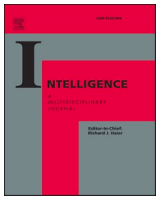
Intelligence 107 (2024) 101857

Contents lists available at ScienceDirect

Intelligence

journal homepage: www.elsevier.com/locate/intell



Sex differences of school grades in childhood and adolescence: A longitudinal analysis

Claire M. Oakley [a,*], Reinhard Pekrun [a,b,c], Gijsbert Stoet [a]

[a] Department of Psychology, University of Essex, United Kingdom
[b] Institute for Positive Psychology and Education, Australian Catholic University, North Sydney, Australia
[c] Department of Psychology, Ludwig Maximilian University, Munich, Germany

ARTICLE INFO

Keywords:
Sex differences
Mathematics achievement
Language achievement
Science achievement
Greater male variability
Development
Childhood
Adolescence
Education

ABSTRACT

We studied trajectories of school achievement in England to determine sex differences in performance and changes in these differences throughout students' development. Using a sample of 5795 children from England born in 2000–2001, this secondary data analysis examined sex differences across a range of school subjects, including differences at the upper and lower tails of the distribution of performance grades. We expected trajectories to differ by subject and to find support for greater male variability in each subject. We found a small male advantage in mathematics at age 11 but no sex differences at ages 7 and 16. Girls achieved higher language grades at each age, but this advantage was notably wider at age 16. Unlike other educational data, there were no sex differences in science achievement at ages 7 and 11 and a small female advantage in science, biology, and chemistry at age 16. Boys' school grades were more variable than girls' in English, reading, and writing at each age. Boys' STEM grades were not consistently more variable than girls' STEM grades. Sex differences were larger at the lower tail in English and the upper tail in mathematics and more balanced in science after age 7. Trajectories of sex differences are age- and subject-specific. By age 16, fewer boys achieved the upper grades, and more boys achieved the lower grades in mathematics and language than at age 11, and we found a female advantage in most school subjects. Implications for practice and directions for future research are discussed.

1. Introduction

Meta-analyses of gender differences in teacher-assigned school grades and multi-year cross-sectional analyses report that girls, on average, outperform boys across most subjects throughout compulsory education. The female advantages reported for mathematics and science grades are smaller than those reported for language subjects (Voyer & Voyer, 2014). These female advantages are, however, contrary to the expected influence of gender stereotypes, lower mathematics or science self-concepts, and less positive mathematics emotions in girls (Eccles et al., 1983; Frenzel et al., 2007). The female advantage in mathematics and science school grades also opposes the male advantage reported in international and national, standardized, large-scale educational assessments (Baye & Monseur, 2016; Reilly et al., 2015; Reilly et al., 2019). These large-scale assessments differ from school grades in that they measure point in time achievement in aptitude tests, whereas school grades also reflect engagement, persistence, and effort over longer periods (Voyer & Voyer, 2014). School grade achievement differences in language, mathematics, and science subjects widen during early to middle adolescence (Cavaglia et al., 2020; O'Dea et al., 2018; Voyer & Voyer, 2014). Similarly, in international assessments, the female advantage in reading is larger in tests of secondary school students than in primary school students (Baye & Monseur, 2016). Depending on the features of an education system, widening achievement gaps during adolescence may influence long-term outcomes in different ways.

National education systems differ in how and when students are channeled from general education to increasingly specialized opportunities. Some countries, such as Germany, Austria, and Hungary, feature early selection (often at the end of primary education) into academic or vocational pathways (van Elk et al., 2011). Others, such as the US and Canada, Denmark, and Finland, emphasize general education and a later choice between academic or vocational options (Burgess et al., 2022; Pekkarinen, 2008; van Elk et al., 2011). The UK education systems are general until age 16, after which vocational options are offered as an alternative to the academic track, with further vocational paths available at or after university (Cavaglia et al., 2020).

* Corresponding author at: University of Essex, Wivenhoe Park, Colchester CO4 3SQ, United Kingdom.
E-mail address: c.oakley@essex.ac.uk (C.M. Oakley).

https://doi.org/10.1016/j.intell.2024.101857
Received 23 February 2024; Received in revised form 26 July 2024; Accepted 18 August 2024
Available online 12 September 2024

The UK education systems feature high-stakes examinations at age 16, which are key determinants of the choice between academic or vocational pathways (Cavaglia et al., 2020). Therefore, learning and preparation for these examinations is an important focus during mid-adolescence. Due to the importance of these examinations in the UK context, understanding and mitigating factors that underlie widening achievement gaps in this age group is especially important. However, in the context of other education systems, particularly where students are making elective subject choices, outcomes are likely to influence the choices made. Here, we extend the literature by establishing the longitudinal pattern of change in a single cohort of subject-specific mean-level and variability achievement differences between girls and boys at year 2 (equivalent to US grade 1, age 7), year 6 (US grade 5, age 11) and year 11 (US grade 10, age 16) focusing on mathematics, language, and science.

In the current study, we evaluate changes in subject achievement trajectories in the context of the biopsychosocial model (Halpern, 2012; Halpern et al., 2004; Miller & Halpern, 2014). This model describes how social or environmental factors interact with biological advantages and developmental change to explain sex differences in mathematics, science achievement, and cognitive abilities (Halpern, 2012; Miller & Halpern, 2014). Focusing on dynamic, reciprocal associations between biological and social/environmental factors, the biopsychosocial model emphasizes the interdependency of nature and nurture but also the potential opportunity to change learning environments to address the impact of social factors on outcomes (Halpern, 2012; Miller & Halpern, 2014). Small biological advantages can be amplified or dampened through approach or avoidance, depending on whether an individual is encouraged or discouraged from engaging in a subject or skill due to social approval or disapproval (Halpern et al., 2004). As such, cultural and environmental factors, social roles, and gender stereotypes will each influence engagement with school subjects and school in general, but biological differences in underlying cognitive abilities, greater male variability, and developmental changes will also contribute to outcomes (Halpern et al., 2004; Miller & Halpern, 2014).

While there is a female advantage in school grades during compulsory schooling in combined, multiple-subject achievement measures such as grade point average, the female advantage varies by subject area or type. Furthermore, effect sizes vary across countries, indicating the influence of cultural, societal, and educational system differences (Voyer & Voyer, 2014; Wu, 2010). The female advantage in language achievement is persistent and well-established across most nations and found consistently across international assessments and school grades (Baye & Monseur, 2016; Stoet, 2015; Stoet & Geary, 2013; Voyer & Voyer, 2014). Multi-year analyses of reading achievement in international assessments report a female advantage at age 10 that is wider in secondary school-age students (Baye & Monseur, 2016). The effect sizes of sex differences reported in international reading assessments are comparable to those reported for school grades in language subjects (Voyer & Voyer, 2014). The female advantage in language and reading achievement is clear, but differences in mathematics and science are less clear.

Boys achieve higher scores than girls in international mathematics assessments, although not in all countries, and there is no clear sex difference in mathematics school grades (Stoet, 2015; Voyer & Voyer, 2014). The Voyer and Voyer (2014) meta-analysis of school grades reported a significant female advantage in mathematics in junior/middle school, high school, and college. Sex differences in elementary school and graduate students were not statistically significant, and, there were no significant sex differences in scholastic mathematics achievement across studies outside North America at all school levels (Voyer & Voyer, 2014). In contrast, the OECD Programme for International Student Assessment (PISA) reported a male advantage in OECD countries, while sex differences vary widely across non-OECD countries (Stoet & Geary, 2013). Importantly, PISA results show that sex differences in international mathematics assessments are larger at the right tail of the distribution (Baye & Monseur, 2016; Stoet & Geary, 2013). In science school grades, female advantages have been reported during adolescence, although there were relatively few studies of sex differences in science achievement included in the meta-analysis (Voyer & Voyer, 2014). Other reports found that sex differences vary between science disciplines (e.g., Deary et al., 2007; Reilly et al., 2015; Stoet, 2015). However, in international science assessments, male advantages are consistently found at higher achievement levels despite mean sex differences being close to zero (Baye & Monseur, 2016). Together, these results highlight a lack of consistency when comparing school grades and international assessments in mathematics and science. The over-representation of North American studies in Voyer and Voyer (2014) meta-analysis may contribute to these inconsistencies - there may be closer alignment between school grades and international assessments within countries. Additionally, the focus on mean sex differences has been criticized, arguing that this focus ignores the contribution of sex differences in the distribution of scores to educational and other outcomes, such as future careers (Baye & Monseur, 2016).

The biopsychosocial model proposes that biological differences in specific abilities and greater male variability (GMV) of intelligence will contribute to mathematics and science achievement, with psychological and social processes potentially amplifying these differences (Halpern, 2012; Halpern et al., 2004; Miller & Halpern, 2014). The GMV hypothesis asserts that males are more variable in many measures of physical and psychological traits, such as intelligence. As a result, male scores are more dispersed, and men/boys are more frequently represented at the tails of the intelligence distribution (Baye & Monseur, 2016; Stoet & Geary, 2013). Several population-sized studies have reported GMV in intelligence. Boys/men are reported to be over-represented at the upper and lower tails of the intelligence distribution, with girls/women overrepresented around population means (Deary et al., 2003; Johnson et al., 2008; Strand et al., 2006). Baye and Monseur (2016) argued that due to the focus on mean sex differences, the literature has neglected to consider that gifted boys may still outperform gifted girls. Differences in the upper portions of the achievement distribution may matter for STEM fields.

The GMV hypothesis is commonly cited as a possible explanation for the significant male majority in STEM careers and at senior levels in tertiary education and academic research (Baye & Monseur, 2016; Hedges & Nowell, 1995; Stoet & Geary, 2013). Proponents of this perspective argue that more men meet the threshold for success in these fields (O'Dea et al., 2018; Spelke, 2005). Some support for this comes from analyses of international assessment scores, which find that sex differences vary across the distribution, being larger in the upper tail in mathematics and science and the lower tail in reading (Baye & Monseur, 2016; Stoet & Geary, 2013). Also, among gifted students taking standardized mathematics aptitude tests, the ratio of boys to girls in the 95th percentile of the mathematics distribution, which is reported to have reduced over time, was most recently reported as 2.5:1 (Benbow, 1988; Makel et al., 2016; Wai et al., 2010). Sex differences in variability in international assessments are reported to increase with age and differ between countries and test providers (Baye & Monseur, 2016; Gray et al., 2019). Also, greater sex differences in variability occur in countries with greater cultural, economic, and political female participation (Gray et al., 2019). In contrast, sex differences in the variability in school grades tend to reduce with age (O'Dea et al., 2018). Boys' grade variability is similar across subjects, and girls' grade variability is greater in non-STEM than in STEM subjects (O'Dea et al., 2018). To our knowledge, GMV in school grades or educational assessments has not been examined longitudinally, and we were unable to identify any studies examining sex differences at the tails of school grade distributions beyond O'Dea et al.'s (2018) meta-analytic comparison.

