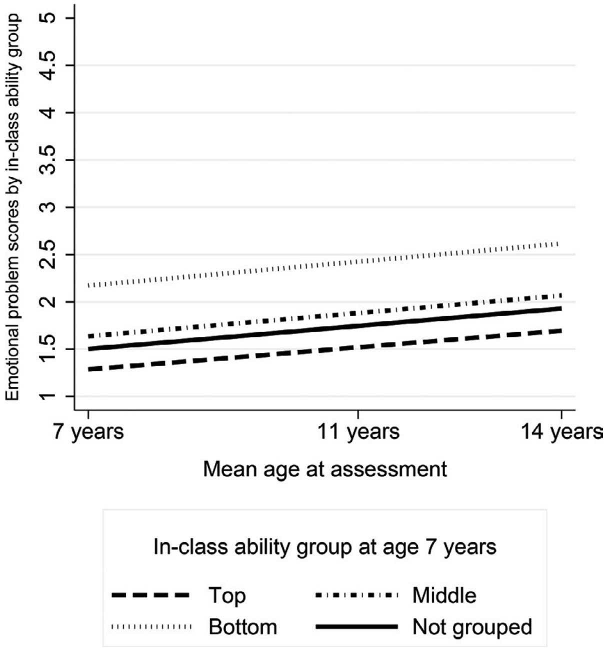

FIGURE 3 Predicted emotional problem scores at ages 7, 11 and 14 years by within-class ability group at age 7 years

other types of ability-grouping (i.e., streaming and setting) and children's earlier emotional and behavioral problems. Increased levels of hyperactivity were also observed for children placed in the middle or bottom sets for maths or literacy. However, these associations did not survive adjustments for the other types of ability-grouping, that is, streaming and within-class grouping. Interestingly, hyperactivity over time in children in the bottom within-class ability group was reduced at a significantly higher rate compared to that of non-grouped children. Nonetheless, their average levels of hyperactivity remained significantly elevated by more than ½ standard deviation compared to those of non-grouped children. In contrast to what some previous studies with gifted pupils have proposed (Becker et al., 2014; S.-Y. Lee et al., 2012), there was no evidence for psychosocial disadvantages for those in the top ability groups.

Our findings about the increased emotional and behavioral problems of children placed in low within-class ability groups highlight an important challenge for the use and implementation of ability-grouping. Whether the academic benefits of within-class ability-grouping reported by some outweigh its shortcomings should be a priority for future research. To date, very little is known about the learning dynamics, peer processes, and subtle effects of in-class ability-grouping (Wilkinson & Penney, 2014), particularly in classes with extensive selective grouping. However, if the associations found in this study are causal, they suggest that children in the lower within-class ability groups require close monitoring and support by their teachers to ensure that their behavioral and emotional development is not compromised.

While the causal mechanisms linking within-class ability-grouping with emotional and behavioral problems have yet to be explored, a plausible explanation for the findings we observe in our study may be social comparison. According to social comparison theory (Festinger, 1954; Huguet et al., 2009; Zell & Alicke, 2009b), we evaluate ourselves by comparing ourselves to our in-group peers. Since pupils placed in within-class ability groups are all members of the same group (the class), it is likely that those in the lower-ability groups feel inferior because any in-group comparison would be unfavorable for them. Previous studies on the effects of between-class ability-grouping suggest that streaming results in negative self-feelings for the lower-stream pupils (Van Houtte & Stevens, 2009). Poor self-image and low self-esteem are known correlates of behavioral problems (Kellison et al., 2010) and depressive symptoms, and thus could be contributing to the high levels of hyperactivity and emotional symptoms we observed for children placed in the low within-class ability group. Our study also showed negative behavioral outcomes associated with being in the bottom set when within-class grouping was not taken into account. But importantly it showed that being placed in the bottom group *within class* is apparently the most damaging context, emotionally and behaviorally, in line with the "local dominance effect" that suggests that the more proximal the unfavorable comparison is, the more powerful its impact (Alicke et al., 2010; Zell & Alicke, 2009a).

