Longitudinal study of the influence of puberty on school engagement and widening achievement gaps during secondary education.

Claire M. Oakley

A thesis submitted for the degree of PhD in Psychology

Department of Psychology
University of Essex

June 2025

Longitudinal study of the influence of puberty on school engagement and widening achievement gaps during secondary education.

General Abstract

A female advantage in education is evident during secondary education, and fewer boys continue into tertiary education. Current educational theories fail to explain the changes observed in STEM subjects. The earlier onset of puberty in girls has been linked to the female advantage. Puberty is associated with a wide range of developmental changes in adolescence. However, the mechanisms connecting these developmental processes to academic achievement remain underexplored. Using the longitudinal panel and educational achievement data of 5,795 UK school students, this thesis investigates the developmental and psychosocial mechanisms underlying educational outcomes during adolescence. Study 1 explores how sex differences in academic achievement evolve from childhood to age 16 across core subjects, revealing a widening female advantage, particularly in language and increasingly in STEM disciplines. The findings challenge the assumption that men are more prevalent at higher levels of STEM achievement. Study 1 also examines the role of teacher grading bias, finding limited evidence that it substantially contributes to sex achievement gaps. Study 2 extends the analysis by examining the association between puberty, academic self-concepts, educational expectations and school engagement. While puberty was associated with declines in school engagement, especially among girls, self-concepts and expectations emerged as stronger predictors of engagement. This thesis highlights that adolescent identity development can be influenced by peers, parents, and teachers, thereby shaping educational trajectories. Additionally, structural and societal factors may drive girls to prioritise their education to achieve a living standard comparable to that of boys. The higher prevalence of special educational needs (SEN) diagnoses among boys likely contributes to sex achievement gaps. This thesis recommends future research to disaggregate school engagement findings by sex and SEN diagnosis to better understand subgroup risks. Further research is needed to clarify the relationship between adolescent development and academic achievement and the longitudinal impact of teacher stereotype endorsement.

Acknowledgements

I would like to express my thanks and gratitude to my supervisors, Professor Gijsbert Stoet and Professor Reinhard Pekrun, for their patience, advice, and support during the last few years. Also, thank you to Dr Helge Gillmeister, Professor Marie Juanchich, and Dr Carina De Klerk for their oversight and guidance during and beyond their roles as Chairperson for the supervisory process. Finally, thank you to my family for their support and endurance during what have been challenging times for us all, but at the same time, have brought us great change and opportunity.

Contents

List of Figures

List of Tables

1 General Introduction.

Sex differences in educational achievement have been the subject of considerable research and debate. Despite the volume of research, it is unclear why sex differences change during adolescence / secondary education. Some observed changes, such as reported female advantages in STEM subject school grades, appear to contradict existing theories that predict male advantages in these subjects (e.g., Eccles et al., 1983). During secondary education, many students fall further behind, and others who achieved high grades at the end of primary school do not reach their potential (Allen, 2015; Montacute, 2018). The Sutton Trust, a UK social mobility charity, reports that up to 15% of previously high-achieving students are not among the upper 25% of achievers at age 16 (Allen, 2015; Montacute, 2018). These underachieving students are more likely to be boys, eligible for free school meals, or have additional educational needs (Allen, 2015; Ermisch & Bono, 2010; Montacute, 2018). The period from age 11 to 14 years old, encompassing the transition to secondary education, has been identified as a critical period for sharp declines in attitudes, mental health, and achievement (Jerrim & Palma Carvajal, 2025).

Generally, the gap in educational achievement favouring girls at the end of primary education increases during secondary school, and more girls go on to enter university (Cavaglia et al., 2020; Stoet & Geary, 2020a; Voyer & Voyer, 2014). These trends are common in most OECD countries (Cavaglia et al., 2020; Welmond & Gregory, 2021). Although some argue that boys' underachievement may not need special attention, the links between low educational achievement and criminality, as well as reduced marriage/cohabitation rates among low socioeconomic groups, suggest a social cost resulting from low male educational achievement (Autor, 2010; Lochner & Moretti, 2004; Machin et al., 2011). In general, increasing the number of years spent in education and increasing educational achievement is beneficial for health and wealth for all and, therefore, benefits the individual and the state (Department for Business, Innovation and Skills, 2013). Yes, despite these social, health, and economic benefits, there is little, if any, political focus on the educational underachievement of boys and, despite many calls for reforms in the provision and funding of special educational needs support, which disproportionately affects boys, little progress has been made (Hillman & Robinson, 2016; Welmond & Gregory, 2021). To be able to address the widening sex differences during secondary education, it is necessary to understand when and why achievement gaps change, the specific factors associated with underachievement in this age group, and how these may differ from those associated with underachievement in primary education. This thesis, therefore, examines longitudinal trajectories of school achievement from mid-primary school to mid-secondary school and how subject-specific trajectories change with age. Through secondary data analysis of a single cohort, this thesis examines the contribution of adolescent development to school achievement. It investigates the influence of puberty on school engagement as a potential mechanism to explain the reported correlation between puberty and educational achievement.

The general intelligence, or Spearman's g, of girls and boys is similar (Halpern & Wai, 2019; Halpern et al., 2007). Underlying g, there are some small differences in the specific cognitive abilities that contribute to sex differences in achievement. Girls tend to excel in verbal skills, and boys often outperform girls in spatial tasks, which contributes to sex differences in language, mathematics and some aspects of science, such as physics. Nevertheless, overall, sex differences in individual cognitive skills do not fully account for the sex differences observed in school subjects (Atit et al., 2020; Collado, 2019; Deary et al., 2007). The female advantage in language subjects is persistent and found across nations; however, sex differences in mathematics and science achievement are unclear (Stoet & Geary, 2015). The results of cross-national standardised tests such as the Programme for International School Assessments (PISA) report male advantages in mathematics and science in most countries (Baye & Monseur, 2016; Gray et al., 2019). In contrast, the Voyer and Voyer (2014) meta-analysis of school grades and the later updated and extended meta-analysis by O'Dea et al. (2018) report female advantages in mathematics and science. However, the majority of effect sizes analysed in both school grade meta-analyses are sourced from North American studies; therefore, findings may not generalise to other countries. For example, effect sizes in mathematics were smaller and sometimes not significant in samples from outside of North America (O'Dea et al., 2018; Voyer & Voyer, 2014). The variation of effect sizes between countries is not fully explained. Differences in science and mathematics curricula and how these subjects are tested or assessed between educational systems likely contribute, as well as national factors such as female economic participation or country-average mathematics achievement (Gray et al., 2019; Marsh et al., 2021; Stoet & Geary, 2015; Wu, 2010). Other evidence reports that sex differences in mathematics and science achievement are larger in the upper tail of mathematics and science distributions, and the opposite pattern is found in language/reading (Baye & Monseur, 2016; Stoet & Geary, 2013). Research that focuses solely on mean sex differences in educational achievement may, therefore, ignore larger differences reported in the upper and lower tails and the contribution of greater male variability to sex differences in academic outcomes (Baye & Monseur, 2016).

Greater male variability (GMV) of intelligence, or individual cognitive skills, may contribute to sex achievement gaps. The GMV hypothesis proposes that more boys/men are observed at the upper and lower tails of the intelligence distribution and more females are observed at the distribution's centre, resulting in the greater availability of men/boys at very high levels of *g* and with intellectual disabilities. If the GMV hypothesis is supported in educational outcomes, then we would expect to observe an overrepresentation of boys at the highest levels of achievement. Some support for GMV has been reported in school grades, with sex differences in grade variability being smaller in STEM subjects than in non-STEM subjects (O'Dea et al., 2018). GMV in *g* is supported in very large or population-sized samples, although evidence from twin studies found that sex differences in intelligence (both mean and variability) fluctuate or change with age (Arden & Plomin, 2006; Martin & Hoover, 1987). Any contribution of GMV to educational outcomes, therefore, may change depending on the age group tested. In their meta-analysis, O'Dea et al. (2018) found no significant association between age and variability in school grades, observing that sex differences in grade variability tended to reduce with age. However, I was unable to identify any research specifically addressing sex differences in grade variability in school grades. It remains unclear how or when changes in school grade variability contribute to changes in sex differences in achievement.

Sex differences in school grades fluctuate with age in subject-specific patterns. From Voyer and Voyer's (2014) meta-analysis, sex differences are widest in junior/middle school samples in mathematics and science and both junior/middle and high school samples in language subjects. Despite this, sex differences in mathematics and science school grades remain unclear, as there is a discrepancy between the reported North American female advantage in mathematics when compared with samples from other countries where sex differences in mathematics are smaller or non-existent (Voyer & Voyer, 2014). Additionally, few samples of science achievement were available for inclusion in Voyer and Voyer's (2014) meta-analysis. Similarly, why a female advantage emerges later in mathematics and science than in language subjects, and why this advantage is reduced in high school samples, is not fully explained. Students' enrolments in optional courses may influence STEM sex achievement gaps later in their education, and these may be influenced by perceptions of competency (e.g., Lindberg et al., 2010; Voyer & Voyer, 2014; Wang et al., 2013). Questions remain about the size of sex achievement gaps in mathematics and science, and also about when and why sex differences change during compulsory education. Puberty, a pivotal developmental stage, has been linked to fluctuations in academic achievement, although the mechanisms remain largely unexplored.

Sex educational achievement gaps are widest during early adolescence and into middle adolescence for global measures of achievement (e.g., GPA - grade point average) and in language subjects (Voyer & Voyer, 2014). Some evidence suggests that puberty and the wider changes associated with adolescent development may contribute to the academic underachievement of some groups of students during adolescence. It is often proposed, for example, that the earlier maturation of girls contributes to achievement gaps in this age group. Several studies have reported links between early puberty for girls and late puberty for boys, with underachievement in education (e.g., Borra et al., 2023; Dreber et al., 2012; Koerselman & Pekkarinen, 2017; Pekkarinen, 2008; Torvik et al., 2021). At present, though, the evidence is mixed - both positive and negative effects have been reported, sometimes effects are not found, and associations may be domain-specific, context-dependent, or indirect via motivation (see Laube & Fuhrmann, 2020, for a review). Puberty, the reproductive maturation process, is one aspect of adolescent development that develops along a separate trajectory to the socio-emotional and behavioural transitions also encompassed within this period (Sisk & Foster, 2004). Although they are separate processes, puberty and adolescent development are inextricably linked through hormonal interactions (Sisk & Foster, 2004). Evidence explaining the mechanism(s) through which puberty is related to academic achievement is currently sparse. However, initial findings suggest an association between puberty and executive function development, and also with engagement with learning (Chaku & Hoyt, 2019; Martin et al., 2022; van Tetering et al., 2020, 2022).

1.1 Outstanding Issues and Research Questions

Through secondary data analysis of a large UK longitudinal survey combined with national educational outcomes, Study 1 examines trajectories of sex achievement gaps in English, mathematics, and science at ages 7, 11, and 16 to answer the research question: how do subject-specific sex differences in school achievement change with age (Research Question (RQ1)? This study will ask whether a female advantage is found during middle adolescence (RQ2), and in which subjects this advantage is found (Voyer & Voyer, 2014), where early adolescence corresponds to ages 11-14 years, middle adolescence to ages 15-17 years, and late adolescence to ages 18-21 years (Muyeed, 2008). These are, however, broad categorisations that do not account for individual differences in developmental trajectories (Muyeed, 2008). The reasons why sex achievement gaps widen between early and mid-adolescence are unclear. This thesis will also clarify sex differences in STEM subjects in each age group (RQ3), where country differences have been reported, and where relatively few studies have examined sex differences in science subjects, particularly individual sciences such as chemistry, physics, and biology (Voyer & Voyer, 2014). Much of the existing literature takes a cross-sectional approach, and existing longitudinal studies examine shorter periods (Voyer & Voyer, 2014). By analysing changes in a single cohort from childhood to mid-adolescence, this thesis excludes the contribution of differences in sample demographics and between education systems to the conclusions drawn. Using a sample from an education system that uses a national, standardised approach to curriculum-based school grades, this thesis asks whether sex differences in achievement differ between standardised school grades and teacher-assessed school grades (RQ4). It is reported that teacher grading biases contribute to the sex differences in school achievement (e.g., Lavy & Megalokonomou, 2019). Therefore, this thesis will clarify whether the reported sex differences are found irrespective of the mechanism through which grades are assessed.

In the context of the psychobiosocial model of mathematics and science achievement (Halpern, 2012), the first study also examines if GMV contributes to sex differences in the upper and lower tails of grade distributions. Sex differences at the upper and lower end of the school grade distribution have not been examined in detail to date, except for O'Dea et al. (2018), who used a meta-analytical approach using the standard deviations reported in prior research. This thesis will therefore examine the proportion of girls and boys represented at the higher and lower levels of the achievement distribution, to ask whether GMV is found in school grades and how this contribution changes with age (RQ5) (Arden & Plomin, 2006; O'Dea et al., 2018). As the school grades literature has been criticised for its focus on mean differences, we compare both mean differences and sex differences at the upper and lower achievement levels against reports from international standardised tests to ask whether the findings from these ability and achievement testing domains align (RQ6) (Baye & Monseur, 2016).

With a focus on widening sex differences in achievement during adolescence, this thesis then examines the reported association between puberty and educational achievement to advance the discussion in this area further. The current literature linking puberty and achievement, and the proposal that the early maturation of girls contributes to the widening achievement gaps, is limited by conflicting results and little evidence demonstrating a female advantage due to earlier maturation. In general, few studies have examined links between puberty and boys' achievement at school (Deardorff et al., 2019). Using longitudinal, multilevel growth models, this thesis asks how school engagement and academic motivation develop from late childhood/early adolescence to middle adolescence and whether puberty, age or a combination of these is associated with the changes during this period (Martin et al., 2017, 2022). Operationalising puberty status as a continuous measure, rather than distinguishing between early and late maturers in comparison to peers, and examining puberty change longitudinally, allows this thesis to address the criticisms of the current body of literature. Both the pubertal timing and the speed of change are relevant when examining puberty outcomes (Mendle & Koch, 2019). Finally, this thesis asks whether perceived puberty status predicts changes in school engagement during adolescence (RQ7) and whether these changes contribute to the female advantage reported in mid-adolescence (RQ8). By examining boys and girls, this thesis adds to the limited literature examining boys' maturation and school achievement. This thesis adds to the very limited number of studies examining possible mechanisms that may explain why puberty and achievement may be related.