1.1. Present study

This study sought to understand better the longitudinal pattern of

change in sex differences in a single cohort at ages 7, 11, and 16. We extended the literature by examining differences in science alongside sex differences in mathematics and English and in a range of optional subjects at age 16. We included the optional subjects in lieu of a GPA equivalent measure to indicate the broader trend in male/female school achievement differences. We examined a single, consistent sample across age groups to isolate how sex differences develop. In this way, sex differences cannot be explained or contributed to by changes in sample membership between age groups. We examined teacher-assessed school grades and standardized school grades from national school tests, reporting mean sex differences. As analyses of international educational assessments indicated sex differences can be larger at the tails of the distribution, we examined sex differences at the upper and lower tails of school grade distributions to determine if the same patterns are observed. Based on the empirical research reviewed above, we hypothesized that:

H1. There are subject-specific trajectories in school grades from age seven to sixteen. We expected sex differences to align with findings from the Voyer and Voyer (2014) school grades meta-analysis, reflecting the observation of no sex differences in school mathematics achievement in countries outside of North America. While these data are a mix of teacher-assigned school grades and standardized school grades, despite differing marking processes, both measure curriculum learned over extended periods. The expectation that sex differences will align with the school grades meta-analysis is also supported by prior results from UK samples (Deary et al., 2007; Haworth et al., 2010; Voyer & Voyer, 2014).

H1a. There is a significant female advantage in English grades at ages 7 and 11, which widens by age 16.

H1b. There are no significant sex differences in mathematics grades at any age.

H1c. There are no significant sex differences in science grades at ages 7 and 11 and a significant female advantage by age 16.

We do not provide subject-specific hypotheses for mean differences in the optional subjects studied from ages 14–16 years old, as these are often chosen by students from different achievement levels. However, we expected that sex differences in sufficiently powered samples would align with those reported by Deary et al. (2007).

H2. Boy's school grades are more variable than girl's school grades at all ages.

H3a. Sex differences in variability are larger in non-STEM subjects and smaller in STEM subjects.

H3b. Girls' grade variability is greater in STEM subjects, and boys' grade variability is similar across subjects.

2. Methods

This study is a secondary data analysis of educational outcomes extracted from the UK Department for Education's National Pupil Database (NPD) of learners in England that has previously been matched to survey data from the Millennium Cohort Study (MCS). The NPD records had been matched and extracted for MCS cohort members whose parents had permitted access to educational records. These extracted NPD data are available for use with, but are distinct from, the MCS survey data (Department for Education, UCL Institute of Education, Centre for Longitudinal Studies, University College London, 2021). The NPD contains detailed information on the school achievement of all students throughout their formal education. The educational outcomes collated in the NPD are school grades measuring performance in national assessments of students' learning of the school curriculum. Some are teacher-assessed school grades using a national framework, others are externally-marked, standardized test grades, and the remainder are standardized examination (qualification) grades. The subjects taught from age 14–18 are either vocational (practical or applied subjects) or academic (traditional subjects including languages, mathematics, sciences, and humanities). We focused on achievement trajectories to age 16; age 18 achievement data for the MCS cohort are currently unavailable.

The MCS is a UK-wide, longitudinal cohort study. We analyzed the school grades of MCS cohort members who had responded to each of three school-age MCS survey waves undertaken when cohort members were seven years old (UCL, IoE, Centre for Longitudinal Studies, 2021a), eleven years old (UCL, IoE, Centre for Longitudinal Studies, 2021b) and fourteen years old (UCL, IoE, Centre for Longitudinal Studies, 2021c). While there is an additional school-age wave, known as MCS Age 17, these data were collected after the last available educational achievement data at age 16. Prior MCS waves are available at nine months and three years old. The MCS surveys and school grades align at ages 7 and 11, but there is a lag between the MCS Age 14 survey and the next set of school grades achieved at age 16.

The MCS study used stratified random sampling of children born between September 1, 2000, and January 11, 2002, and resident in England, Scotland, Wales, and Northern Ireland at nine months old, clustered by the characteristics of electoral wards. Electoral wards are national electoral sub-divisions. In 2018, the average population in an electoral ward was 7065. However, there is a wide variation, with some wards having a population of less than two hundred and others having more than forty thousand (Office for National Statistics, 2018). Electoral wards were divided into three strata for sampling: wards with an ethnic minority population of at least 30%, wards from the poorest 25% group (not previously categorized as an ethnic minority), and the remainder, which were designated as advantaged wards (Plewis, 2007). Children were randomly sampled from these strata, with intentional oversampling from ethnic minority and disadvantaged wards. This oversampling enables analyses of ethnic differences, neighborhood, and disadvantage effects.

2.1. Procedure

The sample comprises MCS cohort members who responded to all three survey waves (Ages 7, 11, and 14) and for whom linked educational achievement data were available at ages 7, 11, and 16. We constrained the sample to ensure that we analyzed the outcomes of the same cohort members at each age. The data merging process automatically excludes Scottish, Welsh, and Northern Irish cohort members, as the educational data was for MCS cohort members from England only. We excluded data relating to MCS cohort members who attended non-mainstream schools focusing on educating pupils with more complex educational support needs; the sample contains cohort members who attended mainstream educational settings only. We report demographic information from the MCS and educational outcomes from the NPD.

Data cleaning, transformation, and analyses were undertaken using R version 4.1.1 with RStudio Team, 2021.09.1 for Windows (R Core Team, 2021; RStudio Team, 2020).

2.2. Participants

Educational achievement data were available for 7975 students for whom there were performance assessments from school year 2 (aged 6–7 years old), year 6 (10–11 years old), and year 11 (15–16 years old). Participants were excluded if they did not attend mainstream school settings ($n = 163$). The resulting matched sample consisted of $n = 5795$ MCS cohort members (2846 boys, 2949 girls). The sex of children was identified from the MCS parental survey entry.

As intended in the survey design, the sample is more ethnically diverse than the UK population was when the cohort members were born (for details, see Table 1).

The MCS classifies each cohort member's family income under five

Table 1
Ethnic background of the sample compared to UK Census 2001 results.

UK Census 2001		Sample		
Ethnic Grouping	%	Ethnic Background	*N*	%
White	91.2%	White	4450	77.0%
Asian	4.8%	Pakistani	393	6.8%
		Indian	220	3.8%
		Bangladeshi	165	2.8%
		Other Asian incl. Chinese	73	1.3%
Black	2.2%	Black African	124	2.1%
		Black Caribbean	79	1.4%
		Other Black	20	0.3%
Mixed/Multiple	1.4%	Mixed	205	3.5%
Other	0.4%	Other Ethnic Group	41	0.7%
		Not specified	10	0.2%

Note. UK Census 2001 (ONS, 2022).

equivalized income categories. Family income was adjusted to account for the number and age of dependents supported by that income,in accordance to OECD equivalence scales (Centre for Longitudinal Studies, 2020). Using family income as reported in the MCS Age 7 wave, we find that the sample is distributed across the five income categories (1 - lowest to 5 - highest), with 19–21% of cohort members classified under each group, with a slight 1–2% bias towards categories 3 and 4. 28% of the cohort members live in low-income households as indicated by the OECD 60% of median income indicator. For comparison, the distribution by disposable income of the UK population is positively skewed, and approximately 20% of children live in low-income households (Office for National Statistics, 2022).

Additional educational needs, such as dyslexia, ADHD, and Autism, are referred to as Special Educational Needs (SEN) under the UK system. This broad term covers many intellectual, physical, and behavioral support needs. We created a single SEN indicator (Yes = 1) for cohort members with support needs indicated in at least one of the three data collections. SEN diagnoses were reported by teachers or parents of 1408 (24.3%) cohort members (872 male, 536 female). SEN diagnosis was more prevalent among boys than girls (30.6% vs. 18.2%). The proportion of cohort members with SEN is higher than in this cohort's overall population of learners (14.1%; Department for Education, 2018).

2.3. Measures

The school grades analyzed are the results of national assessments, which take place in the summer term of UK school years 2, 6, and 11 when most students are aged 7, 11, and 16. For the remainder of the paper, we refer to these assessments by age. These assessments are teacher-assessed at age 7, a mix of teacher-assessed and standardized test grades at age 11, and standardized test grades only at age 16.

2.3.1. Achievement at age 7

Age 7 educational assessments are teacher-assessed using a detailed national assessment framework. Students are graded “Below 1” or from 1 to 4. In addition, grade 2 is split into three sub-grades: a-c, where ‘a’ is high, for mathematics, reading, and writing but not science. A combined grade for reading and writing, which we refer to as English for the remainder of this paper, is available for each student. We converted these grades to numeric values 0–4, with the sub-grades converted to 2c = 2.25, 2b = 2.5, and 2a = 2.75. Grade 2b is the expected level of achievement in this age group. A grading of 2a indicates students working at the top of grade 2, and 2c indicates those working at the bottom of grade 2. Grades were converted to z-scores.

2.3.2. Achievement at age 11

Age 11 educational assessments comprise both teacher-assessed grades and standardized test grades marked anonymously by external examiners. These assessments are reported separately and are not combined. Teacher assessments are graded from 1 to 6, where 6 is the highest level of achievement. Teacher-assessed grades are provided for mathematics, English, writing, and science. Test outcomes, available for English and mathematics only, are graded from 1 to 6. In addition, test grades 3 to 5 are further split into three sub-grades: a – c, where a is high. We converted these grades to numeric values 1–6, with the sub-grades converted to c = 0.25, b = 0.5, and a = 0.75. The expected level of achievement in this age group is grade 4b in English and mathematics and grade 4 in science. Below, in Table 2, we report sex differences in age 11 standardized school test grades for reading, writing, English, and mathematics, and teacher-assessed grades for writing and science. We report age 11 teacher-assessed grades for English and mathematics separately (see Supplementary Materials, Tables SM.B.2–3, p. 11).

2.3.3. Achievement at age 16

The achievement scores for the core subjects of mathematics and English language (reading, writing, and grammar) were graded using 1–9 grades, which we then converted to z-scores using the population data by subject from the 2017 cohort (Department for Education, 2018). The age 16 achievement data for the remaining subjects are alphanumeric grades G - A*. We converted the alphanumeric grades to numeric values for analysis: 1–8, where 8 = A*.

Under England's education system, pupils either study science – a combination of physics, biology, and chemistry – or study physics, biology, and chemistry as separate subjects. Physics, biology, and chemistry are most often studied by the upper third of students, who are less likely to obtain lower grades. Two compulsory options are available for the remaining two-thirds of students: science single-award or science double-award. Students studying for the science double-award learn two-thirds of the content from the physics, biology, and chemistry courses (Ofqual., 2018). Students opting for science single-award learn one-third of the physics, biology, and chemistry content. There is also an optional additional science award. As the students choosing to study science mostly come from the lower two-thirds of achievers at age 11, as a group they are underrepresented at the upper achievement levels. Double awards are graded as A*A*, A*A, AA, AB, BB, etc. These were converted to numeric grades by converting each grade individually, summing the two numeric grades, and calculating the mean result. For example, A*A*: (8 + 8)/2 = 8; BC: (6 + 5)/2 = 5.5. The double-award and single-award science results were combined for analysis as very few cohort members studied for the double-award. Grades for each subject were converted to z-scores. To facilitate comparison across the full distribution for science at age 16, we calculated an average science grade. This measure averaged each student's achievement across science and the optional additional science course or across physics, biology, and chemistry.

A minority of participants sat their mathematics or English examinations earlier than the remainder of their cohort and were therefore awarded G - A* grades for these subjects. The alphanumeric and numeric

Table 2
Standardized means, standard deviations, and percentages of boys and girls observed at the upper and lower tails of the grade distributions in English, Reading, Writing, and Mathematics.