We did not find evidence that pupils in higher ability groups suffered with respect to their emotional and behavioral well-being. The BFLPE predicts that these pupils have a lower academic concept (arguably associated with worse emotional and behavioral outcomes), also because of unfavorable proximal comparisons (e.g., Marsh et al., 2000; Trautwein et al., 2009). Rather than contradicting the BFLPE, our findings suggest that social comparison may also explain our findings about greater emotional and behavioral problems in the bottom within-class ability group. However, as we indicated above, we also think that another social process—in some ways related, but for the most part unique to the children in the bottom *within-class* group —may be at play: stigmatization. We argue that the contained and small-scale physical and social context of the classroom makes these children particularly visible. In turn, the more visible a stigmatizing condition, the greater its (negative) impact on the individual (Kurzban & Leary, 2001). Therefore, these children are likely more stigmatized than those belonging to any other ability group. The data available in MCS do not allow us to test these hypotheses but it will be important for future research to explore and compare the social mechanisms at play across all levels and types of ability-grouping we explored in our study.

While the results found in this study are associative, it appears that inclusive whole-school learning cultures should be re-inforced as they can instill greater fluidity in student identities across attainment groups, in line with what recent literature suggests (Mazenod et al., 2019). A recent qualitative study among teachers (Mazenod et al., 2019) highlighted that teachers themselves recognize the damage to pupils' confidence resulting from being labeled as "low ability" students and the need to compensate for this through pedagogic practices. Assessing students' abilities more regularly and allowing for greater mobility and flexibility in group allocations might help reduce the stigma associated with membership in lower ability groups. Students might also find motivating the belief that if they work hard they will be rewarded by moving up a group. Frequent assessment of students' abilities will also ensure that ability groups reflect attainment levels more accurately. Encouraging a classroom climate where students support one another by using set and mixed ability groups interchangeably might also help cultivate a culture of mutual respect and encourage support among students. Finally, in line with a recent study's recommendation, there is a need for the adoption of more equitable practices in allocating students to different ability groups (Taylor et al., 2018).

Our study has several strengths. We used a large population-based sample of children and adolescents which was constructed to be representative of the total U.K. population (Joshi & Fitzsimons, 2016). The measurement of emotional and behavioral problems also covers a critical developmental period. In addition, we used state-of-the-art methods to impute and analyze the data.