1.2 Literature Review

This section aims to provide a comprehensive overview of existing research relevant to the study's objectives and research questions, with the primary aim of contextualising the research. With a focus on Study 1, this literature review situates the study within the academic discourse on sex differences in educational achievement, giving an overview of the broader set of biopsychosocial theories of sex differences in STEM achievement. The first part of the review, therefore, focuses on Halpern's psychobiosocial model, which also encompasses the influence of developmental change to explain why sex differences in mathematics often favour girls in early childhood but then favour boys in older age groups. In general, biopsychosocial theories would predict that once the male advantage in STEM subjects is observed, this would tend to persist or widen with age (Halpern et al., 2007). However, female advantages are observed in STEM school grades in early to middle adolescence that contradict the predictions of these theories (Voyer & Voyer, 2014).

Given that a general widening female educational advantage is reported during early to middle adolescence, and that the earlier maturation of girls is often indicated in general discourse as a contributing factor, the remainder of the literature is a systematic review of the evidence linking puberty with academic motivation, engagement and educational achievement in support of Study 2. Building on two prior reviews of the links between puberty and a range of outcomes for boys and girls (Mendle & Ferrero, 2012; Mendle et al., 2007), a systematic search of Web Of Science and Google Scholar identified five additional studies beyond those reviewed previously. Using keywords and synonyms for educational achievement, puberty, and adolescence, all the identified evidence is reviewed in this section. The review examines two proposed mechanisms through which puberty may be linked to academic achievement -

executive function development and changes in academic motivation and engagement (Chaku & Hoyt, 2019; Martin et al., 2017, 2022).

The final section of this literature review presents theories of academic motivation and school engagement, highlighting their theoretical links with developmental change and school transitions. Particularly, this section focuses on aspects of the situated expectancy value theory that proposes social, environmental, and individual influences on academic motivation and either encompasses or is linked to the variables of interest in this thesis - academic self-concept, educational expectations as measures of academic motivation (Eccles & Wigfield, 2020). School engagement is also reviewed, which, with academic motivation, is linked to developmental change through stage environment fit theory, which predicts declining academic motivation and engagement as the school environment is less able to meet developmental changes in the psychological needs of the individual (Eccles & Roeser, 2013; Eccles et al., 1993; Ryan & Deci, 2000b). Stage environment fit theory is of particular relevance due to its link to school transitions, as each cohort member will transition from primary to secondary education in early adolescence, and poor transitions may contribute to lower motivation and engagement for some individuals during this period (Eccles et al., 1993). The self-systems model of motivational development is briefly highlighted as this model draws together associations between perceptions of the school environment and self-appraisals, including self-concepts, which lead to school engagement or disengagement (Skinner et al., 2008, 2009). Together, these models are linked with the formation of self-identity, a key process of adolescence, for which meeting the basic psychological needs of the individual is seen as fundamental (Hansen & Jessop, 2017; Ryan & Deci, 2000b).

1.2.1 Biopsychosocial Models of Sex Differences in Mathematics and Science Achievement

Several theoretical models have been proposed that aim to explain sex differences in mathematics achievement or, more broadly, across STEM fields. Each differs slightly in how they emphasise or de-emphasise sex differences in individual cognitive abilities, particularly mathematics and spatial skills. Several models highlight that there is a significant overlap in cognitive abilities and that, on average, sex differences are very small (Ceci, 2017; Ceci & Williams, 2010; Hyde, 2005). Therefore, it is sex differences in ability beliefs, gender stereotypes, and cultural socialisation processes that underlie many of the sex differences reported in mathematics and science achievement, the choice to enter further study in STEM subjects or enter into STEM careers (Ceci, 2017; Ceci & Williams, 2010; Else-Quest et al., 2010; Hyde, 2005). Other theorists emphasise biological differences, including the influence of greater male variability in psychological traits that results in larger differences at the tails of intelligence or cognitive skills distributions. These theories de-emphasise the role of gender stereotypes due to a lack of evidence (Stoet & Geary, 2013, 2018). Instead, it is argued that sex differences in patterns of ability differ, with girls' abilities tilted towards verbal skills and boys' towards mathematics skills in combination with girls' innate preferences toward people-oriented and altruistic choices, that contribute to subject and career choice (Casey & Ganley, 2021; Stoet & Geary, 2013, 2020a, 2020b; Wang et al., 2013; Wigfield & Eccles, 2000). These explanations highlight that girls who are mathematically capable but have high verbal skills are less likely to consider STEM subjects and careers. Girls with high mathematics and moderate verbal abilities are more open to choosing STEM (Stoet & Geary, 2018, 2020a; Wang et al., 2013). Stoet and Geary (2018) also reported that girls were less likely to choose STEM careers in countries with higher gender equality, where STEM engagement for women is actively promoted. In contexts where women have more choices and students are encouraged to follow academic paths based on their strengths, girls will be more likely to choose non-STEM paths because there are more girls with an ability tilt towards the verbal domain than the mathematics domain (Cuff, 2017; Stoet & Geary, 2018; Wang et al., 2013). Recognising many of these influences on STEM achievement, Casey and Ganley (2021) argues that it is sex differences in a verbal vs. spatial skills tilt that contribute to mathematics success. Halpern's psychobiosocial model also includes the influence of developmental change to explain why sex differences in mathematics often favour girls earlier in education and boys in older children and adolescents (Halpern et al., 2007; Miller & Halpern, 2014).

Halpern's psychobiosocial model emphasises the reciprocity between nature and nurture, which are described as inseparable influences that enhance or reduce the effects of small biological differences (Halpern, 2012; Halpern et al., 2004, 2007; Miller & Halpern, 2014). Focused on mathematics and science achievement, this model describes how social and environmental contexts interact with biological differences and developmental changes to influence outcomes (Halpern et al., 2007; Miller & Halpern, 2014). Biological predispositions that may result from the influence of sex hormones in the prenatal period, such as the male advantages in visual-spatial tasks, make it easier to learn tasks that require these skills (Halpern & Wai, 2019; Miller & Halpern, 2014). Positive learning experiences using these skills result in greater interest, motivation, and expectancies of success (Eccles et al., 1983; Frenzel et al., 2007). In addition, social factors support and encourage engagement and perseverance in domains or subjects that the individual is perceived to be predisposed to or perhaps is encouraged to persevere with. Together, psychological, biological, and social influences work to maintain sex differences (Halpern & Wai, 2019). Biological factors alone cannot explain the sex differences reported in school grades (e.g., Atit et al., 2020; Collado, 2019; Deary et al., 2007; Voyer & Voyer, 2014). Socio-environmental factors can emphasise small cognitive differences resulting in increasing sex differences over time, but also have the potential to minimise them and reduce sex differences (Halpern et al., 2007; Miller & Halpern, 2014).

Biological Differences. The biological influences of the psychobiosocial model encompass sex differences in specific cognitive skills and the GMV hypothesis, which proposes that males demonstrate greater variability between individuals in physical and psychological characteristics, including intelligence (Ellis, 1934; Feingold, 1994). The psychobiosocial model describes how sex differences in mathematics ability often favour girls in early childhood (Halpern et al., 2004). In this age group, mathematics problems tend to be more computational and processing speed is relevant, supported by small female advantages in encoding and retrieval from long-term memory (Herlitz & Rehnman, 2008; Hyde, 2005). For the remainder of primary education, few, if any, sex differences in mathematics ability are reported, but later and particularly in secondary education, solving mathematics problems requires visual-spatial representations in working memory (Halpern et al., 2004; Kaufman, 2007; Miller & Halpern, 2014; Voyer et al., 2017). Sex differences in visual-spatial working memory begin to emerge at around 13 years old, in line with the emergence of sex differences in specific visual-spatial tasks. This small to moderate difference was not statistically significant until late adolescence (Voyer et al., 1995, 2017). Supporting the GMV hypothesis, larger sex differences are reported at the tails of intelligence and standardised tests (Baye & Monseur, 2016; Deary et al., 2003; Gray et al., 2019; Strand et al., 2006). However, less support for the contribution of GMV is found in the meta-analysis of school grades variability (O'Dea et al., 2018). Sex differences in the cognitive ability tests that underlie general intelligence indicate that there are more boys/men at the upper and lower tails of some quantitative and spatial abilities, but also more boys/men at the lower tail due to male overrepresentation among those with intellectual disabilities (Halpern & Wai, 2019; Johnson et al., 2008; Voyer et al., 1995, 2017). Halpern's and other biopsychosocial models propose that these sex differences in cognitive skills contribute to sex differences in mathematics achievement (Casey & Ganley, 2021; Geary et al., 2023; Halpern & Wai, 2019; Halpern et al., 2004), and this contribution would be expected to increase with age through continued positive experience and approach.

How sex differences in variability at the upper tail of the intelligence distribution contribute to sex differences at high levels of educational achievement has been debated (Johnson et al., 2008). Sex differences at the tails of the achievement distributions differ by domain, for example. In standardised tests, sex differences are larger in the upper tail in mathematics and science and at the lower tail in reading, but are most pronounced at the lower tail in reading (Baye & Monseur, 2016; Stoet & Geary, 2013). The overrepresentation of boys (2.5 times more boys) in the upper 5% of mathematics has reduced over time, despite being stable for many years (Baye & Monseur, 2016; Benbow, 1988; Makel et al., 2016; Stoet & Geary, 2013; Wai et al., 2010). In contrast, the trend at the lower tail in reading over four rounds of PISA tests has increased from 2.5 to 3.2 times more boys in the lower 5%, and from 3.1 to 5.9 in the lower 1% (Stoet & Geary, 2013). The influence of GMV in standardised tests is not found consistently across nations and may be influenced by factors such as female economic participation (Gray et al., 2019; Machin & Pekkarinen, 2008). Nevertheless, sex differences in intelligence variability are found in young children prior to any educational influence. In a single, longitudinal twins sample, girls, on average, achieved higher scores in intelligence tests of preschool-aged children and were overrepresented at the upper tail from age 2 to 4 years old (Arden & Plomin, 2006). The overrepresentation of girls at the upper tail reduced from 1.8 times more girls at age 2 to 1.2 times more girls than boys by age 4. Later, by age 10, boys were overrepresented at the upper tail as predicted by the GMV hypothesis, observing 1.7 times more boys than girls. GMV was supported in the sample from age 3 until age 10 (Arden & Plomin, 2006) and in 11-year-olds (Johnson et al., 2008). Arden and Plomin (2006) suggested that this may reflect that girls develop at a faster rate than boys during childhood, for example, in language development (e.g., Fenson et al., 1994). Brain growth patterns may differ between boys and girls, but the evidence is sparse in young children, and many samples are cross-sectional or analyse the scans of small samples (Giedd et al., 1999; Lenroot et al., 2007). Therefore, developmental changes, as well as environmental factors, may contribute to sex differences in achievement in children and may also influence inter-individual variability. However, currently, there is limited evidence linking brain development and changes in intelligence during early childhood.

While not addressed by the psychobiosocial model, female advantages are commonly reported in tests of language and verbal skills (Halpern & Wai, 2019; Hedges & Nowell, 1995; Paus et al., 2017). Further, there are moderate to large female advantages reported in reading and writing (Baye & Monseur, 2016; Reynolds et al., 2015; Stoet & Geary, 2013). Despite these findings being common, little attention is given to closing reading and writing achievement gaps between boys and girls, despite the implications these differences have for achievement at school across many subjects (Caponera et al., 2016; GL Assessment, 2020). As described by the psychobiosocial model, both mathematics and language achievement may be influenced by sex differences in underlying skills that make learning in each domain easier, and these effects are amplified by continuing success and approach. However, social factors may further influence sex differences in each domain - either positively, through support and encouragement, or negatively, through low expectations (Eccles et al., 1983; Halpern et al., 2007; Miller & Halpern, 2014). For example, adults who endorse gender stereotypes may have high expectations of boys in STEM subjects and therefore encourage boys in these subjects. However, they may have low expectations of, and give less encouragement to, boys in language-related and other non-STEM subjects. In contrast, the endorsement of gender stereotypes may mean that girls are encouraged and supported in language-related subjects, but receive less encouragement in STEM subjects due to low expectations. Gender stereotypes and gendered social roles, and how these are endorsed by both children and the adults with whom they regularly interact, may result in positive or negative influences on achievement in each domain.

Gender Stereotypes. Education-relevant stereotypes can be subject-specific or more general, reflecting expected attitudes, engagement, and behaviour toward learning. Girls are stereotyped to be less able in mathematics and science subjects, and boys to be less able in language subjects (Alan et al., 2018; Kurtz-Costes et al., 2014; Retelsdorf et al., 2015). The endorsement of subject stereotypes and their transmission to children is found in both parents and teachers (Alan et al., 2018; Kurtz-Costes et al., 2014; Retelsdorf et al., 2015). Beyond the subject-specific stereotypes, parents expect girls to be better students than boys from an early age and matching beliefs are found in children (Åhslund & Boström, 2018; Hartley & Sutton, 2013; Latsch & Hannover, 2014; Wong et al., 2023). Teachers and peers within school settings can also promote gender stereotypes influencing subject choices and achievement (Alan et al., 2018; Carlana, 2019; Retelsdorf et al., 2015; Skipper & Fox, 2022). Additionally, stereotypical beliefs about boys' misbehaviour may also result in increased monitoring of boys in the classroom and boys being treated differently as a result (Åhslund & Boström, 2018; Skipper & Fox, 2022). Exposure to media discourse on the relative underachievement of boys compared to girls may itself result in the perpetuation of gender stereotypes through both the improved achievement of boys and girls in stereotypical domains and impaired performance in non-stereotypical domains, resulting in the alignment of learning goals to the stereotypes (Latsch & Hannover, 2014).

Awareness of gender stereotypes that favour boys in mathematics and science and girls in language abilities is found early in education (Cvencek et al., 2011; Shenouda, 2014). Between 7-8 years of age, both boys and girls endorse stereotypes that boys exhibit poorer conduct and achievement than girls, with the awareness of these stereotypes resulting in poorer performance in boys (Hartley & Sutton, 2013). Yet, individuals differ in their susceptibility to gender stereotypes and gendered social roles, which can depend on how strongly they identify with their gender identity, whether their parents or other influential adults in their social circle endorse or reject the stereotypes, or their peer group membership, and their self-esteem (e.g., Crouter et al., 2007; Santos et al., 2013; Tomasetto et al., 2011; Yu, 2019). Age differences in the effects of gender stereotypes have also been reported (Crouter et al., 2007; Hill & Lynch, 1983; Kurtz-Costes et al., 2014). Some research has found that the endorsement of gendered role expectations can increase during adolescence (Hill & Lynch, 1983). Kurtz-Costes et al.