Subject	Age	Sex	*N*	*M* (Cohen's *d* [CI])	*SD* (σ^2_k *p*)	Percentage of Girls/Boys (χ^2 *p*)			
						Lowest Level	Lower Level	Upper Level	Highest Level
English	7	F	2948	**0.15**	0.94	0.6%	10.5%	**34.6%**	–
	TA			0.31 [0.25, 0.36]	< 0.001	< 0.001	<0.001	<0.001	
		M	2846	−0.15	**1.04**	**1.8%**	**17.4%**	25.8%	–
	11	F	2929	**0.14**	0.94	4.1%	6.2%	**12.1%**	–
				0.29 [0.24, 0.35]	< 0.022	< 0.001	<0.001	< 0.001	–
		M	2825	−0.15	**1.03**	**7.7%**	**11.1%**	7.5%	–
	16	F	2900	**0.30**	0.96	1.1%	5.3%	**11.7%**	**3.9%**
				0.41 [0.36, 0.47]	0.005	< 0.001	<0.001	<0.001	< 0.001
		M	2763	−0.10	**0.99**	**4.2%**	**13.0%**	6.2%	1.7%
Reading	7	F	2949	**0.13**	0.93	0.7%	7.9%	**34.1%**	–
	TA			0.26 [0.21, 0.32]	< 0.001	< 0.001	< 0.001	< 0.001	
		M	2846	−0.13	**1.05**	**2.2%**	**13.6%**	26.1%	–
	11	F	2937	**0.11**	0.94	4.3%	6.1%	**10.4%**	–
				0.23[0.18, 0.28]	< 0.001	< 0.001	< 0.001	< 0.001	
		M	2836	−0.12	1.05	**7.7%**	**10.1%**	6.7%	–
Writing	7	F	2949	**0.16**	0.95	1.1%	10.4%	**17.7%**	–
	TA			0.32 [0.27, 0.37]	< 0.001	< 0.001	< 0.001	< 0.001	
		M	2846	−0.16	1.03	**3.1%**	**17.8%**	11.4%	–
	11	F	2945	**0.16**	0.97	1.0%	8.9%	**40.1%**	**1.7%**
	TA			0.32[0.27, 0.37]	< 0.001	0.005	< 0.001	< 0.001	< 0.009
		M	2846	−0.16	**1.01**	**1.9%**	**16.6%**	26.9%	0.9%
Mathematics	7	F	2949	0.02	0.93	0.7%	5.3%	22.4%	–
	TA			−0.05 [−0.10, 0.01]	. < 0.001	0.160	<0.001	<0.001	
		M	2846	−0.02	**1.06**	1.1%	**7.7%**	**29.7%**	–
	11	F	2940	−0.08	0.97	6.6%	**10.6%**	11.4%	3.3%
				−0.16[−0.21, −0.11]	0.017	0.161	0.018	<0.001	< 0.001
		M	2836	**0.08**	**1.02**	5.7%	8.7%	**17.0%**	**5.6%**
	16	F	2884	0.13	0.97	5.4%	12.1%	12.0%	3.7%
				−0.01 [−0.07, 0.04]	0.051	0.593	0.401	0.109	0.049
		M	2773	0.14	**1.01**	5.8%	12.8%	13.5%	**4.8%**

Note. Sex differences are nested between boys and girls means, standard deviations and percentage representation at each achievement level: Cohen's *d* [95% CI], Levene's (σ^2_k) test *p*-value, and Pearson chi-square (χ^2) test *p*-value for each achievement level. Statistically significant differences are indicated in bold typeface. Data are excluded (–) where the observed count of boys or girls <10. Observed frequencies at the lowest, lower, upper, and highest achievement levels, and which grades each level encompasses are detailed in Supplementary Materials, Table SM.A.1–2. TA indicates teacher-assessed grades.

grading systems are not directly comparable due to changes in the content, method of assessment, and grade distributions, so they are excluded from the main analyses.

2.4. Analyses

We analyzed mean differences in school grades between boys and girls assessed in mathematics, science, and language subjects at ages 7, 11, and 16. We employed independent sample *t*-tests to test group differences and used the Welch modification to estimate degrees of freedom where variances are not similar. We tested sex differences in variability using Levene's test (Delacre et al., 2017). We also calculated Cohen's *d* statistic (Cohen, 1988). Cohen's guidelines for interpreting effect sizes have been reexamined empirically, and revisions have been suggested (e.g., Gignac & Szodorai, 2016; Hemphill, 2003; Rubio-Aparicio et al., 2018). Following a more recent examination of effect sizes across psychological disciplines and study designs, we describe effect sizes using lower, grand, and upper medians reported for between-subjects designs in educational psychology (Schäfer & Schwarz, 2019). Consequently, we describe $d = 0.20$, $d = 0.36$, and $d = 0.62$ as small, medium, and large effects, respectively. We compared male achievement against female results (Female = 1, Male = 0). The sample size for mathematics, English at all ages, and science at ages 7 and 11 provides 98% power to detect mean differences of $d = 0.10$, or 84.5% power in detecting a mean difference as low as $d = 0.07$ (Champely, 2020). Post-hoc achieved power is calculated where elective subject choice results in smaller sample sizes.

We examined the tails of the grade distributions for each subject, using Pearson chi-square tests to report the significance of the observed numbers of boys and girls achieving grades at the upper and lower tails. We selected those grades that accounted for the closest to 10% of the cohort at each tail for mathematics, science, and English – the upper and lower achievement levels. Similarly, we also examined the upper and lower 5% in each subject – the highest and lowest achievement levels. We applied these science grade groupings to define the upper and lower tails in each elective subject where choice may influence outcomes and skew grade distributions.

3. Results

3.1. Longitudinal changes in mean school grades

As expected (H1), the achievement gap between boys and girls changed with age and differed by subject (see Fig. 1). Mean sex differences in English were stable during primary school and widened from small to moderate during secondary school (age 7: $d = 0.31$; age 11: $d = 0.29$; age 16: $d = 0.41$), supporting Hypothesis H1a. Sex differences in reading and writing were small and stable (reading, age 7: $d = 0.26$; age

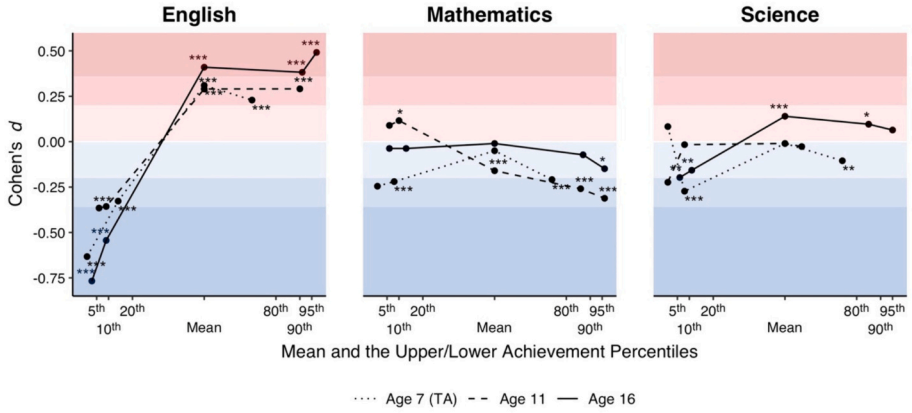


Fig. 1. Effect Size of Sex Differences at the Mean and at the Tails of the English, Mathematics, and Science Distributions by Age.
Note. Each datapoint represents the observed percentile in the sample by subject and age (see Tables SM.A.1–3 for the percentage of cohort members observed at each achievement level). Science outcomes at age 16 are represented by an average grade across science options studied: science and additional science, or biology, chemistry, and physics. Positive values (pink section) indicate an overrepresentation of girls and negative values (blue section) an overrepresentation of boys. Areas of increasing opacity indicate negligible, small, and medium effect sizes respectively. Age 11 science outcomes are teacher-assessed (TA), all other age 11 and age 16 outcomes are standardized grades. *** $p < .001$, ** $p < .01$, * $p < .05$.

11: $d = 0.23$; writing, age 7: $d = 0.32$; age 11: $d = 0.32$). However, we observed significant changes in the overrepresentation of boys at the lower achievement levels between age groups, despite the stability of mean sex differences in English and reading during primary school. We also highlight that the age 11 assessments in reading are standardized school grades, while writing grades are teacher-assessed. We find teacher-assessed grades to be less reliable for examining sex differences at the upper and lower tails of the distribution (see Supplementary Materials, Section B, p. 10–15, for an analysis of teacher-assessed vs. standardized school grades and further discussion in the Limitations section). There were no significant sex differences in mathematics achievement at ages seven and sixteen and a very small, significant male

Table 3
Standardized means, standard deviations, and percentages of boys and girls observed at the upper and lower tails of the grade distributions in combined and single science subjects.

Subject	Age	Sex	*N*	*M* (Cohen's *d* [CI])	*SD* (σ^2_K *p*)	Percentage of Girls/Boys (χ^2 *p*)			
						Lowest Level	Lower Level	Upper Level	Highest Level
Science	7 TA	F	2947	−0.00	0.95	0.4%	5.8%	24.5%	–
				−0.01[−0.06, 0.04]	< 0.001	0.896	<0.001	0.001	
		M	2845	0.00	**1.05**	0.4%	**9.2%**	**28.2%**	–
	11 TA	F	2945	−0.01	0.99	0.6%	7.6%	42.8%	–
				−0.01[−0.06, 0.04]	0.152	0.240	0.817	0.334	
		M	2846	0.01	1.01	0.9%	7.8%	44.1%	–
	16	F	2062	**0.09**	0.98	4.3%	11.4%	**6.9%**	0.8%
				0.19[0.13, 0.25]	0.140	0.032	< 0.001	0.005	0.186
		M	1972	−0.10	1.01	**5.6%**	**14.5%**	5.1%	0.7%
Additional Science	16	F	1955	**0.09**	0.98	4.7%	14.0%	**13.5%**	2.6%
				0.19 [0.12, 0.25]	0.663	0.006	0.002	0.001	0.098
		M	1815	−0.10	1.01	**6.8%**	**17.7%**	9.7%	1.6%
Individual Sciences									
Biology	16	F	826	**0.18**	0.97	–	–	**13.8%**	4.8%
				0.15[0.05, 0.25]	0.629			0.006	0.115
		M	779	0.03	1.01	–	–	11.7%	3.9%
Chemistry	16	F	810	**0.10**	0.96	–	–	13.4%	4.8%
				0.12[0.03, 0.22]	0.748		.	0.021	0.542
		M	780	−0.02	1.03	–	–	11.8%	4.5%
Physics	16	F	817	0.02	0.98	–	–	12.0%	4.8%
				−0.04 [−0.14, 0.06]	0.303		.	0.271	0.480
		M	774	0.06	1.02	–	–	12.7%	5.2%
Across Science Subjects									
Average Science Grade	16	F	2850	**−0.03**	0.99	5.0%	9.6%	**15.9%**	5.4%
				0.14 [0.10, 0.17]	0.096	0.002	0.001	0.021	0.362
		M	2715	−0.19	1.05	**7.0%**	**12.4%**	13.7%	4.9%

Note. Sex differences are nested between boys and girls means, standard deviations and percentage representation at each achievement level: Cohen's *d* [95% CI], Levene's (σ^2_K) test *p*-value, and Pearson chi-square (χ^2) test *p*-value for each achievement level. Achieved power at age 16 – science: 0.99, biology: 0.91, chemistry: 0.77, physics: 0.20 (Champely, 2020). Data are excluded (–) where the observed count of boys or girls <10. We combined the age 16 science and individual science cohorts to calculate a representative percentage at each grade in the sciences: science: $n = 5639$; biology: $n = 5639$; chemistry $n = 5624$, physics $n = 5525$. Averaged science grade combines individual outcomes across science and additional science, or across biology, chemistry, and physics. Observed frequencies at the lowest, lower, upper, and highest achievement levels, and which grades each level encompasses, are detailed in Supplementary Materials, Table SM.A.3. TA indicates teacher-assessed grades.