However, several limitations, mainly pertaining to data availability, should also be acknowledged. The analytic sample comprised a somewhat advantaged group of mainly white children with lower than average internalizing and externalizing scores (as measured with the SDQ) and higher verbal ability. We did not have information on several school compositional characteristics that may be important modifiers of ability-grouping "effects," for example, average school attainment or the proportion of ethnic minority children in the school. These are of particular importance because there is some evidence that in low-performing, low socioeconomic status, and high-minority schools' ability-grouping has no effects for low-ability pupils (Nomi, 2009). There are also various individual unobserved characteristics, such as personal preference and motivation, that can contribute to a selection bias toward certain ability groups. Such differences were not captured in our study and thus we cannot claim that the associations observed are causal (Jackson, 2009). We also did not have information on the quality of teaching delivered to the different levels of sets and streams. As mentioned, differential quality of teaching is one of the proposed reasons for the unequal academic progress observed between streams (Kutnick et al., 2005) which might, in turn, affect emotional and behavioral outcomes. Relatedly, pupil allocation to within-class ability groups in the United Kingdom reflects the class teacher's assessment or view, which have been shown to be subject to expectation bias related to pupil characteristics (Meissel et al., 2017). Such “noise” in group allocation can limit the generalizability of our results with respect to the significant effect of within-class ability-grouping we found. Another, related, limitation is that we did not have information on the size of within-class groups. Within-class grouping size varies and can include large groups, small groups, triads, dyads, and even groups of one (Kutnick et al., 2005). Group size in turn is likely to impact on the child's academic progress but also behavior. It is also worth repeating that there is not a large degree of overlap of relative position by ability-grouping and hence a child, for example, might be in the top stream in her school but in the bottom within-class ability group. In such cases, inferences about the mechanisms explaining the role of ability-grouping are difficult to make. Another limitation is that our analyses were run under the assumption that children stay in the same ability group throughout the study period (i.e., ages 7, 11, and 14 years) or that the effects of ability-grouping at age 7 are long-lasting and detectable in later assessments. In MCS, information on ability-grouping was collected also at age 11, but only for England (and only for streaming and setting), and could not, therefore, be included in our analyses. Nonetheless, we calculated the proportion of children who changed the set and/or stream between the two assessments for those residing in England with complete data on ability-grouping at both assessments (N = 3721). We found that only 1% of the children belonged in a different stream and up to 8% in a different set. The most striking difference was that a significantly higher proportion of children were set and/or streamed at age 11 (27% on average). It is thus possible that rather than demonstrating the effect of early ability-grouping in primary school, our analyses show the effect of continuous or later-ability grouping. However, a lack of mobility within sets and streams in the U.K. schools has been reported in previous literature (Blatchford et al., 2008), which supports our finding from the England subsample that pupils’ positions within in-school hierarchies tend to be largely stable over time. If mobility across assessment waves is indeed negligible it would indicate that it is only ability grouping which affects the levels of emotional and behavioral problems and not vice versa since most associations survived adjustments for pre-existing emotional and behavioral problems at age 5. Nonetheless, in the absence of such information, it appears premature to conclude that higher levels of problem behaviors cannot influence group allocation in later years. Future studies should utilize research designs which would allow establishing the directionality of the associations found in this study. It is also worth noting that between-school mobility in our sample was similarly very low—apart from the transition to secondary school—and, as it was also controlled for, it is unlikely to have affected our findings. Finally, we did not consider additional potential confounders. For example, pupils who are not native English speakers have lower levels of academic self-concept and higher levels of emotional and behavioral problems (Van Landeghem et al., 2002) but very few children in MCS are not competent speakers of English by age 7 since they have all been in the United Kingdom since infancy. Classroom composition may be another important confounder (Hornstra et al., 2015) but MCS has limited data on this. Such “level-2” information would allow for a more comprehensive examination of the associations between ability-grouping and socioemotional outcomes (Müller & Zurbriggen, 2016). We recommend that future research takes a more holistic approach.

CONCLUSION

Children placed in the bottom ability groups, particularly within-class, in U.K. primary schools showed higher levels of emotional symptoms and hyperactivity across primary and secondary school years. This association was independent of important school, individual and family characteristics, associated with both ability-group allocation and emotional and behavioral problems. Our study raises caution that placing primary school pupils in low-ability groups (in contrast to not placing them in any group) is associated with an increase in hyperactivity and emotional symptoms. Hence, closer monitoring and support by teachers are needed to ensure that the behavioral and emotional development of low-attaining in-class grouped pupils is not compromised.

ACKNOWLEDGMENTS

This work was supported by grant ES/N007921/1 from the Economic and Social Research Council.

ORCID

Efstathios Papachristou https://orcid.org/0000-0001-5746-9561
Eirini Flouri https://orcid.org/0000-0001-6207-4847

REFERENCES

Alicke, M. D., Zell, E., & Bloom, D. L. (2010). Mere categorization and the frog-pond effect. *Psychological Science*, *21*(2), 174–177. https://doi.org/10.1177/0956797609357718

Becker, M., Neumann, M., Tetzner, J., Böse, S., Knoppick, H., Maaz, K., Baumert, J., & Lehmann, R. (2014). Is early ability grouping good for high-achieving students’ psychosocial development? Effects of the transition into academically selective schools.