(2014) examined how subject-specific stereotypes changed with age, finding that the endorsement of language stereotypes increased with age from late childhood to early adolescence. Mathematics and science stereotypes, though, were more strongly endorsed in late childhood and were more neutral in older children. The adolescent's perceptions of adult stereotypes matched this pattern (Kurtz-Costes et al., 2014). Kurtz-Costes et al. (2014) concluded that both children and adolescents were influenced by their perceptions of adult gender stereotypes and the comparative achievement of boys and girls they observe at school. Other research has reported that there are substantial within-group differences in gender stereotype endorsement, depending on individual circumstances, which included parents' beliefs, the sex of siblings, and birth order, such that the gendered beliefs of some individuals will increase with age, but will decrease for others (Crouter et al., 2007). This perspective aligns with research that links the academic underachievement of some groups of boys and girls to peer group membership and the alignment of those peer groups to gendered social roles, resulting in patterns of motivation, engagement and school achievement (Santos et al., 2013; Yu, 2019).

The biopsychosocial models would predict that once sex differences in achievement have emerged, achievement gaps will persist or widen in the stereotypical domains of mathematics and science achievement for boys and language achievement for girls. Small biological differences in cognitive skills are amplified over time through psychosocial processes. While meta-analyses of large-scale standardised test results match these expected patterns, meta-analyses of school grades oppose these predictions for mathematics and science achievement during early to mid-adolescence (Baye & Monseur, 2016; O'Dea et al., 2018; Stoet & Geary, 2013; Voyer & Voyer, 2014). Sex differences then reduce in older, more selective samples towards the end of secondary school and into tertiary education. That there appears to be a general widening of female advantage in educational achievement during early-mid adolescence may point to influences specific to the adolescent development period that influence educational outcomes that result in wider female advantages in this age group.

1.2.2 Puberty and Educational Achievement. A Review of the Current Evidence

The earlier maturation of girls during adolescence is often cited as a contributing factor to the general female educational advantage during the adolescent years, possibly influencing education or career-related decision-making, and often with the assumption that boys will catch up later (APPG, Men & Boys, 2023; Pekkarinen, 2008; Tinklin et al., 2001; Torvik et al., 2021). In line with this, smaller sex differences are observed in high school and tertiary education samples (Voyer & Voyer, 2014). However, selection effects will play a part here, as more boys leave full-time education at age 16 to enter apprenticeships or traineeships, and more girls enter university (Cavaglia et al., 2020). The resulting set of students in older samples is more highly motivated than those that include students from across the achievement distribution (Cavaglia et al., 2020; Voyer & Voyer, 2014). The assumption that boys catch up later has also been challenged (Stoet, 2015; Tinklin, 2003).

Intuitively, the earlier average commencement and completion of pubertal maturation for girls and the widening sex difference in academic achievement would appear to align, and links between pubertal maturation and educational outcomes have been reported. However, the evidence linking puberty to school achievement is mixed, and direct puberty-related effects on most outcomes tend to be very small (Smith-Woolley et al., 2017). Pubertal development is one aspect of the broader changes that occur during adolescence. Adolescent development is a set of interrelated processes that encompasses physical, cognitive, and socioemotional change through which children become adults (Sisk & Foster, 2004). The physical changes associated with adolescence include growth spurts and the development of secondary sex characteristics, for example, facial hair growth and voice changes in boys, breast growth and menstruation in girls (Mendle & Ferrero, 2012; Mendle et al., 2007). While these changes and levels of sex hormones, estradiol and testosterone, are commonly associated with sexual maturation, the behavioural and cognitive changes of adolescence are temporally coordinated and interact with hormonal changes (Sisk & Foster, 2004). Cognitive changes result from an extended period of brain reorganisation that continues into early adulthood; these changes are experience-dependent and associated with elevated activation in reward-regions of the brain and linearly maturing cognitive control systems resulting in an adaptive tendency towards increased risk-taking during middle adolescence (Crone & Dahl, 2012; Spear, 2000; Spear, 2013). The socioemotional changes associated with adolescence include an improved understanding of social situations and perspective-taking abilities, the increased salience of peers, a need to feel socially accepted and avoid rejection, and self-identity formation (Crone & Dahl, 2012). Each measure used in the puberty-educational achievement literature may, therefore, encompass different aspects of the changes associated with this developmental period (Mendle & Ferrero, 2012).

Much of the current evidence focuses on within-peer-group differences, categorising individuals as being early, on-time, or late maturers in comparison to their peers rather than using continuous measures of maturation in longitudinal models (Deardorff et al., 2019). These studies link early or late puberty to lower educational achievement, but results can differ depending on which measure of maturation is used. Other research indicates that both the timing (commencement of maturation) and tempo (speed of maturation) are important for outcomes. Typically, puberty completes in four years, but there are substantial individual differences resulting in a tempo as short as one year or as long as seven years (Mendle, 2014). A quicker tempo may be more closely associated with poor outcomes, and a slower tempo may be less noticeable and more easily adapted to (Mendle & Ferrero, 2012; Mendle et al., 2010). Currently, much of the evidence points to an increased risk of poor outcomes for early maturers, more consistently for early-maturing girls. This evidence is not suggestive of a maturity advantage for girls' school achievement. However, with the current evidence's predominant focus on within-sex relative maturity and the increased risks for a small number of early maturers, the body of evidence may say little about the majority, or how boys' and girls' differing maturity trajectories contribute to sex differences in school achievement. There is also less evidence, in general, for boys.

Several hypotheses have been proposed to account for the poor outcomes associated with early pubertal timing (Mendle & Ferrero, 2012; Mendle et al., 2007). Among these, most support is found for the early-timing (or developmental readiness) hypothesis, which proposes that individuals who experience puberty earlier than their peers are not developmentally ready for the demands of this transition (Stattin & Magnusson, 1990). The consequences of early timing may be more problematic for adolescents who already exhibit poor behavioural or emotional control, as individual differences are magnified during this period (accentuation hypothesis: Caspi & Moffitt, 1991). Girls who mature earlier tend to start dating earlier and report that adults expect more mature behaviours from them (Simmons & Blyth, 1987). Early maturing girls tend to associate with older peers, exposing them to situations they are not developmentally ready for, leading to problem behaviours (Eichorn, 1975; Magnusson et al., 1985). Therefore, the developmental readiness hypothesis proposes that the cognitive and emotional development of early maturers lags behind their hormonal and physical development, leaving them at risk of emotional and behavioural disorders and risky behaviours that can persist into adulthood (Ge et al., 1996; Graber et al., 2004). An alternate explanation for the association between puberty and poor outcomes is the maturational deviance hypothesis, which posits that the risk of poor outcomes is associated with deviation from the norm. The maturational deviance hypothesis (or the off-time hypothesis) predicts poor outcomes for early and late maturers (Brooks-Gunn et al., 1985; Petersen & Taylor, 1980). It is also proposed that regardless of timing, all individuals will exhibit some stress during maturational change, and this stress will be most pronounced during the period of greatest change (stressful change hypothesis: Simmons & Blyth, 1987).

More evidence supports the early-timing hypothesis when examining the association between puberty and outcomes including substance use, depression, anxiety, externalising behaviours, and poor academic achievement (Castellanos-Ryan et al., 2013; Dimler & Natsuaki, 2015; Hoyt et al., 2020; Mendle & Ferrero, 2012; Mendle et al., 2007, 2010; Smith-Woolley et al., 2017). However, the majority of the evidence examines female pubertal development, which is more easily measured. There is less evidence for boys, and the findings are mixed, providing some support for the maturation deviance hypothesis or the stressful change hypothesis regarding internalising and externalising behaviours and substance use (Castellanos-Ryan et al., 2013; Deardorff et al., 2019; Dimler & Natsuaki, 2015; Ge et al., 2003; Marceau et al., 2011; Mendle & Ferrero, 2012).

Direct Associations of Puberty with Educational Achievement. The connection between puberty and educational achievement is unclear in the existing literature. There are varying reports of sex differences, and the associations seem to depend on the specific measures used to represent pubertal development and academic achievement. Some studies suggest that later development is linked to higher academic achievement for girls (Torvik et al., 2021) and lower achievement for boys (Dubas et al., 1991; Mendle & Ferrero, 2012). However, other studies have found no association with achievement (Koivusilta & Rimpelä, 2004; Senia et al., 2018) or have identified subject-specific associations (Suutela et al., 2022). Advanced development has been associated with the increasing gender gap in mathematics in standardised tests (Borra et al., 2023). Paradoxically, advanced development has also been linked to higher mathematics school achievement in girls (Suutela et al., 2022). Interestingly, these contradictory results are mirrored when comparing sex differences in mathematics standardised tests and school grades (Baye & Monseur, 2016; Voyer & Voyer, 2014).

For girls, the association between development and higher academic achievement may be specific to advanced biological development. For example, earlier menarche (the commencement of menstruation) and puberty hormones have been associated with lower educational achievement (Cavanagh et al., 2007; Martin et al., 2022; Torvik et al., 2021). In contrast, later age at peak height velocity (growth spurt) has been associated with mathematics achievement but not native language achievement (Suutela et al., 2022). As for boys, early or average spermarche (age at first ejaculation) has been associated with higher school achievement, but the usefulness of spermarche as a measure of relative male maturation has been questioned (Koivusilta & Rimpelä, 2004; Mendle & Koch, 2019). Recently, Suutela et al. (2022) reported that there was no association between male peak height velocity and mathematics or native language achievement, contradicting an earlier study that linked earlier peak height velocity with higher grades for boys (Dubas et al., 1991). Despite the effects of puberty timing being studied for many years, a number of factors may contribute to these inconsistent results. Puberty is one aspect of the wider changes in adolescent development. The measures used may, therefore, capture different aspects of the broader changes of adolescence. It may also matter who assesses or reports the developmental status of the individual, and associations between puberty and outcomes may be indirect.

Studies have reported cognitive benefits for earlier maturation, such as improved attention for both boys and girls (Chaku & Hoyt, 2019). However, earlier maturation is also linked to lower self-control for girls (Chaku & Hoyt, 2019). These seemingly contradictory results align with dual systems theories of adolescent risk-taking, which suggest a gap in the developmental trajectories of self-regulation and sensation seeking (Steinberg et al., 2008, 2018). Dual systems adolescence theories predict that early maturers are at an increased risk of engaging in risky behaviours (Shulman et al., 2016; Steinberg et al., 2018). Pointing to evidence that self-regulation increases linearly with age, while the trajectory of sensation seeking follows an inverted U-shaped pattern, peaking during early to mid-adolescence, these theories propose that there is a developmental gap between self-control and sensation seeking that puts early maturers at an increased risk of making poor choices (Shulman et al., 2016; Steinberg et al., 2018). Even so, adolescence is also an important period for self-identity development and changes in social behaviour, and these changes influence decision-making and self-appraisals through an orientation towards social rewards and others' perceptions of them (Blakemore, 2010; Nelson et al., 2016; Pfeifer & Berkman, 2018).

While some longitudinal surveys utilise medical professionals to evaluate pubertal development (Borra et al., 2023; Koerselman & Pekkarinen, 2017; Torvik et al., 2021), others use self-reported or parent-reported measures (Martin et al., 2017; Senia et al., 2018), and associations may differ between raters (Dorn et al., 2003). Self-reported measures of pubertal development, e.g., pubertal development scale (PDS: Petersen et al., 1988), have been found to be less reliable indicators of timing. Nevertheless, they may be indicative of social comparison processes, an important aspect of adolescent development, and are a valid measure of change for use in longitudinal analyses (Brooks-Gunn et al., 1987; Caspi & Moffitt, 1991; Coleman & Coleman, 2002). Comparisons of the PDS and assessments using Tanner stage pictures find a strong correlation (.80) between the two measures (Marshall & Tanner, 1969, 1970; Paikoff & Brooks-Gunn, 1991). A recent longitudinal study compared both biological (Tanner stages: Marshall & Tanner, 1969, 1970) and self-reported pubertal development with academic achievement, finding that the results differ between the two measures (Goering et al., 2023). Boys' perceived off-time maturation predicted lower academic achievement throughout adolescence, and early perceived maturation predicted lower concurrent achievement in girls (Goering et al., 2023). Different effects for perceived and biological pubertal timing have been reported elsewhere in other contexts (Carter et al., 2017, 2018; Goering & Mrug, 2022; Moore et al., 2014). In summary, there is inconsistency in the current literature between the sexes, between the types of measures used, and between self-report, medical assessments, and biological measures. Focusing on possible mechanisms through which puberty and school achievement are related may help to clarify why there are differences between the measures, as well as explain how and why pubertal maturation and achievement are associated. Currently, changes in hormone levels have been linked to executive function development, which links to cognitive changes in adolescence (Chaku & Hoyt, 2019). Perceived puberty and sex hormones have been associated with academic motivation and engagement, which may link to socioemotional change, for example, peer pressure or changes in self-appraisals (Martin et al., 2017, 2022). However, as associations differ between puberty measures, further evidence is needed. Additional mechanisms that may be influenced by peer-related processes or self-identity formation, for example, could be explored before the links between puberty and achievement can be fully understood.

Indirect Associations of Puberty with Educational Achievement. There is substantial evidence linking puberty with psychological distress and risky or externalising behaviours and their subsequent influence on educational outcomes (Becherer et al., 2021; Dimler & Natsuaki, 2015; Hallfors et al., 2002; Ullsperger & Nikolas, 2017). There is less evidence that explains the mechanisms through which puberty is linked to achievement and how this might influence sex achievement gaps during adolescence. Recent evidence links puberty with executive function development in boys and girls. Executive functions are a set of skills that develop throughout childhood and adolescence, enabling the regulation of attention, self-control, monitoring, planning, and goal-directed behaviour (Diamond, 2013; van Tetering et al., 2020). Early pubertal maturation is associated with faster increases in attention skills during adolescence for boys and girls, while also predicting lower self-control in girls (Chaku & Hoyt, 2019). Lower self-control suggests that early maturers are more sensitive to immediate rewards (Peper & Dahl, 2013), and is linked to substance use and externalising behaviours such as truancy and behaviour problems, which in turn are associated with lower academic achievement (Gottfried, 2009; Hoffmann, 2020; van Tetering et al., 2022). While a recent review found little evidence of sex differences in executive function that persist into adulthood, there may be differences in developmental trajectories (Grissom & Reyes, 2019). Aligning with this perspective, two recent studies examining self-reported executive functions in

10-12-year-olds and 13-16-year-olds found that girls in the older age group reported higher executive functions than boys, but there were no sex differences in the younger age group (van Tetering et al., 2020, 2022). As girls report higher executive functioning than boys during the middle adolescence period, this may indicate that the lower (perceived) executive functioning of boys contributes to sex differences in educational achievement during this period (van Tetering et al., 2022). The links between advanced maturity and improved attention skills would support reports of an academic advantage for more mature boys (Dubas et al., 1991; Koivusilta & Rimpelä, 2004; Torvik et al., 2021). Nevertheless, despite these findings, the differences between typically developing students and off-time maturers are very small. They are unlikely to explain much of the variance in sex differences during adolescence (Koerselman & Pekkarinen, 2018).