Table 4
Standardized means, standard deviations, and percentages of boys and girls observed at the upper and lower tails of the grade distributions in age 16 optional subjects.

Subject	Sex	N	*M* (Cohen's *d* [CI])	P	*SD* (σ_k^2 *p*)	Percentage of boys/girls (χ^2 *p*)			
						Lowest Level	Lower Level	Upper Level	Highest Level
STEM									
Additional Science	F	1955	**0.09**		0.98	4.7%	14.0%	**13.5%**	2.6%
			0.19 [0.12, 0.25]	0.99	0.663	0.006	0.002	0.001	0.098
	M	1815	−0.10		1.01	**6.8%**	**17.7%**	9.7%	1.6%
Computing	F	158	**0.30**		0.88	7.0%	10.1%	**32.9%**	**10.1%**
			0.39 [0.21, 0.57]	0.99	0.019	<0.001	< 0.001	0.001	0.007
	M	584	−0.08		**1.02**	**15.4%**	**21.0%**	20.9%	4.5%
Geography	F	1202	**0.12**		0.97	6.4%	13.2%	**29.5%**	**9.3%**
			0.23 [0.15, 0.31]	1.00	0.805	0.013	< 0.001	< 0.001	0.036
	M	1309	−0.11		1.01	**9.2%**	**19.6%**	21.2%	7.0%
Social Sciences	F	410	**0.10**		0.97	6.1%	14.3%	**29.1%**	–
			0.32 [0.15, 0.49]	0.95	0.842	0.284	0.209	< 0.001	
	M	192	−0.21		1.01	8.8%	18.7%	14.5%	–
Modern Foreign Languages									
French	F	899	0.10		0.98	2.7%	8.0%	**26.6%**	11.4%
			0.25 [0.15, 0.35]	0.99	0.260	0.022	0.002	< 0.001	0.064
	M	637	**−0.15**		1.01	5.0%	**12.9%**	18.5%	8.3%
German	F	260	0.12		0.95	1.5%	4.2%	25.4%	8.1%
			0.25 [0.08, 0.43]	0.86	0.414	0.586	0.011	0.333	0.667
	M	226	**−0.14**		1.04	2.7%	**10.6%**	21.2%	6.6%
Spanish	F	520	**0.14**		0.97	3.1%	7.3%	**31.4%**	**14.0%**
			0.34 [0.21, 0.48]	1.00	0.824	0.167	0.001	0.002	0.013
	M	370	−0.20		1.01	5.1%	**14.6%**	21.6%	8.4%
Humanities									
History	F	1458	**0.11**		0.97	8.1%	15.2%	**31.9%**	**11.3%**
			0.24 [0.16, 0.31]	1.00	0.063	0.005	< 0.001	< 0.001	< 0.001
	M	1266	−0.13		1.02	**11.4%**	**20.5%**	23.8%	6.9%
Religious Studies	F	1780	**0.19**		0.92	5.6%	11.6%	**37.0%**	**12.8%**
			0.41 [0.34, 0.48]	1.00	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
	M	1546	−0.21		**1.04**	**12.8%**	**22.3%**	22.2%	6.9%
Applied Subjects									
Design & Technology	F	728	**0.35**		0.91	3.7%	9.3%	**28.3%**	**9.6%**
			0.61 [0.47, 0.75]	1.00	0.140	< 0.001	< 0.001	< 0.001	< 0.001
	M	959	−0.26		0.98	**10.7%**	**24.9%**	10.3%	2.9%
Film, TV, Media & Office	F	532	**0.22**		0.91	5.3%	10.3%	**24.4%**	**7.3%**
			0.40 [0.28, 0.51]	0.98	0.017	< 0.001	< 0.001	< 0.001	< 0.001
	M	672	−0.17		**1.03**	**11.5%**	**20.4%**	14.9%	2.8%
Physical Education / Sports Studies	F	433	**0.18**		0.98	–	7.2%	**28.6%**	9.5%
			0.29 [0.17, 0.41]	0.99	0.105		0.014	< 0.001	0.085
	M	708	−0.11		1.01	–	**11.9%**	15.3%	6.5%
Arts									
Art & Design	F	1108	**0.16**		0.95	2.1%	6.1%	**28.2%**	**11.7%**
			0.53 [0.42, 0.64]	1.00	0.580	< 0.001	< 0.001	< 0.001	< 0.001
	M	488	−0.36		1.01	**5.5%**	**14.8%**	14.0%	4.7%

Note. Sex differences are nested between boys and girls means, standard deviations and percentage representation at each achievement level: Cohen's *d* [95% CI], Levene's (σ_k^2) test *p*-value, and Pearson chi-square (χ^2) test *p*-value for each achievement level. P indicates achieved power (Champely, 2020). Data are excluded (−-) where the observed count of boys or girls <10. Subject groups: Social Sciences - Sociology, Psychology. Female majority D&T subjects - Food Technology, Textiles Technology. Male majority D&T subjects - Electronic Products, Product Design, Resistant Materials, Systems & Control, Graphic Products, Design & Technology. Film, TV, Media & Office - Office Technology, Film Studies, Information & Communications Technology. Performing Arts including Drama & Theatre Studies, Dance, Music. Tourism, Catering, Home Economics & Care - Catering Studies, Child Development, Food. Observed frequencies at the lowest, lower, upper, highest achievement levels, and which grades each level encompasses, are detailed in Supplementary Materials, Table SM.A.5.

advantage at age eleven (age 7: $d = -0.05$; age 11: $d = -0.16$; age 16: $d = -0.01$), in partial support of Hypothesis H1b. For detailed results for mathematics, English, reading, and writing, see Table 2. There were no significant mean differences in science achievement at ages 7 and 11 (age 7: $d = -0.01$; age 11: $d = -0.01$. We found very small female advantages in science ($d = 0.19$), chemistry ($d = 0.12$), and biology ($d = 0.15$), and no significant sex differences in physics ($d = -0.04$), in partial support of our Hypothesis H1c (for details, see Table 3).

We also analyzed mean differences across a broad range of optional subjects studied in secondary school. Some subjects were grouped due to the low numbers of students enrolled in these subjects. Girls achieved better results, on average, in most optional subjects (for details, see Table 4). Design and technology (D&T) is a broad subject that can be taught at a general level or can be more specialized, so multiple different options are available. Due to the breadth of individual D&T subjects offered, we grouped the options into female and male majority groups to examine whether mean differences were similar depending on male or female participation preferences. We found a large female advantage in the D&T male and female majority groupings (female majority D&T: $d = 0.75$; male majority D&T: $d = 0.61$); these groupings are combined in Table 4 due to comparatively few upper grades achieved by the cohort members who studied these subjects (for details, see Supplementary Materials Table SM.A.5).

3.2. Longitudinal changes in grade variability

Boys' school grades were always significantly more variable than girls' in English, reading, and writing. Science grades were significantly more variable at age 7 but not at ages 11 and 16. Mathematics grades were significantly more variable at ages 7 and 11 but not at age 16. Our Hypothesis H2 was, therefore, partially supported (see Figs. 1, 2 and Tables 2, 3). Sex differences in grade variability were smaller in STEM

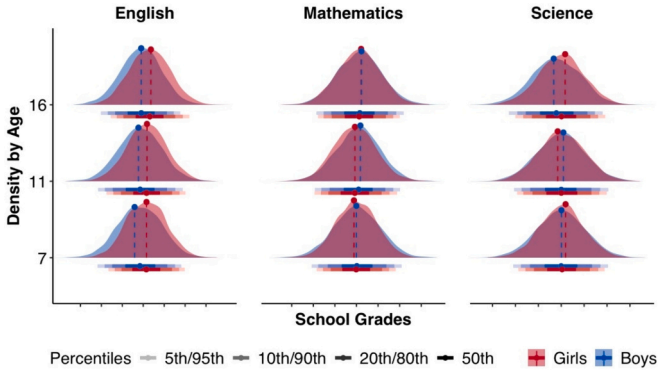


Fig. 2. Implied density distributions, mode, mean, and percentile intervals of boys' and girls' English, mathematics, and science achievement by age group.
Note. Implied distributions generated from means and standard deviations reported in Tables 2 and 3. Mode is indicated as a dotted line and point on each density distribution, with percentile intervals and group means illustrated below these.

subjects in support of Hypothesis H3. Hypothesis H3a was partially supported: girls' grades were more variable in mathematics and science than in English and reading at age 11, but there was no clear pattern at age 16. Boys' grades were most variable in reading and least variable in science at age 11. In general, boys' grades were less variable in English at 16 than at age 11, whereas girls' grades were more variable. Girls' and boys' grade variability did not significantly differ for most optional subjects studied at age 16 (see Table 4 and Supplementary Materials, Table SM.A.5).

The percentage of cohort members at the upper and lower tails differed between subjects at age 16 (see Supplementary Materials, Tables SM.A.1–2). More students were observed at the tails of the distribution in mathematics and science than in English. Grade distributions for mathematics and science were quite similar, with approximately 5% of students observed at the highest and lowest achievement levels and 13% at the upper and lower levels. In English, 3% of students achieved the highest and lowest achievement levels, and 9% achieved the upper and lower levels. A higher percentage of students were observed at the upper and lower tails in the optional subjects at age 16 compared to mathematics, English, and science. In the optional subjects, the proportion of students observed at the upper achievement levels ranged from 19 to 30%, with 9 to 20% observed at the lower levels.

The percentage of students assessed at the lowest achievement levels in science at ages 7 and 11 was substantially lower than at age 16. The reliance on teacher-assessed grades for science assessments in primary school contributes to this difference. Teacher's lesser use of the lowest grades is found consistently across all teacher-assessed grades in all subjects. Similarly, very few students are assessed at the highest available teacher-assessed grade at age 7. The distribution of teacher-assessed grades is compressed around the median grade. At age 11, in test outcomes, an increased percentage of students are observed at the lowest achievement levels and a reduced percentage of students at the upper levels.

In English, boys were significantly overrepresented in the lower tail and girls in the upper tail at each age, indicating stronger female performance in this subject throughout (see Fig. 2). The same pattern was evident in reading and writing at ages 7 and 11. As teacher-assessed grades are less granular than test grades (see Measures section for details), it is impractical to contrast the representation of boys and girls at the upper tail between ages 7 and 11. Similarly, changes at the lower tail may, in part, be due to the change from teacher-assessed grades to test assessments. Among the lower 9% at age 16 in English, there were fewer girls and more boys than the age 11 tests. At the lowest levels of English achievement, the proportion of boys to girls increased from nearly 2:1 to 4:1 between ages 11 and 16.