Journal of Educational Psychology, *106*(2), 555–568. https://doi.org/10.1037/a0035425

Blatchford, P., Hallam, S., Ireson, J., & Kutnick, P. (2008). Classes, groups and transitions: Structures for teaching and learning. Primary Review Research Survey 9/2. http://cprtrust.org.uk/wp-content/uploads/2014/06/research-survey-9-2.pdf

Boaler, J., & Wiliam, D. (2001). Setting, streaming and mixed-ability teaching. In J. Dillon (Ed.), *Becoming a teacher: Issues in secondary teaching*, 2nd ed. (pp. 173–181). Open University Press.

Boaler, J., Wiliam, D., & Brown, M. (2000). Students' experiences of ability grouping-disaffection, polarisation and the construction of failure. *British Educational Research Journal*, *26*(5), 631–648. https://doi.org/10.1080/713651583

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*, 505–522. https://doi.org/10.18546/LRE.15.3.12

Charlton, C., Rasbash, J., Browne, W., Healy, M., & Cameron, B. (2018). *MLwiN version 3.02*. Centre for Multilevel, Modelling University of Bristol.

Deighton, J., Humphrey, N., Belsky, J., Boehnke, J., Vostanis, P., & Patalay, P. (2018). Longitudinal pathways between mental health difficulties and academic performance during middle childhood and early adolescence. *British Journal of Developmental Psychology*, *36*(1), 110–126. https://doi.org/10.1111/bjdp.12218

DeLucia, C., & Pitts, S. C. (2006). Applications of individual growth curve modeling for pediatric psychology research. *Journal of Pediatric Psychology*, *31*(10), 1002–1023. https://doi.org/10.1093/jpepsy/jsj074

Dicke, T., Marsh, H. W., Parker, P. D., Pekrun, R., Guo, J., & Televantou, I. (2018). Effects of school-average achievement on individual self-concept and achievement: Unmasking phantom effects masquerading as true compositional effects. *Journal of Educational Psychology*, *110*(8), 1112–1126. https://doi.org/10.1037/edu0000259

Dishion, T. J., McCord, J., & Poulin, F. (1999). When interventions harm: Peer groups and problem behavior. *American Psychologist*, *54*, 755–764. https://doi.org/10.1037/0003-066x.54.9.755

Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, *7*(2), 117–140. https://doi.org/10.1177/001872675400700202

Flouri, E., Papachristou, E., Midouhas, E., Joshi, H., Ploubidis, G. B., & Lewis, G. (2018). Early adolescent outcomes of joint developmental trajectories of problem behavior and IQ in childhood. *European Child & Adolescent Psychiatry*, *27*(12), 1595–1605. https://doi.org/10.1007/s00787-018-1155-7

Francis, B., Connolly, P., Archer, L., Hodgen, J., Mazenod, A., Pepper, D., Sloan, S., Taylor, B., Tereshchenko, A., & Travers, M.-C. (2017). Attainment grouping as self-fulfilling prophesy? A mixed methods exploration of self-confidence and set level among Year 7 students. *International Journal of Educational Research*, *86*, 96–108. https://doi.org/10.1016/j.ijer.2017.09.001

Francis, B., Craig, N., 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

Gamoran, A., & Nystrand, M. (1994). Tracking, instruction and achievement. *International Journal of Educational Research*, *21*(2), 217–231. https://doi.org/10.1016/0883-0355(94)90033-7

Goodman, R. (1997). The strengths and difficulties questionnaire: A research note. *Journal of Child Psychology and Psychiatry*, *38*(5), 581–586. https://doi.org/10.1111/j.1469-7610.1997.tb01545.x

Hallam, S., & Parsons, S. (2013a). The incidence and make up of ability grouped sets in the UK primary school. *Research Papers in Education*, *28*(4), 393–420. https://doi.org/10.1080/02671522.2012.729079

Hallam, S., & Parsons, S. (2013b). 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

Hanish, L. D., Martin, C. L., Fabes, R. A., Leonard, S., & Herzog, M. (2005). Exposure to externalizing peers in early childhood: Homophily and peer contagion processes. *Journal of Abnormal Child Psychology*, *33*(3), 267–281. https://doi.org/10.1007/s10802-005-3564-6

Hansen, K., Jones, E., Joshi, H., & Budge, D. (2010). Millennium Cohort Study Fourth Survey: A user's guide to initial findings. Centre for Longitudinal Studies, University of London. https://cls.ucl.ac.uk/cls-studies/millennium-cohort-study/mcs-age-7-sweep/

Hanushek, E. A., & Wößmann, L. (2006). Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries. *The Economic Journal*, *116*, C63–C76. https://doi.org/10.3386/w11124

Harlen, W. (1997). *Setting and streaming: A research review*. Scottish Council for Research in Education.