Two recent studies have examined the association between puberty hormones, self-reported pubertal development, academic motivation (academic self-efficacy and valuing of school) and achievement, identifying potentially different mechanisms for boys and girls (Martin et al., 2017, 2022). The initial two-wave longitudinal study reported an indirect association between advanced self-reported pubertal status and lower academic achievement via academic motivation. This initial study also indicated that puberty was more important than age in predicting academic motivation (Martin et al., 2017). The second four-wave study examined how academic engagement and disengagement trajectories were influenced by puberty hormones and academic motivation (Martin et al., 2022). Puberty hormones were more influential for disengagement than for engagement and were more consistently associated with disengagement for boys (Martin et al., 2022). Together, these studies indicate a developmental influence on school engagement and academic motivation over and above the influence of prior achievement, socioeconomic status, and age, providing support for biopsychosocial perspectives. They imply a puberty-specific role in established motivational theories such as the expectancy-value theory, stage-environment fit theory, and the self-systems model of motivational development (Eccles & Wigfield, 2020; Eccles et al., 1983). However, these studies point to lower motivation and school engagement for more mature individuals, which opposes the maturity advantage in executive function development for boys (Chaku & Hoyt, 2019). Nevertheless, these findings align with studies that reported that early puberty is a risk factor for psychological distress and risky behaviours for boys and girls (Mendle & Ferrero, 2012; Mendle et al., 2007) These studies also underline that associations differ between measures and that further research is needed to disentangle these apparent contradictions depending on the type of measure used (Martin et al., 2017, 2022).

1.2.3 Situated Expectancy-Value Theory

Situated expectancy-value theory is a well-established model describing how a student's expectancy of success influences achievement-related choices, persistence, and performance (Eccles & Wigfield, 2020). The academic motivation model describes how ability beliefs and the extent to which a student values an activity are each developed under the influence of a student's social and cultural environment (Eccles & Wigfield, 2020; Eccles et al., 1983). An individual needs to believe that a task or achievement within a domain is achievable, and it is important to be motivated to engage in the effort required to achieve. Expectancies for success are, therefore, a student's evaluative belief about how well they will do in an upcoming task given their perceived competence, which, when combined with the subjective value, has been linked to student choices, effort, and academic achievement (Eccles & Wigfield, 2020; Eccles et al., 1983). Expectancies for success and competence beliefs, while originally conceptualised as distinct concepts, have been shown to be empirically indistinguishable within a given domain (Eccles & Wigfield, 1995; Wigfield, 1994), and are often measured and tested as domain-specific self-concept (Eccles & Wigfield, 2020; Marsh et al., 1984, 1991). Task values and ability self-concepts are empirically distinct but positively related. For example, a student who is good at mathematics will value mathematics and devalue the subject if they find it to be more challenging.

Task values are defined as four distinct concepts (Eccles & Wigfield, 2020; Eccles et al., 1983; Wigfield, 1994). Attainment value is described as the importance of doing well, including whether success in a given task confirms aspects of the individual's identity. Intrinsic value, or subjective interest in a task or domain, is similar to intrinsic motivation from self-determination theory (Deci & Ryan, 1985; Ryan & Deci, 2000a; Wigfield, 1994). Students with high intrinsic value for a given domain receive a positive psychological reward from engagement with that task/domain. Utility value is an evaluation of the usefulness of the subject given future goals and is similar to extrinsic motivation as described by self-determination theory (Deci & Ryan, 1985; Ryan & Deci, 2000a; Wigfield, 1994). Finally, cost or task difficulty describes the student's evaluation of the negative consequences of engaging in the task, for example, performance anxiety, opportunity costs, and the amount of effort required to succeed. A final construct within the expectancy-value model is achievement goals, which include long-term goals, such as needing to do well to support future educational plans or career choices, but also a student's motivation to behave or align with societal expectations, including gendered social roles or gender stereotypes.

The situated expectancy-value theory also describes how aspects of the social and cultural environment can influence the value children place on given tasks or domains. This description encompasses the influence of parents, teachers, peers, and other closely situated socialisers in developing motivations in each task or domain. For example, parents' enjoyment and perceived value of school subjects enable their children to engage with and persist in tasks and activities (Eccles & Wigfield, 2020; Wigfield et al., 2009). How socialisers interact with and influence children will change with age and will be influenced by sociocultural expectations of the appropriateness of their goals and behaviours, for example, gendered social roles and stereotypes (Eccles & Wigfield, 2020; Simpkins et al., 2012). Perceptions of socialisers' beliefs and behaviours are proposed to influence the subjective task value in each domain. As adolescents' social orientation is focused towards peers, peers are likely increasingly important socialisers during this period (Ahmed et al., 2022). Through increased alignment with peers and peer groups, peers may influence achievement goals and task values as adolescents increasingly align themselves with social groups that may be more or less closely aligned with their gender identity (Santos et al., 2013). Domain-specific self-concepts are also influenced through peer comparison and evaluation processes (Marsh et al., 2005).

1.2.4 Educations Expectations

The situated expectancy-value model encompasses the influence of long and short-term goals on achievement-related choices, motivation, and educational achievement (Eccles & Wigfield, 2020; Wigfield & Eccles, 2000). Within this thesis, long-term goals are operationalised as educational expectations, which combine both aspiration and an evaluation of whether that aspiration is achievable given ability beliefs and potentially other costs, such as the economic and social costs of going to university. Both parents and their children will mostly align, but may have distinct educational expectations. While related, parental and child educational expectations each make a unique contribution to the child's academic achievement (Zhang et al., 2011). Parental educational expectations of their child influence their child's expectations, motivation and achievement. These parental expectations are influenced by their child's prior achievement, self-concepts, and expectations in a dynamic, reciprocal relationship throughout the child's education (Buchmann et al., 2022; Pinquart & Ebeling, 2020; Zhang et al., 2011). The literature is divided on whether parental expectations are more influential for boys than girls (Zhang et al., 2011) or the opposite (Flouri & Hawkes, 2008). Nevertheless, it is sometimes reported that girls have higher educational expectations than boys. However, this may be influenced by special educational needs status, which is more prevalent in boys and parental occupations, particularly the sons of fathers in trade occupations (Dockery et al., 2022; Koshy et al., 2019; Rampino & Taylor, 2013).

1.2.5 Academic Self-Concept and Interest

Academic self-concept is an individual's perceived competence in a particular domain, whereas interest results from positive and rewarding engagement with that domain. The two concepts are related as students are more interested in the domains in which they feel competent (Jacobs et al., 2002). Interest is a weak predictor of self-concept, and self-concept has a stronger association with achievement (Denissen et al., 2007; Marsh et al., 2005). Both self-concept and interest are domain-specific. Girls report higher language/verbal self-concept and interest, and boys report higher mathematics and science self-concept and interest (Dai, 2001; Denissen et al., 2007; Goetz et al., 2013; Pennington et al., 2018). Girls report lower mathematics emotions than boys, which is linked to their lower self-concept but high valuing of achievement in mathematics (Frenzel et al., 2007). Higher self-concept predicts higher achievement, and associations are stronger when measuring matching, domain-specific self-concept, and achievement (Huang, 2011).

The association between self-concept and achievement is gender invariant, despite the mean differences between girls and boys (Kriegbaum et al., 2018; Marsh et al., 2005). Self-concept and achievement develop in a reciprocal relationship over time, where self-concept predicts subsequent achievement and prior achievement predicts subsequent self-concept (Marsh et al., 2005, 2018). Self-concept and interest reduce with age during adolescence, especially in mathematics, although gender differences remain stable (De Fraine et al., 2007; Frenzel et al., 2010; Jacobs et al., 2002; Watson et al., 2019; Wilkins, 2004). The association between academic self-concept and achievement also changes with age, becoming stable between early and mid-adolescence (Davis-Kean et al., 2008; Susperreguy et al., 2018). Much of the decline in interest is explained by the decrease in self-concept, with curvilinear trends suggesting that some recovery of interest and self-concept may occur during mid- to late-adolescence (Frenzel et al., 2010; Jacobs et al., 2002). The curvilinear trend is robust and is found in boys and girls at all achievement levels and in different cultures (Davis-Kean et al., 2008; Frenzel et al., 2010; Jacobs et al., 2002). Therefore, the decline in self-concepts and interest may be a normal consequence of adolescent development, as individuals develop increasingly realistic self-concepts in the context of their comparative progress in other domains and the performance of their peers while facing increasingly complex academic content (Marsh & Hau, 2003; Marsh et al., 2018; Möller & Marsh, 2013; Möller et al., 2009). Academic self-concept has been demonstrated as an important antecedent of school engagement, the observable manifestation of students' academic emotions and motivation (Bakadorova et al., 2020; Green et al., 2012; Pekrun & Linnenbrink-Garcia, 2012). Students with a higher academic self-concept are more likely to exhibit moderate to high school engagement, and together these constructs are associated with better academic achievement (Schnitzler et al., 2021; Veiga et al., 2015). In line with the declines in self-concept, declines in the emotional and behavioural dimensions of school engagement have also been reported during early to middle adolescence (Bakadorova et al., 2020).

1.2.6 School Engagement

School engagement is influenced by the degree to which the school environment matches the developmental needs of the individual, which is of particular relevance during adolescence (Eccles et al., 1993; Fredricks et al., 2004). School engagement is a multidimensional concept that encompasses externally observable presentations of an individual's internal academic emotions and motivation (Pekrun & Linnenbrink-Garcia, 2012; Reeve, 2012). School engagement comprises a student's behavioural, cognitive, emotional, and agentic engagement with school and learning (Fredricks et al., 2004; Reeve, 2012). The behavioural component measures effort, attention, and participation in learning. Emotional engagement measures a student's feelings towards the social and learning environment that includes their school, peers, and teachers (Fredricks et al., 2004). Cognitive engagement measures a student's cognitive investment and willingness to invest time and effort in their learning (Fredricks et al., 2004). Together, these three components of school engagement are thought to be particularly important during early adolescence, corresponding to their developmental needs during this period (Fredricks et al., 2004). Since the original conceptualisation of school engagement, two additional components have been encompassed into the concept. Agentic engagement, which refers to an individual's constructive and proactive contribution to learning (Reeve, 2012), and more recently, a social component referring to an individual's interaction with peers both within and outside the classroom (Wang et al., 2017). In general, higher levels of school engagement are related to higher academic achievement and educational expectations in a reciprocal relationship, although individual trajectories can differ (Chase et al., 2014; Li & Lerner, 2011). High levels of school engagement, achievement, and expectations can be a risk for emotional distress and exhaustion for some students, for example (Widlund et al., 2021). School engagement tends to decline during the adolescent years, with particular changes observed around school transitions between educational stages (Wang et al., 2015). When a school provides an optimal environment for individual students - one in which they feel supported in learning, think that what they are learning is relevant, and have some autonomy in decision-making - then students are more interested, engaged with, and value their learning activities (Wang & Eccles, 2013).

1.2.7 Stage Environment Fit Theory and the Self-Systems Model of Motivational Development

A mismatch between the opportunities and experiences offered by the school environment and the developmental needs of individual students contributes to motivational declines during adolescence (stage environment fit theory: Eccles et al., 1993; Symonds & Galton, 2014). During adolescence, three important developmental tasks take place - the formation of self-identity, a sense of self-agency, and improved self-regulation and decision-making (Hansen & Jessop, 2017). The three basic psychological needs of competence, autonomy, and relatedness are seen as fundamental to meeting these developmental milestones (Self-Determination Theory: Hansen & Jessop, 2017; Ryan & Deci, 2000a). Stage environment fit theory proposes that the gap between students' developmental needs and opportunities to meet those needs widens after the transition to middle school, which corresponds to the transition to UK secondary school (Eccles et al., 1993; Symonds & Galton, 2014). Stage environment fit theory is linked to both self-determination theory and the self-systems model of motivational development (Skinner et al., 2008, 2009). Students need to feel related and connected to others and the environment, and perceive themselves as competent, capable, and effective. Students also need to feel autonomous and in control (self-agency) to be able to maintain motivation and motivated behaviour. School transitions can, therefore, influence an individual's perceptions of how well their school meets their psychological needs, as transitions are characterised by the disruption of existing friendship networks and reduced family involvement (Symonds & Galton, 2014). In particular, secondary education school cultures focus on academic performance, in which students are taught by more teachers who know them less well than in primary education, and students have reduced opportunities to self-direct and feel autonomous. It is only through experiencing a sense of control that self-agency can be developed (Hansen & Jessop, 2017). The self-systems model proposes dynamic relations between an individual's experience of the context (the school environment), their self-appraisals (academic motivation and self-concepts), and their resulting engagement/disaffection, which influences their outcomes. Due to feed-forward effects between model components, students who are more highly engaged are proposed to become more so, whereas students who are less engaged would become even less so, over time (Skinner et al., 2008, 2009).

1.3 Chapter Summary

Understanding sex differences in educational achievement requires a multifaceted approach that considers cognitive abilities, societal influences, developmental stages, and direct environmental influences. While certain robust patterns emerge, such as larger female advantages in language subjects and other non-STEM subjects, sex differences in STEM subjects remain unclear, especially during the adolescent period. A substantial body of literature reports on sex differences in mathematics, language, and reading. Fewer studies have examined differences in science achievement, and even fewer have examined sex differences in science domains such as the physical and natural sciences (O'Dea et al., 2018; Voyer & Voyer, 2014). Study 1 examines sex differences in science achievement, including sex differences in chemistry, biology, and physics, while answering this thesis's first research question (RQ1), which asks how sex differences in achievement change with age, by subject. Voyer and Voyer's (2014) meta-analysis of school grades reports how sex differences change between primary, secondary, and tertiary education levels, as reported from predominantly North American studies and cross-sectional data. There are smaller sex differences in mathematics in non-North American samples, and the results differ from the conclusions drawn from analyses of large-scale standardised tests. Sex differences in STEM subjects are therefore unclear. Study 1 examines how sex differences change longitudinally in a single sample of British students at ages 7, 11, and 16 years old in mathematics, language, and science. Study 1, therefore, seeks to confirm the increasing female advantage in early to middle adolescence (RQ2) and clarify sex differences in STEM subjects (RQ3), in a single, longitudinal sample. As the data used here are sourced from a longitudinal panel study, the Millennium Cohort Study, in Chapter 2 - Methodology, the data gathering process is outlined, providing further details of the Millennium Cohort sample and the National Pupil Database from which the educational data were extracted. As the educational data include both teacher-assessed and test results for national, standardised curriculum tests, this thesis also examines whether conclusions drawn from teacher-assessment differ from test-based assessments (RQ4). Being a large sample of more than 5000 boys and girls, Study 1 also examines sex differences and the upper and lower levels of the subject achievement distributions at each age, to examine whether GMV is supported in school grades, and how this may change with age (RQ5). Finally, Study 1 compares these findings to results from analyses of international standardised tests to ask whether the conclusions drawn from these two types of tests align (RQ6). Having confirmed that the female advantage in education increases between early and middle adolescence, Study 2 examines one mechanism that, if linked to puberty, could contribute to explanations for this phenomenon. Currently, the reasons why puberty, a physical maturation process, and achievement are linked remain unclear.