In mathematics, boys were significantly overrepresented in the upper and lower tails at age 7 only. At age 11, boys were significantly overrepresented among the upper 13% and 5%, girls in the lower 13%, and there was parity in the lower 5%. Effect sizes were small in the upper and highest grades and very small in the lower grades. At age 16, boys were significantly overrepresented at the highest grade only, a very small effect among these top 5% achievers. Neither sex was significantly overrepresented in the upper 13% in mathematics nor at the lower 13% or 5%.

In science, at age 7, boys were significantly overrepresented at the upper and lower tails. As science is teacher-assessed at ages 7 and 11, we observe the same reduced differentiation between students and lesser use of the lowest grades by teachers, and we can compare changes between the two assessments. We find that teachers had allocated upper and lower grades as frequently for boys as for girls at age 11. However, there is a small, but not statistically significant, overrepresentation of boys in the lowest achievement level. At age 16, in science and biology, we find a very small, significant female overrepresentation in the upper tail (upper 13% in science and biology). Otherwise, differences in the upper 13% and 5% in chemistry and physics and the upper 5% in biology were not significant. At the lower tail, indicated by achievement in science only, there was a very small, significant overrepresentation of boys in the lower 13% and 5%.

Among the optional subjects, GMV was supported in computing and religious studies but not in most other subjects. Girls achieved a greater share of the highest grade (the upper 5–12%) in most subjects, except physical education/sports studies, additional science, French, and German, where there was parity. Significant effect sizes at the upper tail were small in the natural sciences, computing, history, and modern foreign languages and moderate in the social sciences, religious studies, and applied subjects. Girls were overrepresented in the upper achievement level in all subjects except German. Similarly, boys were overrepresented in the lower achievement level in most subjects except for the social sciences (psychology and sociology) and modern foreign languages, with effect sizes ranging from small to moderate.

4. Discussion

Overall, our results were consistent with our hypothesis (H1) that trajectories of changes in sex differences differ by subject and provided support for greater male variability in language subjects at all ages. However, there were no sex differences in science grade variability after age 7 or mathematics grade variability at age 16.

4.1. Longitudinal changes of sex differences in mean school grades

Girls achieved significantly higher school grades in English, reading, and writing at each age. In mathematics, we found a male advantage at age 11, but there were no significant sex differences in mathematics at ages 7 and 16. There were no significant sex differences in science achievement at ages 7 and 11, but at age 16, a female advantage in science, biology, and chemistry. The English and science results fully support our Hypotheses H1a and H1c. Our Hypothesis H1b is partially supported as the significant male advantage at age 11 was unexpected.

In each subject, across the three age levels, sex differences varied - either the mean difference or at the tails of the achievement distribution. Between ages 7 and 11, there were no significant changes in mean sex differences in English, reading, science, and writing, and a significant male advantage opened in mathematics. There was a non-significant trend of reducing mean sex differences in English and reading, whereas science and writing remained flat despite apparent changes in the distribution of grades. By age 16, the female advantage in English widened, the very small age 11 male advantage in mathematics was no longer evident, and there were very small female advantages in science, biology, and chemistry. In addition, we found female advantages in most optional subjects at age 16, indicating a broad female advantage in

education in this age group.

The magnitude and direction of change in sex differences in English and reading align with analyses of international assessments, the Voyer and Voyer (2014) meta-analysis of school grades, and US results from national assessments (Baye & Monseur, 2016; Reilly et al., 2019; Voyer & Voyer, 2014). Like analyses from US national assessments, we find that the female advantage in writing is larger than in reading. The small effect sizes we report in writing are smaller than those reported in the US (Reilly et al., 2019). The reliance on teacher-assessed grades at age 11 for writing may influence our results. We found that sex differences in teacher-assessed school grades are similar to sex differences in standardized test grades.

Given the results from the Voyer and Voyer (2014) meta-analysis and national outcomes in the UK for the population cohort from which the MCS participants have been sampled (see Supplementary Materials, SM. B.1), the male advantage in mathematics at age 11 was unexpected. Despite this, our results align with prior reports from analyses of nationally representative US datasets of a male advantage in mathematics emerging by the end of elementary school despite no mean differences at school entry (Cimpian et al., 2016; Fryer Jr. & Levitt, 2010; Robinson & Lubienski, 2011). As our sample is more ethnically diverse and has a higher proportion of SEN students than the population cohort, this finding may support accounts of sex differences in mathematics varying across student characteristics, for example, ethnic group membership (Hsieh et al., 2021). However, as the age 11 male advantage is eliminated by age 16, and overrepresentation of either sex at the tails is reduced at that age, the relative contribution of the factors influencing mathematics achievement at age 11 may change, or other factors may become more important in this period. For example, children's endorsement of gender stereotypes may develop based on perceptions of adults' beliefs in primary-age children, such as teachers underrating girls' abilities in mathematics, which has been linked with the widening male advantage in mathematics (Kurtz-Costes et al., 2014; Robinson-Cimpian et al., 2014). However, gender stereotypes may be perceived and endorsed differently by younger and older children (Kurtz-Costes et al., 2014). Adolescents are increasingly influenced by their peers and their alignment with social groups that may conform with or reject gendered roles (Ahmed et al., 2022; Santos et al., 2013; Smyth, 2020). Additionally, future educational or career plans develop and crystallize during secondary school, with girls reporting higher educational expectations, which may have an important influence on educational outcomes (Platt & Parsons, 2017).

In the UK, all students study science until age 16. The choice between studying science vs. studying biology, chemistry, and physics as individual subjects is influenced by prior achievement. The individual science subjects tend to be studied by more able pupils, most often those who had exceeded the benchmark standard at age 11. Science is studied more often by cohort members from the middle and lower tail of the achievement distribution. Other than mathematics, the smallest sex differences at age 16 are found in the sciences, including biology and chemistry. There were no sex differences in physics. Few studies have examined sex differences in science achievement, and even fewer examined differences in biology, chemistry, and physics. Yet, important gender differences exist in the uptake of the life sciences and physical sciences in higher education. Our age 16 physics and science results align with prior reports from a larger UK sample, although the female advantage we report for chemistry is smaller (Deary et al., 2007). Our finding of no sex differences in science achievement at ages 7 and 11 aligns with prior reports in children (aged 9–12) (Haworth et al., 2010). Our science results align with the Voyer and Voyer (2014) school grades meta-analysis but are inconsistent with a US report of a male advantage at grade 8 in NAEP (National Assessment Educational Progress) science assessments (Reilly et al., 2015) and international assessments (Baye & Monseur, 2016).

Across the optional subjects assessed at age 16, a female advantage was found in all subjects except economics. The economics result lacked statistical power, indicating that this may have been statistically significant in a larger sample. Our reported mean differences for history and religious studies align with those reported previously (Deary et al., 2007). Our effect sizes are slightly larger in geography and smaller in Spanish and Art & Design, although Deary and colleagues' results sit within the confidence intervals we report. Our effect sizes for French and German are smaller, and our effect size for physical education is larger than those previously reported (Deary et al., 2007). There are also some differences between the studies for design and technology and the performing arts vs. drama and music, which we have had to group differently due to our smaller sample size compared with the much larger prior study (Deary et al., 2007). We conclude that despite this sample being intentionally more ethnically diverse than England's population, having a larger proportion of students with SEN, and being unrepresentative of the income distribution in England, our results, for the most part, align with those of the larger representative sample from age 16 achievement in England, 2002 (Deary et al., 2007). The lack of any substantial variation in a non-representative sample highlights the consistency of the female advantage reported in most school subjects in secondary school.

4.2. Longitudinal changes of sex differences in grade variability

We find greater male variability in school grades in English at all ages but in mathematics at ages 7 and 11 only and in science at age 7 only. Hypothesis H2 is, therefore, partially supported. In line with prior reports and our Hypothesis H3, sex differences were smaller in science and mathematics than in English, reading, and writing. Hypothesis H3a is partially supported. Girls' grades were more variable in mathematics and science than in English at age 11, and boys' grades were more variable in English and reading at age 11 than in mathematics. At age 16, boys' grades were more variable in STEM subjects, but the pattern for girls was unclear.

Our results in English and mathematics support prior findings from analyses of international assessments that sex differences in variability are larger at the tails of the distribution - the upper tail in mathematics and the lower tail in reading (Baye & Monseur, 2016; Stoet & Geary, 2013). However, the distribution of our science results does not follow a pattern like that of mathematics, which is inconsistent with the Baye and Monseur (2016) findings. The trend in science is of female overrepresentation in the upper 10% and 5% in science, biology, and chemistry at age 16. This sex difference is significant in science and biology in the upper 10%. Boys are significantly overrepresented at the lower tail in science at age 16.

The ratio of boys to girls in English achievement at age 16 in the upper grades aligns with those reported for age 15 students based on PISA tests (Stoet & Geary, 2013). Converting between odd ratios and Cohen's *d*, we find that our reported sex differences in the lower grades (the lower 10% of students) align with analyses of PIRLS and PISA data (Baye & Monseur, 2016; Ben-Shachar et al., 2020). While our lowest grades in English represent the lowest 3% of students and are not directly comparable with the results reported from international assessments, we can draw some conclusions here. We find a greater overrepresentation of boys in the lowest grades than in the lower 5% in PIRLS and PISA assessments (Baye & Monseur, 2016; Stoet & Geary, 2013). Compared to the lower 1% in PISA tests, we report an overrepresentation of males of similar magnitude, noting that the ratio of boys to girls increased between the PISA surveys reported (Stoet & Geary, 2013). Together, our results and those from international assessments indicate an increasing overrepresentation of boys at the lower tail in reading and English, and the effect size of this difference is moderate to large.

The sex differences at the tails of the mathematics school grades distribution partially align with some reported differences at the tails in international assessments. However, patterns and magnitudes differ between tests and age groups. We find alignment at the lower tail at age

16 with PISA results that boys are marginally overrepresented here (Baye & Monseur, 2016; Stoet & Geary, 2013). The very small effect size we report in the highest grade, the upper 5% of achievers in age 16 mathematics, is larger than the negligible effect size reported for TIMSS tests from 2000 onwards and smaller than the effect in PISA and SAT-M tests (Baye & Monseur, 2016; Stoet & Geary, 2013; Wai et al., 2010). At the lower tail, the marginal sex differences in mathematics we report at age 16 align with PISA test outcomes but oppose TIMSS secondary age tests that tend to find more girls in the lower scores, where we observe more boys (Baye & Monseur, 2016; Stoet & Geary, 2013). Like the primary age TIMSS findings, we find more girls at the lower tail at age 11, although our difference is of small magnitude at the lower 10% vs a negligible male overrepresentation in TIMSS. We find a moderate male advantage at the upper tail, which is negligible in TIMSS (Baye & Monseur, 2016).