Hartas, D. (2017). Setting for English and Maths: 11-year-olds' characteristics and teacher perceptions of school attitudes. *Research Papers in Education*, *33*(3), 393–410. https://doi.org/10.1080/02671522.2017.1329338

Henry, L. (2015). The effects of ability grouping on the learning of children from low income homes: A systematic review. *The STeP Journal*, *2*, 79–87.

Hornstra, L., van der Veen, I., Peetsma, T., & Volman, M. (2015). Does classroom composition make a difference: Effects on developments in motivation, sense of classroom belonging, and achievement in upper primary school. *School Effectiveness and School Improvement*, *26*(2), 125–152. https://doi.org/10.1080/09243453.2014.887024

Huang, Y., & Gatenby, R. (2010). *Millennium Cohort Study Sweep 4 teacher survey technical report*. Centre for Longitudinal Studies, Institute of Education.

Huguet, P., Dumas, F., Marsh, H., Régner, I., Wheeler, L., Suls, J., Seaton, M., & Nezlek, J. (2009). Clarifying the role of social comparison in the big-fish–little-pond effect (BFLPE): An integrative study. *Journal of Personality and Social Psychology*, *97*(1), 156–170. https://doi.org/10.1037/a0015558

Ireson, J., & Hallam, S. (1999). Raising standards: Is ability grouping the answer? *Oxford Review of Education*, *25*(3), 343–358. https://doi.org/10.1080/030549899104026

Ireson, J., & Hallam, S. (2001). *Ability grouping in education*. Paul Chapman Publishing, A SAGE Publications Company.

Ireson, J., & Hallam, S. (2005). Pupils' liking for school: Ability grouping, self-concept and perceptions of teaching. *British Journal of Educational Psychology*, *75*(2), 297–311. https://doi.org/10.1348/000709904X24762

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

Jackson, C. K. (2009). Ability-grouping and academic inequality: Evidence from rule-based student assignments. *National Bureau of Economic Research Working Paper*. https://doi.org/10.3386/w14911

Joshi, H., & Fitzsimons, E. (2016). The UK Millennium Cohort: The making of a multipurpose resource for social science and policy. *Longitudinal and Life Course Studies*, *7*, 409–430. https://doi.org/10.14301/llcs.v7i4.410

Kellison, I., Bussing, R., Bell, L., & Garvan, C. (2010). Assessment of stigma associated with attention deficit hyperactivity disorder:

Psychometric evaluation of the ADHD stigma questionnaire. *Psychiatry Research*, *178*(2), 363–369. https://doi.org/10.1016/j.psychres.2009.04.022

Kurzban, R., & Leary, M. R. (2001). Evolutionary origins of stigmatization: The functions of social exclusion. *Psychological Bulletin*, *127*(2), 187–208. https://doi.org/10.1037//0033-2909.127.2.187

Kutnick, P., Sebba, J., Blatchford, P., Galton, M., Thorp, J., MacIntyre, H., & Berdondini, L. (2005). The effects of pupil grouping: Literature review (Research report). UCL Institute of Education, Department for Education and Skills. http://dera.ioe.ac.uk/id/eprint/18143

Lee, E. J., & Stone, S. I. (2012). Co-occurring internalizing and externalizing behavioral problems: The mediating effect of negative self-concept. *Journal of Youth and Adolescence*, *41*(6), 717–731. https://doi.org/10.1007/s10964-011-9700-4