Several studies have linked puberty to achievement in education, school dropout, the decision not to continue into non-compulsory education, and delinquent behaviour. Puberty is a sexual maturation process, but it is also one of the broader set of developmental processes associated with adolescence. The literature explaining how and why adolescent development, which also encompasses physical, cognitive, and socio-emotional change, is linked to educational achievement is currently limited, particularly for boys. Given that boys and girls develop at different rates and boys' pubertal maturation lags behind that of girls, any link between puberty and achievement may contribute to widening achievement gaps during adolescence. At present, the evidence is mixed, although this may partly result from the use of different measures to represent relative maturity in this age group. Study 2 builds on prior evidence of links between puberty, achievement motivation, school engagement and disengagement, and achievement through a longitudinal examination of how changes in puberty influence changes in behavioural and emotional engagement and disengagement between ages 11 and 14 years old and how these influences differ between boys and girls. Study 2, therefore, asks whether changes in perceived puberty status predict changes in school engagement during early to mid-adolescence (RQ7) and whether these changes contribute to the widening female advantage during secondary education (RQ8). Finally, in Chapter 5, the findings from Chapters 3 and 4 are discussed. The relevance and importance of the findings are evaluated in the wider context of existing adolescent development and educational theories are highlighted.

2 Methodology

This thesis is a longitudinal secondary data analysis of the educational outcomes and school engagement of Millennium Cohort Survey (MCS) participants. The Millennium Cohort is a large, longitudinal panel survey with, currently, seven waves of survey data collected between 9 months and 17 years old that can be accessed through the UK Data Archive. Additional datasets extracted from linked administrative databases, such as health and educational outcomes, are also available. This thesis analyses data collected in the MCS4 / Age 7, MCS5 / Age 11 and MCS6 / Age 14 survey waves in combination with linked educational data previously extracted from the National Pupil Database (NPD). This methodology chapter provides a more detailed description of the MCS and the NPD, describing the data-gathering procedure that took place prior to the data analyses reported in Chapters 3 and 4. This chapter also outlines the general methodological approach, including the decision not to use the sample weights in this thesis's analyses. Finally, I provide an overview of the measures used within this thesis, ethics approvals and consents obtained for the MCS data collection, and ethics approval for this thesis.

2.1 General Methodological Approach

The decision to use secondary data analysis, and particularly the Millennium Cohort Study, to answer the research questions addressed in this thesis was driven by both the limitations and criticisms of the current body of research and the advantages offered by analysing data collected as part of large-scale surveys. Much of the current educational literature, and most of the literature encompassed within school grades meta-analyses, is sourced from cross-sectional studies (O'Dea et al., 2018; Voyer & Voyer, 2014). While Voyer and Voyer (2014) included results from longitudinal studies, only the initial reported sex differences in school grades were included in the analyses when the same group of individuals was measured more than once. Therefore, the sex differences reported for each age group are repeated cross-sectional measures of sex differences that identify population trends. We, therefore, cannot be certain that the observed changes in sex differences reflect changing sex differences with age, are an artefact of the differing samples or reflect changes in how achievement is measured in each age-group (Gorard, 2003). Where longitudinal studies of sex differences report how sex differences change with age, often the same students' achievement is tracked over shorter periods (e.g., two to three years) than is achievable using large-scale survey data. To report on a longer range of age-groups, multiple cohorts are often examined over two to three consecutive years, rather than a single cohort examined over six successive years (e.g., Ding et al., 2006). In short, secondary data analysis of large-scale panel data allows researchers to answer longitudinal research questions using a single cohort over longer periods than is practical in most contexts, but particularly within the context relevant to this thesis - a doctoral research programme of three years. Secondary data analysis also allows researchers to examine change at an individual level, which is not possible in cross-sectional or repeated cross-sectional research designs (Gorard, 2003).

Beyond this, the Millennium Cohort Study is well-suited to supporting the aims of this thesis. The MCS is a contemporary cohort who completed their secondary education prior to the COVID-19 pandemic lockdowns, which further influenced sex and socioeconomic differences in educational outcomes (e.g., Oostdam et al., 2024). Additionally, the academic outcomes data available alongside the MCS datasets are national results from standardised curriculum assessments/tests, allowing this thesis to examine whether the sex differences observed are like those reported when school grades are teacher-assessed. As teacher-assessed grades are reported to be biased in favour of girls, or along subject-stereotypical lines, this bias may contribute to the differing conclusions drawn when examining sex differences in school grades and sex differences in international standardised assessments in mathematics and reading (e.g., Baye & Monseur, 2016; Ding et al., 2006; Lavy & Megalokonomou, 2019; Lavy & Sand, 2018). Therefore, one of the aims is to confirm the female advantage in educational achievement, especially in secondary education, in school grades that were anonymously and externally marked, therefore protected against possible teacher marking biases. Also, this thesis examines the earlier maturation of girls during adolescence and whether this is associated with the female advantage, while also addressing some of the limitations of conclusions drawn from cross-sectional studies and analyses.

An alternative methodological approach to secondary data analysis (data that is already available) is primary data collection and analysis, which provides researchers with more control and flexibility (Wyse et al., 2017). However, primary data collection would reduce the period over which an individual's school performance can be tracked. Typically, this would involve recruiting and engaging with local primary and secondary schools to collect the data of interest. Primary data collections allow researchers close control over what is measured and how, but in the context of doctoral research, would limit the longitudinal period to 2-3 years. In contrast, when using secondary data, the available data/measures constrain researchers, but allow change to be examined over longer periods, for example, using the MCS, birth to 17 years old at present (Gorard, 2003; UCL Centre for Longitudinal Studies, 2020, 2021c). In most research carried out using secondary data, the data is likely to have been collected without the researchers' input. How variables of interest are measured in large-panel surveys may, therefore, introduce additional limitations to analyses and conclusions that can be drawn from them, as these studies are often balancing covering a breadth of measures while managing the length of interviews to a pre-determined time limit (Wyse et al., 2017). Primary data collection, while more flexible, can be time-consuming, challenging - often it is difficult to recruit schools to take part due to schools being busy and under-resourced - and costly

(Gorard, 2003). The resulting samples from primary data collection are likely to be smaller, leading to lower statistical power, and it can be difficult to recruit a broad range of students to ensure that the sample is representative (Wyse et al., 2017). Large-panel survey data, however, address these latter limitations by recruiting substantially larger samples than is practical within most research projects, whether grant-funded or otherwise. In sum, analysing secondary data has limitations, including a lack of flexibility; the available measures may not match the researchers' preferred approach, for example. However, using a secondary data source for this enables a longer period of change to be examined than is practical when using primary data collection. It provides a rich sample that is broadly representative, which can be time-consuming and costly to replicate in research using primary data only.

2.2 The Millennium Cohort Study

The Millennium Cohort Study recruited children who were born across the UK (England, Scotland, Wales, and Northern Ireland) between 1 September 2000 and 11 January 2002 (Plewis, 2007). The MCS recruited nine-month-old UK resident children, surveying their parents and the children themselves in later surveys since then. The MCS used a stratified random sampling approach based on the characteristics of electoral wards. Electoral wards are subdivisions of the 650 constituencies that are each represented by an elected Member of Parliament in the UK Parliament. The UK Census, which is undertaken every ten years, collates a range of demographic data about each resident in the UK. This information can be collated by several types of geographic areas, which include electoral wards. The population of electoral wards can vary substantially. In 2018, for example, there were 7,065 residents, on average, in each ward. Some wards had more than 40,000 residents, whereas others had fewer than 200 (Office for National Statistics, 2018). For sampling, the electoral wards were divided into three strata - wards with a 30% or more ethnic minority population, wards in the lowest 25% by income that had not previously been categorised as having a large ethnic minority population, and the remaining wards, categorised as advantaged. Children were randomly sampled from these strata, intentionally oversampling from ethnic minority and disadvantaged (low-income) wards (Plewis, 2007).

To undertake this thesis's secondary data analyses, this survey collated the survey response data from the school-age surveys. As this thesis examines how the cohort members (the sample children) progressed during their education, the data-gathering process focused on the surveys, in which there was a self-completed cohort member survey. The first of these was encompassed within the MCS4 / Age 7 survey (Plewis, 2007; UCL Centre for Longitudinal Studies, 2021d). Three further survey waves took place within the age range of compulsory education in the UK (5 - 18 years old) - the MCS5 / Age 11, the MCS6 / Age 14, and the MCS7 / Age 17 survey waves (UCL Centre for Longitudinal Studies, 2021a, 2021b, 2021c). As the available educational data from national assessments recorded each cohort member's school achievement at ages 6-7, 10-11, and 15-16 years old, the MCS7 wave was not used as these data were collected after the last available set of educational achievement data. The UK Data Archive manages access to the MCS data.

2.3 National Pupil Database

The National Pupil Database, maintained by the UK Department for Education, collates the educational achievement data of all pupils except for those who are educated in Scotland. Scotland has a separate education system, which is distinct from the rest of the UK. During the Age 7 wave, the parents of cohort members were asked for their consent to link to their child's educational records. Of the cohort members in the Age 11 wave, permission was given for n = 7,508 cohort members had valid consent, and the NPD team successfully linked data for n = 7,252 of these. Matching was semi-automated using full name, sex, date of birth, and postcode initially, and reverting to manual matching for close matches of name or date of birth (Institute of Education, 2019). Consent to access the post-16 educational data was requested during the Age 17 wave, but these data have not yet been made available in the UK Data Archive. The linked NPD dataset contains the results of Key Stage 1, Key Stage 2, and Key Stage 4 assessments, which take place in the final term of school years 2, 6, and 11 when the cohort members were aged 6-7, 10-11, and 15-16 years old. Throughout this thesis, we refer to these educational data as the Age 7, 11, and 16 data, respectively.

The Age 7 educational data are the results of the national Key Stage 1 assessments that take place in late Spring. These assessments are teacher-assessed using a detailed national framework. Details of the Key Stage 1 assessments are made available online. They are moderated internally within the school, and schools are encouraged to moderate them externally. Local education authorities, who are responsible for education within their jurisdictions, moderate 25% of schools. A sample of outcomes in English reading, writing, and mathematics are moderated to ensure consistency with national standards (e.g., Testing and Standards Agency, 2018). Both teacher-assessed grades and grades achieved in national tests comprise the Age 11 educational data, and national tests only at Age 16. The Age 11 and 16 tests are anonymously and externally marked. The Age 11 assessments are the same across the country. The Age 16 tests are qualifications awarded by multiple examination boards, with processes in place to ensure the awarding of grades is standardised across providers. The extract used in this thesis contains the educational achievement data of those cohort members who were educated in England (Department for Education, 2021). The UK Data Archive also manages access to the linked educational data. Due to the sensitive nature of the data, permission to use these datasets is subject to secure access protocols.

2.4 Data Gathering

The data gathering process first drew together education-relevant data from the parent and child MCS questionnaires, creating a single row for each cohort member for each wave - Ages 7, 11, and 14. Following this, the NPD extracts were reorganised to create a single row per assessment subject for each assessment age for each cohort member - Age 7, 11, and 16. The third stage combined the survey data and the educational data into a combined dataset, merging Age 7 survey data with Age 7 educational data, Age 11 survey and educational data, and Age 14 survey with Age 16 educational data.

2.4.1 Millennium Cohort Study Data

The datasets for each survey wave comprised the responses from several questionnaires, each of which was stored in a separate data file. There are also additional data files that contain derived data, such as calculated results of cognitive skills tests, alternative ethnicity categorisations, and family income. These data files were combined into a single row per cohort member for each survey wave. The data files required for the analyses in Chapters 3 and 4 and the data they provide are summarised in Table 2.1. For each survey wave, the required data were merged from the six data files, creating a single row for each cohort member at each wave. A range of non-response codes (don't know, refusal, and no response) were removed during this process.

Three types of questionnaires were used to collect socio-economic and attitudinal data about the main caregivers of each cohort member, the cohort member, and the household more generally. One caregiver was designated the main respondent, who would complete the most detailed questionnaire (the CAPI Questionnaire), and their partner would complete the partner questionnaire. If the partner of the main respondent was unavailable to complete the partner questionnaire, then a separate partner-proxy questionnaire was completed by the main respondent. Sometimes, there was an inconsistency between the waves in terms of who completed the main and partner questionnaires. Therefore, the parent-derived data file provided additional indicators to clarify who completed each set of questions. The main caregiver also completed a questionnaire about the cohort member.

The CAPI questionnaire asks the main respondent to answer a broad set of questions covering household demographics and general family context, including the age and ethnic

Table 2.1

Data Sourced from the Millennium Cohort Study Data Files for each Wave

Data File	Description	Used for	Rows
cm_interview	CM interview responses	School engagement, domain-specific interest (Age 7, 11) and self-concept (Age 11, 14), PDS (Age 14), and educational expectations (Age 14)	1 per CM
cm_derived	CM derived data	Ethnic group	1 per CM
hhgrid	Derived household structure	CM age, age at interview, sex	1 per HH member
parent_derived	Parent derived data	Parent qualifications, relationship to CM	2 per CM
parent_cm_intervew	Parent interview responses about CM	School type, special educational needs, free school meals, school year group, PDS (Age 11), and educational expectations (Age 11)	1 per CM
family_derived	Family derived data	Disposable family income	1 per HH

Note. CM - Cohort Member, HH - household. PDS - Puberty Development Scale (Petersen et al., 1988).

group of the main respondent and information on their personal circumstances. The CAPI questionnaire goes on to capture information about the cohort member's early education, schooling, and childcare, family activities and child behaviour, their health and their child's health, their employment, income, and their education, and information about their neighbourhood and housing. The main respondent reported basic details about each child, including their sex, date of birth, and ethnicity. The main respondent also reported whether the cohort members had a special educational needs (SEN) diagnosis in each wave and whether their child received free school meals. The main respondent reported their educational expectations for their child and their child's puberty status in the Age 11 wave.