In science, there is little alignment between sex differences at the tails of school grade distributions and international assessments. Particularly, international assessments often report more girls at the lower tail and more boys at the upper tail, whereas we find the opposite. We observe more boys in the lower grades across science, biology, chemistry, and physics at age 16. The difference is of small magnitude, except among the higher achievers in the individual science subjects, where the difference is negligible and not significant. Our results at the lower tail align with the earlier TIMSS (1995) data, but this trend was reversed in TIMSS over time, with more girls reported at the lower tail in the 2007 tests. PISA always finds more girls at the lower tail, and the difference is most commonly of small magnitude (Baye & Monseur, 2016). We found more girls at the upper tail in science, biology, and chemistry but not in physics, where there are marginally more boys. Our results at the upper tail are inconsistent with both PISA and TIMSS, in which male advantages are reported throughout (Baye & Monseur, 2016). The negligible female overrepresentation in the lower 8% aligns with the primary TIMSS outcomes at the lower tail in science (Baye & Monseur, 2016). However, the age 11 variability results we report in science are less reliable as these are teacher-assessed (see Supplementary Materials, SM.B.3).

4.3. Implications

The biopsychosocial model outlines the contribution of developmental and cognitive differences, including GMV, to differences in mathematics and science achievement. The model also highlights that psychosocial factors are particularly important and, therefore, will influence any biological differences (Halpern et al., 2007; Miller & Halpern, 2014). Gender stereotypes, lower expectancies and values, and less positive emotions are each predicted to influence girls' outcomes in mathematics and science negatively. The general pattern of achievement in mathematics from ages 7 to 11 would align with this account, as despite parity in achievement at age 7, there is a small gap favoring males at age 11. As reported elsewhere, this gap could result from lower female mathematics self-concept, interest, and less positive emotions (Frenzel et al., 2007; Jacobs et al., 2002; Wilkins, 2004).

That girls close the male advantage in mathematics school grades and open a female advantage in most science subjects between ages eleven and sixteen opposes the influence of the negative predictors of female outcomes in these subjects. No evidence suggests that girls' mathematics or science self-concepts improve during adolescence. In fact, both male and female mathematics self-concepts has been found to reduce during this period, although the male advantage is maintained (Jacobs et al., 2002; Wilkins, 2004). Therefore, we consider whether other factors may be more influential in adolescence, which may explain why girls' achievement is, on average, better than that of boys across most subjects at age 16.

We highlight that there is substantial within-sex variation in achievement. We do not find that all boys underachieve or that all girls do well. At age 16, boys and girls are equally represented at the highest grades in biology, physics, and chemistry, for example. In secondary school mathematics, only in the upper 5% of achievers were girls underrepresented. Otherwise, a similar percentage of boys and girls are represented in the upper and lower grades in mathematics. However, certain groups do less well, as evidenced by the increased representation of girls in the lower grades in mathematics at age 11, and the substantial increase in the overrepresentation of boys in the lower grades in English at age 16. Our finding that the overrepresentation of boys in the upper grades in mathematics reduces between ages 11 and 16 suggests that some boys who do very well at the end of primary school do not maintain this level of achievement at age 16.

One proposal that attempts to explain the underachievement of some groups of boys and girls is conformity to gendered social roles linked to patterns of motivation, engagement, and achievement, which are pertinent during adolescence (Santos et al., 2013; Yu et al., 2020). The influence of gendered social roles is encompassed within the biopsychosocial account, although it does not address how the influence of these may change with age or in the context of peer groups (Halpern et al., 2004, Halpern et al., 2007). The recently outlined developmental cascades model of educational attainment proposes that the influence of social and environmental factors will change during adolescence as the home environment becomes less influential and peers, school, and the wider neighborhood become increasingly important as individuals spend less time at home (Ahmed et al., 2022). Peers are a particularly important source of influence on motivation and achievement during adolescence (Ryan, 2001; Santos et al., 2013; Yu et al., 2020).

An alternate but not mutually exclusive factor may be the influence of educational expectations and career plans on outcomes. While continuing education is compulsory in the UK from sixteen until eighteen years old, this may take the form of remaining in full-time education to study for academic or vocational qualifications or enrolling in an apprenticeship or traineeship. A good grade (grade 4 or above) in mathematics and English is a prerequisite for many post-age sixteen opportunities in the UK, so the influence of post-16 plans may be a potential source of motivation to improve grades. Post-16 opportunities may provide an additional focus, resulting in individuals employing extra effort to ensure they achieve the grades they need in the important subjects for their future. The influence of post-16 opportunities would be expected to apply equally to both sexes. However, boys have been reported to have lower educational expectations than girls (Platt & Parsons, 2017). The higher rates of girls choosing academic options or attending university are well established in the UK and in many other Western countries (Broeke & Hamed, 2008; Cavaglia et al., 2020; Lundberg, 2020; Rampino & Taylor, 2013; Wiseman & Zhao, 2020). Not all educational systems feature high-stakes tests at age 16 that influence future options, and the widening of achievement gaps in adolescence is not peculiar to countries where post-16 options are contingent on achievement at age 16. However, even in those countries without high-stakes examinations, many education systems measure consistency of achievement through combined measures such as grade point averages, requiring focused effort across a broad range of subjects. If, in addition, boys prioritize male-stereotyped subjects, including mathematics and physics, then the overall lower achievement of boys may, therefore, result from lower engagement and expectations in some subjects and selective de-prioritization of these subjects (Wirthwein et al., 2020).

The developmental cascades model also highlights the contribution of biological cascades, which include pubertal maturation, to educational achievement (Ahmed et al., 2022). Several studies have linked puberty timing, often late puberty in boys and early puberty in girls, with underachievement in education and a reduced likelihood of boys choosing to continue in education after age sixteen (Cavanagh et al., 2007; Koerselman & Pekkarinen, 2018; Koivusilta & Rimpelä, 2004; Pekkarinen, 2008, 2012). The effects of puberty on achievement are likely to be indirect, possibly resulting from an association with academic motivation, and may differ depending on which measure of maturation is used (Martin et al., 2022; Martin & Steinbeck, 2017;

Torvik et al., 2021). Pekkarinen (2008, 2012) argued that the change in the Finnish education systems from an early (pre-puberty) to a late (post-puberty) choice between vocational or academic options has contributed to the lower educational achievement of boys compared to girls due to the later maturation of boys. An association between pubertal status and academic motivation has also been reported, finding that higher levels of pubertal hormones predict increased academic disengagement, particularly for boys (Martin et al., 2022; Martin & Steinbeck, 2017). Evidence in this area is sparse and requires further investigation to establish how puberty may be linked to educational outcomes.

4.4. School grades vs. international assessments

The discrepancies between sex differences in school grades, TIMSS, and PISA scores in mathematics and science tests and other standardized aptitude tests such as SAT-M are poorly understood. Comparisons of the 2003 TIMSS and PISA mathematics assessments report that sex differences in mathematics, like sex differences in school grades, vary across countries and between test providers (Voyer & Voyer, 2014; Wu, 2010). Students from Western countries perform better in PISA tests, and students from Asian and Eastern European countries perform better in TIMSS tests, for example. Differences in curricula and the number of schooling years between countries are thought to contribute to country differences in performance in standardized tests (Wu, 2010). There are also notable differences in the balance of content between different standardized tests. Girls and boys perform better in some mathematics content areas within the TIMSS tests, but the male advantage is found across all mathematics content areas of the PISA test in most countries (Wu, 2010). For example, girls perform equally or better in algebra content, but this accounts for a much lower proportion of PISA tests. Other research indicates that the format of the tests may also contribute to sex differences, with boys achieving higher scores in multiple-choice questions and girls in constructed response items (Reardon et al., 2018; Shear, 2023).

One explanation of the inconsistency of mathematics achievement gaps between international assessments and school grades proposes a female disadvantage when tests contain novel problems versus problems that reflect content learned at school (Kimball, 1989; Willingham & Cole, 1997). It is highlighted that TIMSS tests are oriented towards school mathematics, whereas PISA tests are oriented towards basic mathematical competencies (Wu, 2010). An alternate explanation centers around sex differences in learning styles (Kenney-Benson et al., 2006). Girls prioritize mastery over performance achievement goals and are less disruptive in classroom settings, resulting in better school grades. However, mathematics self-concept is a better predictor of outcomes in standardized aptitude tests, contributing to the male advantage (Kenney-Benson et al., 2006).

Despite our finding of no sex differences in achievement in school mathematics at age 16, there remains an overrepresentation of boys in the upper 5%, although this is smaller than we report at age 11. While individual tests and school grades in each country may disagree on the size of the male overrepresentation among the highest achievers, this is often found across many different types of tests but seems to vary depending on the focus and mathematical content tested (Baye & Monseur, 2016; Benbow, 1988; Stoet & Geary, 2013; Wu, 2010). Benbow and colleagues' research, focused on the upper 3% of gifted US students, reports that there are no sex differences in attitudes towards mathematics among talented populations (Benbow, 1988). Across a multitude of studies of this population, the authors conclude that biological factors, lower female self-confidence, sex-differentiated socialization, and greater male variability contribute to the overrepresentation of boys at the highest levels of mathematical reasoning ability (Benbow, 1988). However, there is little evidence here that differences in mathematics or science ability justify the underrepresentation of girls at increasingly higher levels of mathematics after age 16. The underrepresentation of girls in the highest grades in mathematics is very small, and girls are equally represented across the highest levels in science. The more substantial overrepresentation of girls in the upper grades in English and larger female advantages in most non-STEM subjects would support accounts of female abilities being more balanced or perhaps tilted towards non-STEM subjects (Valla & Ceci, 2014; Wang et al., 2013).

4.5. Strengths and limitations

The use of the MCS cohort for longitudinal outcomes analysis has clear strengths; the sample size is large, diverse, and spread across socioeconomic groups. While there has been attrition between waves, and permission to access educational data is optional, the final sample size remains large. However, selection biases may limit our ability to generalize these findings to the wider population, as may our decision to exclude those cohort members attending special educational settings, which may exclude some individuals with moderate to severe intellectual disabilities. The resulting sample is almost equally split across the OECD family income quintiles, so it is not unduly biased towards a particular family income group but also does not reflect the population's income distribution. The sample has slightly higher rates of cohort members from disadvantaged backgrounds than the national cohort: 34% were eligible for free school meals in at least one of the three surveys vs. 27% of the national cohort were categorized as disadvantaged (Department for Education, 2018). The disadvantaged categorization identifies those pupils eligible for free school meals (FSM) at some point during the prior six years of schooling, so it will exclude those receiving FSM before age ten but not since. Our FSM indicator will reflect those who met the criteria at any time between ages seven and fourteen. Therefore, some variation in the percentage of FSM in this sample compared to the national cohort would be expected. However, this sample also has a higher percentage of financially disadvantaged cohort members. Finally, as highlighted in Table 1, the sample has more cohort members from UK ethnic minority groups. Despite these limitations, the overall trends identified are mostly consistent with findings reported elsewhere, and results at age sixteen do not substantially deviate from those of the representative sample who sat their age 16 examinations in 2002 (Deary et al., 2007).