Lee, S.-Y., Olszewski-Kubilius, P., & Thomson, D. T. (2012). Academically gifted students' perceived interpersonal competence and peer relationships. *Gifted Child Quarterly*, *56*(2), 90–104. https://doi.org/10.1177/0016986212442568

Lipps, G. E., Lowe, G. A., Halliday, S., Morris-Patterson, A., Clarke, N., & Wilson, R. N. (2010). The association of academic tracking to depressive symptoms among adolescents in three Caribbean countries. *Child and Adolescent Psychiatry and Mental Health*, *4*(1), 16. https://doi.org/10.1186/1753-2000-4-16

MacIntyre, H., & Ireson, J. (2002). Within-class ability grouping: Placement of pupils in groups and self-concept. *British Educational Research Journal*, *28*(2), 249–263. https://doi.org/10.1080/01411920120122176

Malmberg, L. E., & Flouri, E. (2011). The comparison and interdependence of maternal and paternal influences on young children's behavior and resilience. *Journal of Clinical Child & Adolescent Psychology*, *40*(3), 434–444. https://doi.org/10.1080/15374416.2011.563469

Marks, G. N. (2015). Are school-SES effects statistical artefacts? Evidence from longitudinal population data. *Oxford Review of Education*, *41*(1), 122–144. https://doi.org/10.1080/03054985.2015.1006613

Marsh, H. W., & Hau, K.-T. (2003). Big-Fish–Little-Pond effect on academic self-concept: A crosscultural (26-country) test of the negative effects of academically selective schools. *American Psychologist*, *58*(5), 364. https://doi.org/10.1037/0003-066X.58.5.364

Marsh, H. W., Köller, O., & Baumert, J. (2001). Reunification of East and West German school systems: Longitudinal multilevel modeling study of the big-fish-little-pond effect on academic self-concept. *American Educational Research Journal*, *38*(2), 321–350. https://doi.org/10.3102/00028312038002321

Marsh, H. W., Kong, C.-K., & Hau, K.-T. (2000). Longitudinal multilevel models of the big-fish-littlepond effect on academic self-concept: Counterbalancing contrast and reflected-glory effects in Hong Kong schools. *Journal of Personality and Social Psychology*, *78*(2), 337. https://doi.org/10.1037/0022-3514.78.2.337

Marsh, H. W., Seaton, M., Trautwein, U., Lüdtke, O., Hau, K.-T., O'Mara, A. J., & Craven, R. G. (2008). The big-fish–little-pond-effect stands up to critical scrutiny: Implications for theory, methodology, and future research. *Educational Psychology Review*, *20*(3), 319–350. https://doi.org/10.1007/s10648-008-9075-6

Mazenod, A., Francis, B., Archer, L., Hodgen, J., Taylor, B., Tereshchenko, A., & Pepper, D. (2019). Nurturing learning or encouraging dependency? Teacher constructions of students in lower attainment groups in English secondary schools. *Cambridge Journal of Education*, *49*(1), 53–68. https://doi.org/10.1080/0305764X.2018.1441372

Meissel, K., Meyer, F., Yao, E. S., & Rubie-Davies, C. M. (2017). Subjectivity of teacher judgments: Exploring student characteristics that influence teacher judgments of student ability. *Teaching and Teacher Education*, *65*, 48–60. https://doi.org/10.1016/j.tate.2017.02.021

Moilanen, K. L., Shaw, D. S., & Maxwell, K. L. (2010). Developmental cascades: Externalizing, internalizing, and academic competence from middle childhood to early adolescence. *Development and Psychopathology*, *22*(3), 635–653. https://doi.org/10.1017/S0954579410000337

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

Mostafa, T., & Ploubidis, G. (2017). Millennium cohort study. Sixth Survey 2015-2016: Technical report on response (Age 14). Centre for Longitudinal Studies. https://cls.ucl.ac.uk/cls-studies/millennium-cohort-study/mcs-age-14-sweep/

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/10.1080/00131881.2010.524750