Cohort members completed questionnaires from the Age 7 wave onwards. Measures from these questionnaires were used to source the cohort members' school emotions and engagement in all waves and self-concepts in the Age 11 and 14 waves. In the Age 14 wave, the cohort members were asked to report their educational expectations and their puberty status. The Puberty Development Scale was used to report puberty status by parents and cohort members (Petersen et al., 1988).

2.4.2 National Pupil Database Extract

Table 2.2

Data Sourced from National Pupil Database Data Files at each Wave

Data File	Description	Used for	Rows
qualcodes_restricted	Qualifications reference data	Qualification Codes and Types, School Subjects, Subject Groupings	1 per assessment type, subject
ks4_exams_2017	CM test grades	Subject, qualification type, grade, exam season and year, subject grouping	1 per CM, subject, exam period
ks2_exams_2012	CM test and teacher-assessed outcomes	Subject, grade, fine test grade	1 per CM, subject
ks1_2008	CM teacher-assessed grades	Subject grades	1 per CM

Note. CM - Cohort Member, HH - household. PDS - Puberty Development Scale (Petersen et al., 1988).

The linked data sets extracted from the NPD contain details of the Age 7, 11, and 16 educational assessments and tests, which had been matched and extracted by the NPD team. Four data files were used to source the data required for the analyses described in Chapters 3 and 4 (see Table 2.2 for details). The Age 7 data was converted to a long format (1 row per cohort member/subject), and the qualifications reference data was used to identify all academic General Certificate of Secondary Education (GCSE) qualifications at age 16 (GCSE and International GCSE (IGCSE)).

2.4.3 Data Merging

A combined cohort across the three survey waves was created by selecting the data to be held constant across the survey waves (ethnic group, special educational needs status, free school meals status) from each wave-specific dataset and joining these together where the cohort member data was present in each wave-specific dataset. This joining process resulted in a combined set of n = 10,230 cohort members present in all three waves. To account for late diagnosis, special educational needs status was checked across waves and set if reported in any of the waves. Similarly, free school meal status was set to indicate that the cohort member had been entitled to receive free school meals in at least one of the survey waves. Three ethnic group categorisations were provided in the MCS data files - 6-category, 9-category, and 11-category versions. To maintain as many ethnic groups as possible given the reduced MCS dataset, the 9-category version was chosen. The 11-category version was not used due to very low numbers in some groups within the sample.

The wave-specific time-varying data were combined with the three age-group educational data using the cohort member identifier - a combination of family identifier and the cohort member's number within that family (to account for twins and other multiple births). The data of any cohort members who did not attend mainstream education was removed at this stage (Age 7: $n = 49$; Age 11: $n = 165$; Age 7: $n = 163$). These wave + education data sets were then combined into a single dataset - 1 row per wave, subject - and then any cohort members without at least one row at each wave were removed. Consent for access to educational data had been sought earlier in the MCS data collection process (Age 7), and there had been sizeable attrition (n = 2268 cohort members) since the Age 7 wave. The NPD extracts and the MCS retained cohort members at Age 14 were, therefore, not aligned. Although the NPD extracts contained n = 8042 Age 7 results, n = 7974 Age 11 results, and n = 8004 results, only n = 7530 of these were cohort members who had a result in each of these age-group assessments, prior to matching back to the retained MCS sample. These were the exam results of all cohort members for whom a valid Age 7 consent was available. The resulting dataset used for this thesis was of cohort members who were present in each MCS survey wave and for whom at least one subject's educational data was available at each assessment age, n = 5,795. This sample was used for Chapter 3. The sample used in Chapter 4 focused on those cohort members who had sat their mathematics and language Age 16 examinations at the normal time for their age group (Summer 2017), resulting in a sample of n = 5,617 cohort members.

Finally, a number of duplicate educational rows were removed - some cohort members had the same educational outcomes recorded against multiple schools. Where cohort members had multiple Age 16 subject outcomes, these were retained if they took place in different exam seasons, which is likely due to resits taking place, but duplicates with the same grade in the same exam season and year due to school change were removed. Some cohort members had two grades recorded against the same subject in the same exam season and year. Where one grade was higher than the other, the higher grade was retained, as these may result from re-marking.

2.5 Sampling Weights

The MCS team - the UCL Centre for Longitudinal Studies - provides sample design weights (probability weights) to correct for the biased likelihood of being selected under the clustered and stratified sampling design (UCL Centre for Longitudinal Studies, 2020). The weights used differ depending on whether the analyses required to test the analysts' research questions are country-specific or relate to the whole of the UK. The weights also adjust for non-response / attrition. Weighted analyses are advised where researchers want to calculate population estimates (UCL Centre for Longitudinal Studies, 2020). In general, not using the provided sample weights, and therefore treating the sample as if it were randomly sampled and observations are independent, risks underestimating standard errors (Ketende & Jones, 2011). Further, the MCS documentation also advises against subgroup analysis as, again, standard errors are likely to be underestimated in these situations. However, this thesis, as a direct result of using the educational data, analyses the educational outcomes and predecessors of these outcomes of a subgroup of Age 7 cohort members for whom parental permission to access the educational data had been given and who remained present in the Age 14 survey.

The sampling weights are not used in the main analyses within this thesis for two reasons. Firstly, the MCS does not provide sampling weights for the educational data extracted from the NPD. Secondly, the Study 2 (Chapter 4) analyses are a Bayesian implementation of longitudinal growth models, which prevents the use of complex survey weights in Study 2. As this thesis analyses data from multiple waves of the MCS, the sampling weights from the most recent wave (Age 14/ MCS6) are the appropriate weights to use. This thesis analysed a sub-sample of the Age 14 sample for whom Age 7, Age 11, and Age 16 educational data were available. In general, deleting unused group data and performing sub-sample analyses that are not advised when working with complex survey analyses using survey weights, as this too can bias standard errors (Jones & Ketende, 2010; Ketende & Jones, 2011). Applying sample weighting to analyses comes at a cost - substantially lower statistical power and lower efficiency, and the differences between unweighted and weighted analyses are, in most cases, very small and comparative analyses find that sample average effects from unweighted analyses do not significantly differ from population average effects (Haddad et al., 2022; Miratrix et al., 2018) Other evidence suggests that weighting, in general, should be treated with caution as the effects of weighting can differ between datasets (Haddad et al., 2022). Miratrix et al. (2018) conclude that for high-quality, broadly representative samples, such as the Millennium Cohort Study, for most purposes, unweighted analyses are sufficient and avoid the cost of using weighted analyses for limited gain – unless precise population estimates are required. However, in addition to disagreement within the methodological literature on the benefits vs. the costs of weighted vs. unweighted analyses, at present, there is no agreed solution to how weighting should be implemented in Bayesian analyses due to increased complexity (Bollen et al., 2016; Gelman, 2007; Miratrix et al., 2018). While the Study 1 independent sample t-tests are not unduly influenced by sample size or efficiency concerns and could be analysed using survey weights, the Study 2 Bayesian Growth Models, using the brms package in R, do not support complex survey weights (Bürkner, 2017, 2018). Therefore, in the absence of a complex survey weight implementation in R-based Bayesian analyses, the main study chapters remain as unweighted analyses to maintain consistency across the two studies. However, Study 1 weighted analyses are reported in Appendix B and compared to the main study results. Given the research questions being examined in this thesis - confirming a change in sex differences with age and whether puberty influences school engagement, while coefficients may change marginally between weighted and unweighted samples, the choice to use or not to use the sampling weights was not expected to substantially influence the answers to the research questions being asked.

2.6 The Thesis Sample

While there are technical reasons not to use complex survey weights for the analyses undertaken and reported in this thesis, this section provides an overview of the thesis sample demographics. It also examines how the decision not to use the complex survey weights may have biased the results.

The MCS team provided multiple weights to account for attrition and sampling design at each wave of the study (Jones & Ketende, 2010; UCL Centre for Longitudinal Studies, 2020). Three options are available to researchers. Two weights (weight1 and weight2) are provided for researchers who want to account for the sampling design only in their analyses. Otherwise, longitudinal weights are provided for each MCS wave, one for country-specific and one for whole-of-UK analyses. As this study uses educational data from England only, the single-country weights are used - aovwt1 for MCS1 / Age 9 months, and eovwt1 for MCS6 / Age 14 years. Other variables needed for the complex survey design weighting are pptype2 and sptn00, which indicate the cluster and stratum that each cohort member was sampled from (Jones & Ketende, 2010). To understand the effect of using or not using the complex survey weights for the Study 1 & 2 analyses, the ethnic background and income distribution of the thesis sample was compared with its weighted version in Tables 2.3 - 2.6. For illustrative purposes, the weighted and unweighted MCS1 sample is also reported.

The MCS1 sample of cohort members living in England comprised *n* = 11,676 children (*n* = 5963 boys, *n* = 5713 girls). Applying weightings to the MCS1 sample resulted in a weighted n_w = 5874 children (n_w = 2991 boys, n_w = 2883 girls). The thesis sample comprised *n* = 5697 cohort members (*n* = 2878 boys, *n* = 2910 girls). Applying weights to this thesis sample resulted in a weighted sample representing n_w = 5229 MCS cohort members (n_w = 2594 boys, n_w = 2635 girls).

Applying the survey weights to the thesis sample increased the proportion of White cohort members and reduced some other groups accordingly. The achieved proportion of White cohort members in the weighted thesis sample remains substantially below the weighted proportion of White cohort members in the MCS1 sample, or indeed in the population around the time the cohort members were born (91.2%; ONS, 2022). There are 2% more White boys than White girls, and therefore slightly more girls from several of the ethnic groups in the sample. Overall, both the unweighted and weighted versions of the thesis sample underrepresent the proportion of White cohort members and overrepresent the proportion of Asian and Black cohort members, as well as those from multiple ethnic backgrounds in comparison to the population (see Chapter 3, Table 3.1). Applying the sampling weights to the MCS1 sample resulted in a sample that represents the White population but substantially underrepresents the Black and South Asian populations at that time (Black: 2.2%; Asian 4.8%), according to UK 2001 Census results (ONS, 2022).

In MCS4, 19-21% of the sample was observed in each of the five OECD equivalised

Table 2.3

Ethnic Background of Female Cohort Members in the Weighted MCS1 Sample from England, Compared with the Unweighted and Weighted Thesis Sample.

	MCS1 Sample (England)				Thesis Sample			
	Unweighted		Weighted		Unweighted		Weighted	
Ethnic background	*n*	%	n_w	%	*n*	%	n_w	%
White	4174	73.1%	2654	92.1%	2240	76.0%	2180	82.7%
Pakistani	435	7.6%	16	0.6%	210	7.1%	122	4.6%
Indian	219	3.8%	49	1.7%	109	3.7%	58	2.2%
Bangladeshi	178	3.1%	6	0.2%	95	3.2%	44	1.7%
Other Asian incl. Chinese	73	1.3%	17	0.6%	38	1.3%	26	0.7%
Black African	181	3.2%	23	0.8%	67	2.3%	53	2.0%
Black Caribbean & Other Black	131	2.3%	17	0.6%	51	1.7%	41	1.6%
Mixed or Multiple	255	4.5%	87	3.0%	110	3.7%	90	3.4%
Other ethnic group	48	0.8%	12	0.4%	28	0.9%	21	0.8%
Not specified	19	0.3%	–	–	–	–	–	–

Note. Observed *n*'s are suppressed where *n* < 10 in the unweighted samples, and merged into the Other ethnic group. – indicates observations excluded when using the sample weights, or otherwise merged.

family disposable income categories (see Study 2 Measures Overview section below for details of this measure). The income distribution had changed in the MCS6 survey. Tables 2.5 and 2.6 below report the family income distribution of boys and girls in the original MCS1 England sample and the sample used in both study chapters. For comparison, the sample's unweighted income distribution at MCS4 is also provided to illustrate that the income distribution of this single sample changes, and an increase in mean family income is observed between the two surveys (also, see Study 2).

Applying the complex survey weights added additional bias to the sample. For example,

Table 2.4

Ethnic Background of Male Cohort Members in the Weighted MCS1 Sample from England, Compared with the Unweighted and Weighted Thesis Sample.

	MCS1 Sample (England)				Thesis Sample			
	Unweighted		Weighted		Unweighted		Weighted	
Ethnic background	*n*	%	n_w	%	*n*	%	n_w	%
White	4401	73.8%	2751	92.0%	2222	78.1%	2199	84.7%
Pakistani	436	7.3%	20	0.7%	183	6.4%	99	3.8%
Indian	229	3.8%	64	2.1%	111	3.9%	60	2.3%
Bangladeshi	178	3.0%	4	0.1%	70	2.5%	35	1.4%
Other Asian incl. Chinese	70	1.2%	8	0.3%	35	1.3%	18	0.7%
Black African	195	3.3%	30	1.0%	57	2.0%	36	1.4%
Black Caribbean & Other Black	148	2.5%	20	0.7%	49	1.8%	39	1.5%
Mixed or Multiple	235	2.9%	82	2.8%	94	3.3%	88	3.4%
Other ethnic group	53	0.9%	13	0.4%	23	0.8%	20	0.8%
Not specified	18	0.3%	–	–	<10	<0.3%	–	–

Note. Observed *n*'s are suppressed where *n* < 10 in the unweighted samples, and merged into the Other ethnic group. – indicates observations excluded when using the sample weights, or otherwise merged.

applying weights to the MCS1 results in an income distribution where fewer than 2% of cohort members are from the lowest income quintile. This falls substantially short of the 20% of children known to live in low-income households in the UK and is therefore not population representative (Francis-Devine, 2024; ONS, 2022). The substantial down-weighting of the cohort members from Bangladeshi families likely contributes to this, as Bangladeshi families are overrepresented in the lowest income quintile (Ethnicity Facts and Figures Service, 2023b). While the thesis sample became increasingly affluent, on average, between MCS4 and MCS6, applying the weights to the thesis sample does not result in a similar effect as observed for the

MCS1 sample. In fact, for most quintiles, the distribution of the sample between weighted and unweighted versions of the sample is very similar. There is, however, a < 2% reduction in the representation of male and female cohort members in the lowest income quintile in the weighted thesis sample.

Table 2.5

Family Income Groups of Female Cohort Members in the Weighted MCS1 Sample from England, Compared with the Unweighted and Weighted Thesis Sample.