A further limitation is that all assessments at age 7 and the science and writing assessments at age 11 are teacher-assessed. Our analyses in Table 2 (see also Supplementary Materials, Table SM.A.1) demonstrate that at age 7 teacher-assessed grades are less reliable than standardized test grades. The teacher-assessed grades were systematically less granular than test grades, reducing differentiation between students. However, there is also evidence that they are less reliable at the upper and lower tails. Students are rarely assessed at the highest available grades, and comparatively few are assessed at the lowest. For example, in reading at age 11, 4.0% of students were assessed in the lowest grades through externally marked tests, whereas in teacher-assessed writing, only 1.4% were assessed at the lowest grades. The percentage of students assessed in the lowest grades in writing at age 11 is similar to the percentage of students assessed at the lowest grades at age 7 in reading and writing, indicating that this trend is independent of age group - substantially fewer students are assessed at the lower grades by their teachers. Even fewer students were assessed at the lowest grades in mathematics. However, it can be argued that other factors contribute to the difference at the lower tails, such as standardized school test outcomes not reflecting performance and achievement in class. In supplementary analyses (see Supplementary Materials, Table SM.B.2, p. 10–15), we compared age 11 teacher-assessed grades and standardized school grades in English and mathematics. Like the age 7 assessments, fewer students were observed at the tails of the grade distribution in teacher-assessed grades, and more students were observed at the tails in standardized school grades.

Finally, we highlight that attrition between survey waves, separate

permission to access educational records, and our exclusion of cohort members from non-mainstream educational settings may have resulted in some portions of the distribution being over or under-sampled. Selection biases may, therefore, limit the generalizability of age-based conclusions regarding representation at the upper and lower tails. Similarly, the comparison of grades rather than test scores limits the ability to differentiate at percentiles above that represented by the highest achievement level. Upper grade boundaries may also be high, resulting in ceiling effects. However, there is no evidence that this applies in the age 16 examinations; for example, there is a 25% total marks gap above the grade 9 mathematics boundary and a 23% gap for English (AQA Education, 2017). Despite these limitations, we highlight how the representation at the tails changes over time. There is instability between age groups, particularly in STEM subjects, but also at the lower tail in language domains. We conclude that the composition at the upper and lower tails is unstable, and the group of students at the extremes can change markedly over time in school grades.

5. Conclusion

Sex differences in school grades change with age, and they are subject-specific. Interventions must, therefore, be targeted accordingly. Sex differences at the lower tail in English are especially problematic and suggest a disproportionate number of boys leave school with low literacy levels (4% of boys vs. 1% of girls). Focused attention is needed to ensure that all students leaving mainstream education are able to read and write effectively. Interventions to improve boys' achievement in language subjects must be broadly applied to reduce the size of the language achievement gap that persists throughout compulsory education. Girls' math achievement at age 16, given prior achievement at age 11, is inconsistent with the expected effects of expectancy-value beliefs and lower emotions in male-stereotyped subjects. Girls achieve a greater share of upper grades, and boys the lower grades, in most subjects studied, including male-stereotyped subjects such as science and chemistry, despite no sex differences in science at age 11. More nuanced theoretical explanations are required to explain the general pattern of girls' and boys' achievement in early to middle adolescence.

Author Note

This study is a secondary data analysis of the educational achievement trajectories of Millennium Cohort Study (MCS) cohort members for whom age 7, 11, and 16 educational outcomes data are available. Educational outcomes were sourced from the UK Department for Education's National Pupil Database (NPD) of England and Wales. Access to MCS & MCS-linked NPD educational data is controlled and managed by the UK Data Service (https://ukdataservice.ac.uk/).

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Claire M. Oakley: Writing – original draft, Visualization, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Reinhard Pekrun:** Writing – review & editing, Supervision. **Gijsbert Stoet:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.intell.2024.101857.

References

Ahmed, S. F., Chaku, N., Waters, N. E., Ellis, A., & Davis-Kean, P. E. (2022). Chapter ten—Developmental cascades and educational attainment. In C. S. Tamis-Lemonda, & J. J. Lockman (Eds.), *Vol. 64. Advances in child development and behavior* (pp. 289–326). JAI. https://doi.org/10.1016/bs.acdb.2022.10.006.

AQA Education. (2017). Grade boundaries—June 2017 exams. GCSE reformed (subjects which use the 9 to 1 grade scale). https://filestore.aqa.org.uk/over/stat_pdf/AQA-GCSE-RF-GDE-BDY-JUN-2017.PDF.

Baye, A., & Monseur, C. (2016). Gender differences in variability and extreme scores in an international context. *Large-Scale Assessments in Education, 4*. https://doi.org/10.1186/s40536-015-0015-x

Benbow, C. P. (1988). Sex differences in mathematical reasoning ability in intellectually talented preadolescents: Their nature, effects, and possible causes. *Behavioral and Brain Sciences, 11*(2), 169–183. https://doi.org/10.1017/S0140525X00049244

Ben-Shachar, M. S., Lüdecke, D., & Makowski, D. (2020). Effectsize: Estimation of effect size indices and standardized parameters. *Journal of Open Source Software, 5*(56), 2815. https://doi.org/10.21105/joss.02815

Broeke, S., & Hamed, J. (2008). *Gender gaps in higher Education participation DIUS Research Report 08–14* (p. 42). Department for Innovation, Universities and Skills. https://dera.ioe.ac.uk/8717/1/DIUS-RR-08-14.pdf.

Burgess, S., Hauberg, D. S., Rangvid, B. S., & Sievertsen, H. H. (2022). The importance of external assessments: High school math and gender gaps in STEM degrees. *Economics of Education Review, 88*, Article 102267. https://doi.org/10.1016/j.econedurev.2022.102267

Cavaglia, C., Machin, S., McNally, S., & Ruiz-Valenzuela, J. (2020). Gender, achievement, and subject choice in English education. *Oxford Review of Economic Policy, 36*(4), 816–835. https://doi.org/10.1093/oxrep/graa050

Cavanagh, S., Reigle-Crumb, C., & Crosnoe, R. (2007). Puberty and the education of girls. *Social Psychology, 70*(2), 186–198. doi:10.1177%2F019027250707000207.

Centre for Longitudinal Studies. (2020). *Millennium cohort study, user guide (surveys 1–5)* (9th ed.). Institute for Education, University of London.

Champely, S. (2020). Pwr: Basic functions for power analysis. https://CRAN.R-project.org/package=pwr.

Cimpian, J. R., Lubienski, S. T., Timmer, J. D., Makowski, M. B., & Miller, E. K. (2016). Have gender gaps in math closed? Achievement, teacher perceptions, and learning behaviors across two ECLS-K cohorts. *AERA Open, 2*(4), Article 2332858416673617. https://doi.org/10.1177/2332858416673617

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge. https://doi.org/10.4324/9780203771587

Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence, 35*(1), 13–21. https://doi.org/10.1016/j.intell.2006.02.001

Deary, I. J., Thorpe, G., Wilson, V., Starr, J. M., & Whalley, L. J. (2003). Population sex differences in IQ at age 11: The Scottish mental survey 1932. *Intelligence, 31*(6), 533–542. https://doi.org/10.1016/S0160-2896(03)00053-9

Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use welch's t-test instead of student's t-test. *International Review of Social Psychology, 30*(1), 1. https://doi.org/10.5334/irsp.82

Department for Education. (2018). *SFR01/2018 subject tables (statistical first release 01/2018; revised GCSE and equivalent results in England: 2017 to 2017)*. Department for Education. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/784809/SFR01_2018_Subject_tables.xlsx.

Department for Education, UCL Institute of Education, Centre for Longitudinal Studies, University College London. (2021). *Millennium Cohort Study: Linked Education Administrative Datasets (National Pupil Database), England: Secure Access. [Data collection]* (2nd ed.). UK Data Service. https://doi.org/10.5255/UKDA-SN-8481-2. SN: 8481.

Eccles, J., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1983). Expectancies, values and academic behaviors. In J. T. Spence (Ed.), *Achievement and Achievement Motives*. W. H. Freeman.

van Elk, R., van der Steeg, M., & Webbink, D. (2011). Does the timing of tracking affect higher education completion? *Economics of Education Review, 30*(5), 1009–1021. https://doi.org/10.1016/j.econedurev.2011.04.014

Frenzel, A. C., Pekrun, R., & Goetz, T. (2007). Girls and mathematics—A "hopeless" issue? A control-value approach to gender differences in emotions towards mathematics. *European Journal of Psychology of Education, 22*(4), 497–514. https://doi.org/10.1007/BF03173468

Fryer, R. G., Jr., & Levitt, S. D. (2010). An empirical analysis of the gender gap in mathematics. *American Economic Journal: Applied Economics, 2*(2), 210–240. https://doi.org/10.1257/app.2.2.210

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences, 102*, 74–78. https://doi.org/10.1016/j.paid.2016.06.069

Gray, H., Lyth, A., McKenna, C., Stothard, S., Tymms, P., & Copping, L. (2019). Sex differences in variability across nations in reading, mathematics and science: A meta-analytic extension of Baye and Monseur (2016). *Large-Scale Assessments in Education, 7*(1), 2. https://doi.org/10.1186/s40536-019-0070-9

Halpern, D. F. (2012). *Sex differences in cognitive abilities* (4th ed.). Psychology Press.

Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, R. C., Hyde, J. S., & Gernsbacher, M. A. (2007). The science of sex differences in science and mathematics. *Psychological Science in the Public Interest: A Journal of the American Psychological Society, 8*(1), 1–51. https://doi.org/10.1111/j.1529-1006.2007.00032.x

Halpern, D. F., Wai, J., & Saw, A. (2004). A Psychobiosocial model: Why females are sometimes greater than and sometimes less than males in math achievement. In A. M. Gallagher, & J. C. Kaufman (Eds.), *Gender differences in mathematics: An integrative psychological approach* (pp. 48–72). Cambridge University Press. https://doi.org/10.1017/CBO9780511614446.004.

Haworth, C. M. A., Dale, P. S., & Plomin, R. (2010). Sex differences in school science performance from middle childhood to early adolescence. *International Journal of Educational Research, 49*(2), 92–101. https://doi.org/10.1016/j.ijer.2010.09.003

Hedges, L. V., & Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science, 269*(5220), 41–45. https://doi.org/10.1126/science.7604277

Hemphill, J. F. (2003). Interpreting the magnitudes of correlation coefficients. *American Psychologist, 58*(1), 78–79. https://doi.org/10.1037/0003-066X.58.1.78

Hsieh, T., Simpkins, S. D., & Eccles, J. S. (2021). Gender by racial/ethnic intersectionality in the patterns of Adolescents' math motivation and their math achievement and engagement. *Contemporary Educational Psychology, 66*, Article 101974. https://doi.org/10.1016/j.cedpsych.2021.101974

Jacobs, J. E., Lanza, S., Osgood, D. W., Eccles, J. S., & Wigfield, A. (2002). Changes in Children's self-competence and values: Gender and domain differences across grades one through twelve. *Child Development, 73*(2), 509–527.

Johnson, W., Carothers, A., & Deary, I. J. (2008). Sex differences in variability in general intelligence: A new look at the old question. *Perspectives on Psychological Science, 3*(6), 518–531.