Müller, C. M., & Hofmann, V. (2016). Does being assigned to a low school track negatively affect psychological adjustment? A longitudinal study in the first year of secondary school. *School Effectiveness and School Improvement*, *27*(2), 95–115. https://doi.org/10.1080/09243453.2014.980277

Müller, C. M., & Zurbriggen, C. (2016). An overview of classroom composition research on socialemotional outcomes—Introduction to the Special Issue. *Journal of Cognitive Education and Psychology*, *15*(2), 163–184. https://doi.org/10.1891/1945-8959.15.2.163

Nomi, T. (2009). The effects of within-class ability grouping on academic achievement in early elementary years. *Journal of Research on Educational Effectiveness*, *3*(1), 56–92. https://doi.org/10.1080/19345740903277601

Parsons, S., & Hallam, S. (2014). The impact of streaming on attainment at age seven: Evidence from the Millennium Cohort Study. *Oxford Review of Education*, *40*(5), 567–589. https://doi.org/10.1080/03054985.2014.959911

Preckel, F., Götz, T., & Frenzel, A. (2010). Ability grouping of gifted students: Effects on academic self-concept and boredom. *British Journal of Educational Psychology*, *80*(3), 451–472. https://doi.org/10.1348/000709909X480716

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons Inc. https://doi.org/10.1002/9780470316696

StataCorp. (2011). *College station*. StataCorp LP.

Taylor, B., Francis, B., Craig, N., Archer, L., Hodgen, J., Mazenod, A., Tereshchenko, 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*, *67*(1), 5–24. https://doi.org/10.1080/00071005.2018.1424317

Televantou, I., Marsh, H. W., Kyriakides, L., Nagengast, B., Fletcher, J., & Malmberg, L.-E. (2015). Phantom effects in school composition research: Consequences of failure to control biases due to measurement error in traditional multilevel models. *School Effectiveness and School Improvement*, *26*, 75–101. https://doi.org/10.1080/09243453.2013.871302

Trautwein, U., Lüdtke, O., Marsh, H. W., & Nagy, G. (2009). Within-school social comparison: How students perceive the standing of their class predicts academic self-concept. *Journal of Educational Psychology*, *101*, 853. https://doi.org/10.1037/a0016306

Van Houtte, M., & Stevens, P. A. (2008). Sense of futility: The missing link between track position and self-reported school misconduct. *Youth & Society*, *40*(2), 245–264. https://doi.org/10.1177/0044118X08316251

Van Houtte, M., & Stevens, P. A. (2009). Study involvement of academic and vocational students: Does between-school tracking sharpen the difference? *American Educational Research Journal*, *46*(4), 943–973. https://doi.org/10.3102/0002831209348789

Van Landeghem, G., Van Damme, J., Opdenakker, M.-C., De Frairie, D. F., & Onghena, P. (2002). The effect of schools and classes on noncognitive outcomes. *School Effectiveness and*

School Improvement, *13*(4), 429–451. https://doi.org/10.1076/sesi.13.4.429.10284

Wilkinson, S., & 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., Penney, D., & Allin, L. (2016). Setting and within-class ability grouping: A survey of practices in physical education. *European Physical Education Review*, *22*(3), 336–354. https://doi.org/10.1177/1356336X15610784

Zell, E., & Alicke, M. D. (2009a). Contextual neglect, self-evaluation, and the frog-pond effect. *Journal of Personality and Social Psychology*, *97*(3), 467–482. https://doi.org/10.1037/a0015453

Zell, E., & Alicke, M. D. (2009b). Self-evaluative effects of temporal and social comparison. *Journal of Experimental Social Psychology*, *45*, 223–227. https://doi.org/10.1016/j.jesp.2008.09.00

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Papachristou, E., Flouri, E., Joshi, H., Midouhas, E., & Lewis, G. (2021). Ability-grouping and problem behavior trajectories in childhood and adolescence: Results from a U.K. population-based sample. *Child Development*, 00, 1–18. https://doi.org/10.1111/cdev.13674