OECD Income Quintile	MCS1 Sample (England)				Thesis Sample					
					MCS4		MCS6			
	Unweighted		Weighted		Unweighted		Unweighted		Weighted	
	n	%	n_w	%	n	%	n	%	n_w	%
1 (Lowest)	1428	25.0%	56	1.9%	600	20.3%	484	16.4%	398	15.1%
2	1276	22.3%	278	9.7%	586	19.9%	504	17.1%	472	17.9%
3	1045	18.3%	685	23.8%	573	19.4%	573	19.5%	519	19.7%
4	982	17.2%	928	32.2%	630	21.4%	677	23.0%	593	22.5%
5 (Highest)	958	16.8%	935	32.4%	559	19.0%	710	24.1%	653	24.8%
No data	24	0.4%	–	–	–	–	–	–	–	–

Note. Observed n's are suppressed where n < 10 in the unweighted samples, and merged into the 3rd income quintile. – indicates observations excluded when using the sample weights, or otherwise merged.

Applying the complex survey weights to account for the sampling design and attrition between survey waves resulted in a lower weighted sample size for analysis, and the sample remained biased. The sample remained unrepresentative of the population's ethnic group distribution, and additional income-related and SEN status bias was introduced. For Study 1 (Chapter 3), changes to the income distribution and a reduced sample size would further reduce the number of optional subjects that can be analysed, particularly where choices are influenced by family disposable income. For example, lower-income students are less likely to go to university, and this choice/decision will influence subject choice at age 14 in the UK (e.g., Anders, 2012; Department for Education, 2019). Applying survey weights to the thesis sample also further increased the already high proportion of SEN students in the sample, more so for boys than for girls. Using the weighted sample for analysis would increase the proportion of male SEN cohort members represented in the thesis sample from 30.6% to 32.3%.

Table 2.6

Family Income Groups of Male Cohort Members in the Weighted MCS1 Sample from England, Compared with the Unweighted and Weighted Thesis Sample.

OECD Income Quintile	MCS1 Sample (England)				Thesis Sample MCS4		Thesis Sample MCS6			
	Unweighted		Weighted		Unweighted		Unweighted		Weighted	
	n	%	n_w	%	n	%	n	%	n_w	%
1 (Lowest)	1469	24.6%	42	1.4%	509	17.9%	440	15.5%	366	14.1%
2	1357	22.8%	291	9.8%	529	18.6%	426	15.0%	410	15.8%
3	1106	18.5%	741	24.8%	624	21.9%	552	19.4%	521	20.1%
4	1023	17.2%	936	31.3%	608	21.4%	713	25.1%	645	24.9%
5 (Highest)	983	16.5%	981	32.8%	574	20.2%	713	25.1%	652	25.1%
No data	25	0.4%	–	–	–	–	–	–	–	–

Note. Observed n's are suppressed where $n < 10$ in the unweighted samples, and merged into the 3rd income quintile. – indicates observations excluded when using the sample weights, or otherwise merged.

In summary, applying the MCS6 sample weights to the thesis samples does not meet the aim of achieving a population-representative sample from an ethnic background perspective, and the sample is, through observed change, tending to become more affluent over time. Applying the sample weights to the MCS6 sample restricted by the available educational data further reduces the representation from the lowest income quintile and increases the proportion of SEN cohort members in the sample, introducing further bias to educational outcomes. Given that this thesis examines sex differences between boys and girls, which persist across ethnic groups (e.g., Strand, 2014), there appears to be little benefit to analysing these data using the MCS6 sampling weights - even if this were practically achievable for Study 2. Knowing that the use of sampling weights will reduce efficiency and statistical power (Miratrix et al., 2018), their use in educational-related analyses should be driven by whether the sampling weights achieve a sample that fully reflects the population's demographic, income, and SEN profile -

important influences on educational outcomes (Strand, 2011, 2014). This is not achieved for this sample. Therefore, the use of weights for these analyses is not justified for this thesis.

2.7 Measures

The first study chapter of this thesis tested how sex differences in school achievement between girls and boys, by subject and age group, changed during compulsory education within England's education system. The second study chapter examines associations of puberty, income, special educational needs status, parental education, and English and mathematics self-concepts on school engagement. The following section provides an overview of the measures used in each study chapter.

2.7.1 Study 1 Measures Overview

Chapter 3 quantifies how sex differences in English, mathematics, and science change during compulsory education. Each measure of school achievement is the result of an age-specific, curriculum-based national assessment. Teachers assess their students' achievement when students are aged 6-7 years old using detailed national assessment criteria (Testing and Standards Agency, 2018). These teacher assessments are then internally and externally moderated to ensure consistency between schools. At age 10-11 years old, towards the end of primary education, students are assessed using national, standardised curriculum tests in English and mathematics and by their teachers in science, again using national assessment criteria (GOV.UK, n.d.; Testing and Standards Agency, 2025). At the end of compulsory school-based education (15-16 years old), students sit national, standardised curriculum examinations (GOV.UK, n.d.). Chapter 3, therefore, tests sex differences in achievement in each of these national assessments. In addition to sex differences in mathematics, English, and science, we also examined sex differences in the optional subjects that students study from the age of 14-15 years old, and they are then tested at the age of 15-16 years old.

2.7.2 Study 2 Measures Overview

Chapter 4 examined how school engagement changes between the end of primary education, 10-11 years old, and mid-adolescence, 14-15 years old, and whether there was an association between puberty (Puberty Development Scale (PDS): Petersen et al., 1988) and school engagement. An association between puberty hormone levels and school engagement has been reported previously (Martin et al., 2022). The early to middle adolescence period corresponds with MCS waves 5 and 6. This chapter, therefore, tested whether changes in school engagement were associated with changes in pubertal development between the Age 11 and Age 14 MCS survey waves. Six questions measuring school engagement were available in MCS5 and MCS6. As this set of questions did not match a full published scale, the literature was searched to identify which dimensions of school engagement the items aligned with. Using Wang et al. (2019), three of the items were identified as measuring emotional and behavioural engagement and behavioural disengagement. The remaining three questions measured the emotional disengagement. These measures represent two of the three originally identified dimensions of school engagement (Fredricks et al., 2004). Confirmatory factor analysis (CFA) was used to clarify how best to structure school engagement for analysis (see Study 2 and Appendix C for details). There is some debate in the literature on whether school engagement should be treated as a single measure of school engagement, a dichotomised measure of school engagement and disengagement, or as two (or more) dimensions of behavioural and emotional (dis)engagement (Fredricks & McColskey, 2012; Skinner et al., 2008, 2009; Wang et al., 2017). This thesis followed Wang et al. (2017), dichotomising each dimension as emotional engagement and emotional disengagement, behavioural engagement and behavioural disengagement. The CFA results (see Appendix C) indicated that this was the most appropriate approach for each survey wave. As a result of the available items, emotional and behavioural engagement and behavioural disengagement were operationalised as single-item measures and emotional disengagement as a three-item measure. The internal consistency of emotional disengagement in MCS5 and MCS6 was tested by calculating Cronbach's α and McDonald's ω (Dunn et al., 2014; McDonald, 2013; Revelle, 2024). The internal consistency results were at the lower end of the acceptable range (MCS5: α = 0.62, ω = 0.63; MCS6: α = 0.62, ω = 0.61), with a lower bound of 0.6 - 0.7 suggested in the methodological literature (Hulin et al., 2001; Tavakol & Dennick, 2011). The alpha coefficient is influenced by the number of test items, item interrelatedness, and heterogeneity (Tavakol & Dennick, 2011). While the individual items are weakly correlated (MCS5: r = 0.31 - 0.43; MCS6: r = 0.31 - 0.38), correlations of each item with the total score are stronger (MCS5: r = 0.68 - 0.79; MCS6: r = 0.75 - 0.79), indicating that all items should be retained (Tavakol & Dennick, 2011).

Control Variables. In addition to the main variables being examined in this chapter - puberty development status and its association with school engagement, the following variables were included in the Bayesian Growth Models to understand whether puberty associations are still found when controlling for known associations with school engagement (Skinner & Raine, 2022, e.g.,).

1. **Age (time-varying).** The age of the cohort member on the day they were interviewed/surveyed in each survey wave.

2. **Special Educational Needs (SEN) (time-invariant).** Whether the cohort member's parent had reported a SEN diagnosis in one of the school-age survey waves (MCS4 through MCS6).

3. **Prior Achievement (time-invariant).** Mean achievement in language and mathematics at the end of primary education. This was calculated using the Age 11 educational data reported in Study 1.

4. **Parental Educational Qualifications (time-invariant).** The highest educational

qualifications level reported by either of the parents of the cohort member in MCS4, MCS5, or MCS6.

5. **Family Income (time-varying).** The OECD UK Income Quintile - an equivalised measure of family disposable income measured in each survey wave. This measure is derived from the economic data reported by the main and partner parents in the study. It is pre-calculated and forms part of the available MCS datasets.
6. **Mathematics Self-Concept (time-varying).** Cohort members' self-belief in their mathematics ability reported in MCS5 and MCS6.
7. **Language Self-Concept (time-varying).** Cohort members' self-belief in their language ability reported in MCS5 and MCS6.
8. **Educational Expectations (time-varying).** Parental educational expectations of their child, reported in MCS5, and the cohort members' educational expectations, reported in MCS6.

2.8 Ethical Considerations

The MCS surveys collect a range of personal data for each cohort member and their immediate family members. These data include detailed information on family income, cognitive assessments, and physical measurements of the cohort members, and a broad set of attitudinal and behavioural measures (UCL Centre for Longitudinal Studies, 2020). Additionally, other linked administrative data are available for research use that extend the available data further, for example, into the educational measures used in this thesis. This section outlines the ethical considerations and approvals obtained to enable these studies to proceed, outlining the approach taken by the MCS team and considerations specific to this thesis.

2.8.1 MCS Ethical Approvals

The MCS ethical approval process is consistent for each MCS survey wave. Approval for each wave is given by NHS multicentre Research Ethics Committees (MRECs), who are appointed by the Strategic Health Authorities, or their equivalents in Scotland, Northern Ireland, and Wales (Shepherd & Gilbert, 2019). For the data accessed in this thesis, the MREC ethical approvals given are outlined in Table 2.7.

Table 2.7

MCS Ethical Approvals Relevant to this Thesis

Survey Wave	Age	Survey Year	MREC	Approval
MCS1	9 months	2000/1	South West MREC	MREC/01/6/19
MCS4	7 years	2007/8	Yorkshire MREC	07/MRE03/32
MCS5	11 years	2011/12	Yorkshire and The Humber – Leeds East	11/YH/0203
MCS6	14 years	2015/16	London – Central MREC	13/LO/1786

Note. MREC - Multicentre Research Ethics Committee. Sourced from Shepherd and Gilbert (2019).

2.8.2 Consent

The MCS survey approach to getting consent is also consistent across survey waves. At each wave, explicit written consent is obtained for each survey administered, with separate permissions obtained for any requests for data linkages, for example, to the National Pupil Database for educational data (Shepherd & Gilbert, 2019). Letters and leaflets are sent in advance of the surveys to ensure informed parental consent. Written consent is obtained from parents for their participation and the participation of their child(ren). Siblings may also be surveyed in addition to the child subjects of the study. While parents may give consent for their children to participate, written consent was also obtained from the children when practical to do so, with interviewers asked to ensure child subjects were as fully informed as possible, with the support of child-focused leaflets. In line with other cohort studies, individuals can refuse to participate in individual elements of the survey, or withdraw from the study, at any time (Shepherd & Gilbert, 2019).

Table 2.8

Summary of Consents for MCS Survey Waves Applicable to Thesis.

Wave	Obtained From	Purpose
MCS1	Parent(s)	Administer survey (main & partner parents)
MCS4	Parent(s)	Administer survey (main & partner parents). Permission to link to the main parent's economic records and the child's educational records, and to contact the child's teacher. Administer child survey and cognitive assessments
MCS5	Parent(s)	Administer survey (main & partner parents). Permission to link to the main parent's economic records and contact the child's teacher. Obtain consent from the cohort member to administer the child survey and cognitive assessments.
MCS5	Cohort Member(s)	Administer complete self-completion questionnaire and cognitive assessments. Consent to be physically measured and to contact their class teacher.
MCS6	Parent(s)	Administer survey (main & partner parents), permission to link to main parent economic data, and obtain consent from the child subject(s) to administer child survey and cognitive assessments.
MCS6	Cohort Member(s)	Administer complete self-completion questionnaire and cognitive assessments.

Note. Sourced from Shepherd and Gilbert (2019), where further details of the consent process for each wave are also provided.

2.8.3 Data Access Protocols

Access to the MCS datasets is managed under two different access protocols. The main data collected in the MCS surveys is designated as "safeguarded" data. These data are accessed through a registration process, including a licensing agreement that defines the terms and conditions for using these data. More sensitive data is designated as "controlled" and managed under more stringent secure access protocols. The NPD data used in this thesis is designated as controlled data. These sensitive or confidential data are only accessible through a secure environment - the UK Data Service Secure Lab (see https://ukdataservice.uk/find-data/access-conditions, for further details). Data analyses must be published from the Secure Lab in the form of written reports and checked by two UKDS researchers prior to approval for release and onward publishing. This process is in place to ensure the confidentiality and anonymity of cohort members are maintained.

2.8.4 Thesis Ethical Considerations and Approvals

While all MCS data are stored in an anonymised form, with consistent personal identifiers for cohort members across survey waves, additional measures are put in place to ensure that individual cohort members cannot be identified. Where low numbers of cohort members in a subgroup analysis could lead to possible identification of an individual, data is suppressed in published outputs. Examples of this can be seen throughout Chapters 3 and 4, for example, in Table 3.2.

In addition to the ethical approvals sought as part of the data collection for the Millennium Cohort, the research encompassed within this thesis received ethics approval ETH2021-0417 from Ethics Sub Committee 3 of the University of Essex.

2.9 Chapter Summary

This thesis takes a secondary data analysis approach to answer two main research questions: How do sex differences in subject-specific school achievement change during compulsory education, and does puberty influence school engagement during the early adolescence period, contribute to the reported link between puberty and educational achievement and the widening female advantage in secondary education? Using large-scale panel data from the Millennium Cohort Study, this thesis gathered together multiple data files from three survey waves. Once this process was completed, these data were then combined with the available educational records previously extracted from the National Pupil Database. The educational data had been extracted based on consent given in the Age 7 wave, which was given for 60% of cohort members at that time. Our resulting sample represents 57% of the

Age 14 cohort, which aligns well with the Age 7 sample's proportion with consent. In the next chapter, Study 1, this thesis examined sex differences in achievement and how these changed with age for this cohort.