Kenney-Benson, G. A., Pomerantz, E. M., Ryan, A. M., & Patrick, H. (2006). Sex differences in math performance: The role of children's approach to schoolwork. *Developmental Psychology, 42*(1), 11–26. https://doi.org/10.1037/0012-1649.42.1.11

Kimball, M. M. (1989). A new perspective on women's math achievement. *Psychological Bulletin, 105*(2), 198–214. https://doi.org/10.1037/0033-2909.105.2.198

Koerselman, K., & Pekkarinen, T. (2018). Cognitive consequences of the timing of puberty. *Labour Economics, 54*, 1–13. https://doi.org/10.1016/j.labeco.2018.05.001

Koivusilta, L., & Rimpelä, A. (2004). Pubertal timing and educational careers: A longitudinal study. *Annals of Human Biology, 31*(4), 446–465. https://doi.org/10.1080/03014460412331281719

Kurtz-Costes, B., Copping, K. E., Rowley, S. J., & Kinlaw, C. R. (2014). Gender and age differences in awareness and endorsement of gender stereotypes about academic abilities. *European Journal of Psychology of Education, 29*(4), 603–618. https://doi.org/10.1007/s10212-014-0216-7

Lundberg, S. (2020). Educational gender gaps. *Southern Economic Journal, 87*(2), 416–439. https://doi.org/10.1002/soej.12460

Makel, M. C., Wai, J., Peairs, K., & Putallaz, M. (2016). Sex differences in the right tail of cognitive abilities: An update and cross cultural extension. *Intelligence, 59*, 8–15. https://doi.org/10.1016/j.intell.2016.09.003

Martin, A. J., Balzer, B., Garden, F., Handelsman, D. J., Hawke, C., Luscombe, G., … Steinbeck, K. (2022). The role of motivation and puberty hormones in adolescents' academic engagement and disengagement: A latent growth modeling study. *Learning and Individual Differences, 100*, Article 102213. https://doi.org/10.1016/j.lindif.2022.102213

Martin, A. J., & Steinbeck, K. (2017). The role of puberty in students' academic motivation and achievement. *Learning and Individual Differences, 53*, 37–46. https://doi.org/10.1016/j.lindif.2016.11.003

Miller, D. I., & Halpern, D. F. (2014). The new science of cognitive sex differences. *Trends in Cognitive Sciences, 18*(1), 37–45. https://doi.org/10.1016/j.tics.2013.10.011

O'Dea, R. E., Lagisz, M., Jennions, M. D., & Nakagawa, S. (2018). Gender differences in individual variation in academic grades fail to fit expected patterns for STEM. *Nature Communications, 9*(1), 3777. https://doi.org/10.1038/s41467-018-06292-0

Office for National Statistics. (2018). *Population estimate by output areas, electoral, health and other geographies, England and Wales: Mid 2017* (p. 20). Statistical Bulletin.

Office for National Statistics. (2022). *Households below average income: An analysis of the UK income distribution: FYE 1995 to FYE 2022*. Department for Work & Pensions. www.gov.uk/government/statistics/households-below-average-income-for-financial-years-ending-1995-to-2022/households-below-average-income-an-analysis-of-the-uk-income-distribution-fye-1995-to-fye-2022#children-in-low-income-households.

Ofqual.. (2018). *Grading the new GCSE science qualifications*. The Ofqual Blog. March 23 https://ofqual.blog.gov.uk/2018/03/23/grading-the-new-gcse-science-qualifications/.

ONS. (2022). *Ethnicity facts and figures: By ethnicity over time [Governmental web site]*. Office for National Statistics. https://www.ethnicity-facts-figures.service.gov.uk/uk-population-by-ethnicity/national-and-regional-populations/population-of-england-and-wales/latest.

Pekkarinen, T. (2008). Gender differences in educational attainment: Evidence on the role of tracking from a Finnish quasi-experiment. *The Scandinavian Journal of Economics, 110*(4), 807–825. https://doi.org/10.1111/j.1467-9442.2008.00562.x

Pekkarinen, T. (2012). Gender differences in education. *Nordic Economic Policy Review, 1*.

Platt, L., & Parsons, S. (2017). *Is the future female? Educational and occupational aspirations of teenage boys and girls in the UK (working paper 17/2017)*. Centre for Longitudinal Studies.

Plewis, I. (2007). The millennium cohort study: Technical report on sampling. In *Centre for longitudinal studies*. http://doc.ukdataservice.ac.uk/doc/4683/mrdoc/pdf/mcs_technical_report_on_sampling_4th_edition.pdf.

R Core Team. (2021). R: A language and environment for statistical computing. https://www.R-project.org/.

Rampino, T., & Taylor, M. P. (2013). *Gender differences in educational aspirations and attitudes* (ISER working paper 2013–15). In *Institute for social and economic research (ISER)*. University of Essex.

Reardon, S. F., Kalogrides, D., Fahle, E. M., Podolsky, A., & Zairate, R. (2018). The relationship between test item format and gender achievement gaps on math and ELA tests in 4th and 8th grade. *Educational Researcher, 47*(5) (47(5)) https://cepa.stanford.edu/content/relationship-between-test-item-format-and-gender-achievement-gaps-math-and-ela-tests-4th-and-8th-grade.

Reilly, D., Neumann, D. L., & Andrews, G. (2015). Sex differences in mathematics and science achievement: A meta-analysis of national assessment of educational Progress assessments. *Journal of Educational Psychology, 107*(3), 645–662. https://doi.org/10.1037/edu0000012

Reilly, D., Neumann, D. L., & Andrews, G. (2019). Gender differences in reading and writing achievement: Evidence from the national assessment of educational Progress (NAEP). *American Psychologist, 74*(4), 445–458. https://doi.org/10.1037/amp0000356

Robinson, J. P., & Lubienski, S. T. (2011). The development of gender achievement gaps in mathematics and Reading during elementary and middle school: Examining direct cognitive assessments and teacher ratings. *American Educational Research Journal, 48*(2), 268–302. https://doi.org/10.3102/0002831210372249

Robinson-Cimpian, J. P., Lubienski, S. T., Ganley, C. M., & Copur-Gencturk, Y. (2014). Teachers' perceptions of students' mathematics proficiency may exacerbate early gender gaps in achievement. *Developmental Psychology, 50*(4), 1262–1281. https://doi.org/10.1037/a0035073

RStudio Team. (2020). *RStudio: Integrated development for R*. RStudio, PBC. http://www.rstudio.com.

Rubio-Aparicio, M., Marín-Martínez, F., Sánchez-Meca, J., & López-López, J. A. (2018). A methodological review of meta-analyses of the effectiveness of clinical psychology treatments. *Behavior Research Methods, 50*(5), 2057–2073. https://doi.org/10.3758/s13428-017-0973-8

Ryan, A. M. (2001). The peer group as a context for the development of young adolescent motivation and achievement. *Child Development, 72*(4), 1135.

Santos, C. E., Galligan, K., Pahlke, E., & Fabes, R. A. (2013). Gender-typed behaviors, achievement, and adjustment among racially and ethnically diverse boys during early adolescence. *The American Journal of Orthopsychiatry, 83*(2 Pt 3), 252–264. https://doi.org/10.1111/ajop.12036

Schäfer, T., & Schwarz, M. A. (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology, 10*. https://doi.org/10.3389/fpsyg.2019.00813

Shear, B. R. (2023). Gender Bias in test item formats: Evidence from PISA 2009, 2012, and 2015 math and Reading test. *Journal of Educational Measurement*. https://doi.org/10.1111/jedm.12372

Smyth, E. (2020). Shaping educational expectations: The perspectives of 13-year-olds and their parents. *Educational Review, 72*(2), 173–195. https://doi.org/10.1080/00131911.2018.1492518

Spelke, E. S. (2005). Sex differences in intrinsic aptitude for mathematics and science?: A critical review. *American Psychologist, 60*(9), 950–958. https://doi.org/10.1037/0003-066X.60.9.950

Stoet, G. (2015). British male students continue to fall behind in secondary education. *New Male Studies, 4*(3), 23–49.

Stoet, G., & Geary, D. C. (2013). Sex differences in mathematics and Reading achievement are inversely related: Within- and across-nation assessment of 10 years of PISA data. *PLoS One, 8*(3), Article e57988. https://doi.org/10.1371/journal.pone.0057988

Strand, S., Deary, I. J., & Smith, P. (2006). Sex differences in cognitive abilities test scores: A UK national picture. *British Journal of Educational Psychology, 76*(3), 463–480. https://doi.org/10.1348/000709905X50906

Torvik, F. A., Flatø, M., McAdams, T. A., Colman, I., Silventoinen, K., & Stoltenberg, C. (2021). Early puberty is associated with higher academic achievement in boys and girls and partially explains academic sex differences. *Journal of Adolescent Health, 69*(3), 503–510. https://doi.org/10.1016/j.jadohealth.2021.02.001

University of London, Institute of Education, Centre for Longitudinal Studies. (2021a). *Millennium cohort study: Age 7, sweep 4, 2008. [Data collection]* (8th ed.). UK Data Service. https://doi.org/10.5255/UKDA-SN-6411-8. SN: 6411.

University of London, Institute of Education, Centre for Longitudinal Studies. (2021b). *Millennium cohort study: Age 11, sweep 5, 2012. [Data collection]*. (5th ed.). UK Data Service. https://doi.org/10.5255/UKDA-SN-7464-5. SN: 7464.

University of London, Institute of Education, Centre for Longitudinal Studies. (2021c). *Millennium cohort study: Age 14, sweep 6, 2015. [Data collection]* (7th ed.). UK Data Service. https://doi.org/10.5255/UKDA-SN-8156-7. SN: 8156.

Valla, J. M., & Ceci, S. J. (2014). Breadth-based models of women's underrepresentation in STEM fields: An integrative commentary on Schmidt (2011) and Nye et al. (2012). *Perspectives on Psychological Science: A Journal of the Association for Psychological Science, 9*(2), 219–224. https://doi.org/10.1177/1745691614522067

Voyer, D., & Voyer, S. D. (2014). Gender differences in scholastic achievement: A meta-analysis. *Psychological Bulletin, 140*(4), 1174–1204. https://doi.org/10.1037/a0036620

Wai, J., Cacchio, M., Putallaz, M., & Makel, M. C. (2010). Sex differences in the right tail of cognitive abilities: A 30year examination. *Intelligence, 38*(4), 412–423. https://doi.org/10.1016/j.intell.2010.04.006

Wang, M.-T., Eccles, J., & Kenny, S. (2013). Not lack of ability but more choice. *Psychological Science, 24*. https://doi.org/10.1177/0956797612458937

Wilkins, J. L. M. (2004). Mathematics and science self-concept: An international investigation. *The Journal of Experimental Education, 72*(4), 331–346. https://doi.org/10.3200/JEXE.72.4.331-346

Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment.* Routledge.

Wirthwein, L., Sparfeldt, J. R., Heyder, A., Buch, S. R., Rost, D. H., & Steinmayr, R. (2020). Sex differences in achievement goals: Do school subjects matter? *European Journal of Psychology of Education, 35*(2), 403–427. https://doi.org/10.1007/s10212-019-00427-7

Wiseman, A., & Zhao, X. (2020). Parents' expectations for the educational attainment of their children: A cross-national study using PIRLS 2011. *Prospects*. https://doi.org/10.1007/s11125-020-09460-7

Wu, M. (2010). *Comparing the similarities and differences of PISA 2003 and TIMSS (32; OECD education working papers)*. Assessment Research Centre, University of Melbourne. https://doi.org/10.1787/5km4psnm13nx-en

Yu, J., McLellan, R., & Winter, L. (2020). Which boys and which girls are falling behind? linking adolescents' gender role profiles to motivation, engagement, and achievement. *Journal of Youth and Adolescence, 50*, 336–352. https://doi.org/10.1007/s10964-020-01293-z