3 Sex Differences of School Grades in Childhood and Adolescence.

A version of this chapter is published as: Oakley, C. M., Pekrun, R., & Stoet, G. (2024). Sex Differences of School Grades in Childhood and Adolescence: A Longitudinal Analysis. *Intelligence*. https://doi.org/10.1016/j.intell.2024.101857.

3.1 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.

3.2 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, standardised, large-scale educational assessments (Baye & Monseur, 2016; Reilly et al., 2015, 2019). These large-scale assessments differ from school grades in that they measure point-in-time achievement in aptitude tests. In contrast, 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 channelled from general education to increasingly specialised 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, emphasise general education and a later choice between academic or vocational options (Burgess et al., 2022; Pekkarinen, 2008; van Elk et al., 2011). The UK's educational 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).

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. We sought to establish how subject-specific sex differences change with age (Research Question RQ1), and whether a widening female advantage is observed in secondary education (RQ2). We also aimed to clarify sex differences in STEM subjects, given that these remain unclear in the literature to date (RQ3), and how the conclusions drawn differ when comparing outcomes from school grades with teacher-assessed grades (RQ4), international standardised tests (RQ6). Finally, we examined sex differences at the upper and lower extremes of grade distributions to clarify how the variability of boys' and girls' school grades contributes to sex differences and how variability changes with age (RQ5).

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 emphasises 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, 2012; 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 overrepresentation of North American studies in the 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 criticised, 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 overrepresented 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 standardised mathematics aptitude tests (SAT Math), 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). However, it should be highlighted that the use of standardised aptitude tests, from which conclusions about the overrepresentation of men at upper achievement levels are drawn, are often criticised within the literature as being biased against some groups (Mattern & Patterson, 2014; Willingham & Cole, 2013). For example, while SAT scores are good predictors of eventual college grades, a sex difference is observed when examining how well SAT scores predict male and female students' outcomes (Mattern & Patterson, 2014). Men tend to achieve higher scores on the SAT test, but do not achieve higher grades. This sex difference is documented in several studies, sometimes referred to as the female underprediction effect, under which aptitude scores overpredict men's grades and underpredict women's grades (e.g., Duckworth & Seligman, 2006; Kling et al., 2013). In addition to these concerns about the conclusions drawn from standardised aptitude tests, the variability of international standardised assessments is not stable over time. 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.

3.2.1 Present Study

This study sought to better understand 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 standardised 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 hypothesised 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 standardised 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. Boys' school grades are more variable than girls' school grades at all ages.

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

H3b. Girls' STEM grades are more variable than their non-STEM grades. The variability of boys' grades is similar in STEM and non-STEM subjects.

3.3 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, 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, standardised test grades, and the remainder are standardised 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 (Department for Education, 2021), eleven years old (MCS5: UCL Centre for Longitudinal Studies, 2021a) and fourteen years old

(MCS6: UCL Centre for Longitudinal Studies, 2021b). 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 subdivisions. 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 categorised 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, neighbourhood, and disadvantage effects.

3.3.1 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 analysed 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).

3.3.2 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 equivalised income categories. Family income was adjusted to account for the number and age of dependents supported by that income, in accordance with OECD equivalence scales (UCL 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 (ONS, 2022).

Table 3.1

Ethnic Background of the Sample Compared to UK Census 2001 Results.

UK Census 2001 Ethnic Grouping	%	Sample Ethnic background	**n**	%
White	91.2%	White	4450	77.0%
Asian ethnic groups	4.8%	Pakistani	393	6.8%
		Indian	220	3.8%
		Bangladeshi	165	2.9%
		Other Asian incl. Chinese	73	1.3%
Black ethnic groups	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).

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 behavioural 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).

3.3.3 Measures

The school grades analysed 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 standardised curriculum test grades at age 11, and standardised curriculum test grades only at age 16.

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.

Achievement at Age 11. Age 11 educational assessments comprise both teacher-assessed grades and standardised curriculum 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 standardised 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 Appendix A, Tables A.9-A.10, pg. 150-151).

Achievement at Age 16. The achievement scores for the core subjects of mathematics and English language (reading, writing, and grammar), the results from standardised curriculum tests, 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*, also the results from standardised curriculum tests. 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 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.

3.3.4 Analyses

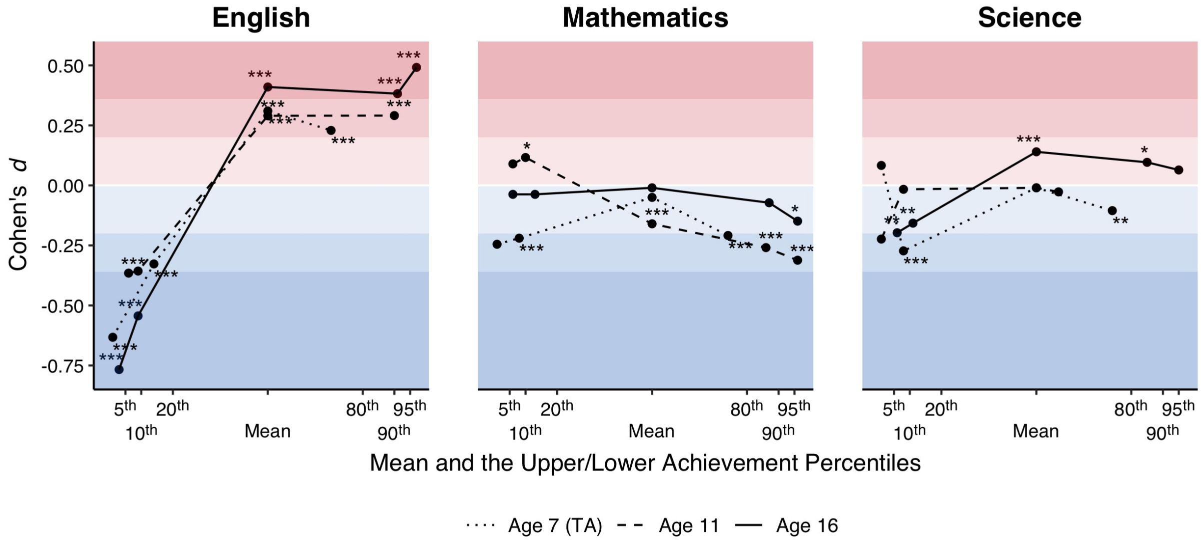
We analysed mean differences in school grades between boys and girls assessed in mathematics, science, and language subjects at ages 7, 11, and 16. We employed independent sample t-tests to test group differences and used the Welch modification to estimate degrees of freedom where variances are not similar. 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.

3.4 Results

Figure 3.1

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 A.2–3 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 standardised grades. *** p < .001, ** p < .01, * p < .05.

3.4.1 Longitudinal Changes in Mean School Grades

As expected (H1), the achievement gap between boys and girls changed with age and differed by subject (see Figure 3.1). Mean sex differences in English were stable during

Table 3.2

Standardised Means, Standard Deviations, and Percentages of Boys and Girls Observed at the Upper and Lower Tails of the Grade Distributions in English, Reading, and Writing.

Subject	Age	Sex	*N*	*M* (Cohen's *d* [CI])	*SD* (σ_k^2 *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]	<.001	<.001	<.001	<.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]	<.001	<.001	<.001	<.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]	.005	<.001	<.001	<.001	<.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]	<.001	<.001	<.001	<.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]	<.001	<.001	<.001	<.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]	<.001	<.001	<.001	<.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]	<.001	.005	<.001	<.001	<.009
		M	2846	-0.16	1.01	1.9%	16.6%	26.9%	0.9%

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_k^2) 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 Appendix A, Table A.2. TA indicates teacher-assessed grades.

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 (see Table 3.2). Sex differences in reading and writing were small and stable (reading, age 7: *d* = 0.26; age 11: *d* =

Table 3.3

Standardised Means, Standard Deviations, and Percentages of Boys and Girls Observed at the Upper and Lower Tails of the Grade Distributions in Mathematics.

Subject	Age	Sex	*N*	*M* (Cohen's *d* [CI])	*SD* (σ_k^2 *p*)	Percentage of Girls/Boys (χ^2 *p*)			
						Lowest Level	Lower Level	Upper Level	Highest Level
Mathematics	7	F	2949	0.02	0.93	0.7%	5.3%	22.4%	–
	TA			-0.05[-0.10, 0.01]	.<.001	.160	<.001	<.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]	.017	.161	.018	<.001	<.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]	.051	.593	.401	.109	.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_k^2) *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 Appendix A, Table A.1. TA indicates teacher-assessed grades.

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 in English and reading between ages 7 and 11, despite the stability of mean sex differences observed in these subjects during primary school. We also highlight that the age 11 assessments in reading are standardised 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 Appendix A, pg. 148–153, for an analysis of teacher-assessed vs. standardised 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 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 (see Table 3.3). Boys may have made more progress than girls overall, given the higher achievement of boys compared to girls in mathematics and the reduced overrepresentation of boys at the lower achievement levels. 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, which was as expected. 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 Tables 3.4 and 3.5).

Table 3.4

Standardised Means, Standard Deviations, and Percentages of Boys and Girls Observed at the Upper and Lower Tails of the Grade Distributions in Combined 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	F	2947	-0.00	0.95	0.4%	5.8%	24.5%	–
	TA			-0.01[-0.06, 0.04]	< .001	.896	<.001	.001	
		M	2845	0.00	1.05	0.4%	9.2%	28.2%	–
	11	F	2945	-0.01	0.99	0.6%	7.6%	42.8%	–
	TA			-0.01[-0.06, 0.04]	.152	.240	.817	.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]	.140	.032	< .001	.005	.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]	.663	.006	.002	.001	.098
		M	1815	-0.10	1.01	6.8%	17.7%	9.7%	1.6%
Average Science Grade	16	F	2850	-0.03	0.99	5.0%	9.6%	15.9%	5.4%
				0.14 [0.10, 0.17]	.096	.002	.001	.021	.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. Achieved power at age 16 – science: 0.99, biology: 0.91, chemistry: 0.77, physics: 0.20 (Champely, 2020). Data are excluded (–) where n < 10. Averaged science grade combines individual outcomes across science and additional science, or across biology, chemistry, and physics (see Table 3.5). Observed frequencies at the lowest, lower, upper, and highest achievement levels, and which grades each level encompasses, are detailed in Appendix A, Table A.3. TA indicates teacher-assessed grades.

We also analysed mean differences across a broad range of optional subjects studied in

Table 3.5

Standardised Means, Standard Deviations, and Percentages of Boys and Girls Observed at the Upper and Lower Tails of the Grade Distributions in Single Science Subjects.

Subject	Age	Sex	*n*	*M* (Cohen's *d* [CI])	*SD* (σ_k^2 *p*)	Percentage of Girls/Boys (χ^2 *p*)			
						Lowest Level	Lower Level	Upper Level	Highest Level
Biology	16	F	826	0.18	0.97	–	–	13.8%	4.8%
				0.15[0.05, 0.25]	.629			.006	.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]	.748		.	.021	.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]	.303		.	.271	.480
		M	774	0.06	1.02	–	–	12.7%	5.2%

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. Achieved power at age 16 – science: 0.99, biology: 0.91, chemistry: 0.77, physics: 0.20 (Champely, 2020). Data are excluded (–) where n < 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. Observed frequencies at the lowest, lower, upper, and highest achievement levels, and which grades each level encompasses, are detailed in Appendix A, Table A.3. TA indicates teacher-assessed grades.

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 Tables 3.6 and 3.7). 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 3.7 due to comparatively few upper grades achieved by the cohort members who studied these subjects (for details, see Appendix A, Table A.4-A.5). Together with the results in the core subjects of mathematics, English, and science, these findings indicate a general female advantage in secondary education across most subjects.

Table 3.6

Standardised 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: STEM and Modern Foreign Languages.

Subject	Sex	*n*	*M* (Cohen's *d* [CI])	*SD* (σ_k^2 *p*)	Percentage of Girls/Boys (χ^2 *p*)				
					Lowest Level	Lower Level	Upper Level	Highest Level	
Additional Science	F	1955	0.09		0.98	4.7%	14.0%	13.5%	2.6%
			0.19 [0.12, 0.25]	0.99	.663	.006	.002	.001	.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	.019	<.001	< .001	.001	.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	.805	.013	< .001	< .001	.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	.842	.284	.209	< .001	
	M	192	-0.21		1.01	8.8%	18.7%	14.5%	–
French	F	899	0.10		0.98	2.7%	8.0%	26.6%	11.4%
			0.25 [0.15, 0.35]	0.99	.260	.022	.002	< .001	.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	.414	.586	.011	.333	.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	.824	.167	.001	.002	.013
	M	370	-0.20		1.01	5.1%	14.6%	21.6%	8.4%

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. Observed frequencies at the lowest, lower, upper, highest achievement levels, and which grades each level encompasses, are detailed in A.4.

Table 3.7

Standardised 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: Humanities, Arts, Sport, and Applied Subjects.

Subject	Sex	n	M	SD		Percentage of Girls/Boys			
			(Cohen's d [CI])	(σ^2_k p)		(χ^2 p)			
					Lowest Level	Lower Level	Upper Level	Highest Level	
History	F	1458	0.11		0.97	8.1%	15.2%	31.9%	11.3%
			0.24 [0.16, 0.31]	1.00	.063	.005	< .001	< .001	< .001
	M	1266	-0.13		1.02	11.4%	20.5%	23.8%	6.9%
Religious	F	1780	0.19		0.92	5.6%	11.6%	37.0%	12.8%
Studies			0.41 [0.34, 0.48]	1.00	< .001	< .001	< .001	< .001	< .001
	M	1546	-0.21		1.04	12.8%	22.3%	22.2%	6.9%
Design &	F	728	0.35		0.91	3.7%	9.3%	28.3%	9.6%
Technology			0.61 [0.47, 0.75]	1.00	.140	< .001	< .001	< .001	< .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	.017	< .001	< .001	< .001	< .001
	M	672	-0.17		1.03	11.5%	20.4%	14.9%	2.8%
Physical	F	433	0.18		0.98	–	7.2%	28.6%	9.5%
Education /			0.29 [0.17, 0.41]	0.99	.105		.014	< .001	.085
Sports Studies	M	708	-0.11		1.01	–	11.9%	15.3%	6.5%
Art &	F	1108	0.16		0.95	2.1%	6.1%	28.2%	11.7%
Design			0.53 [0.42, 0.64]	1.00	.580	< .001	< .001	< .001	< .001
	M	488	-0.36		1.01	5.5%	14.8%	14.0%	4.7%

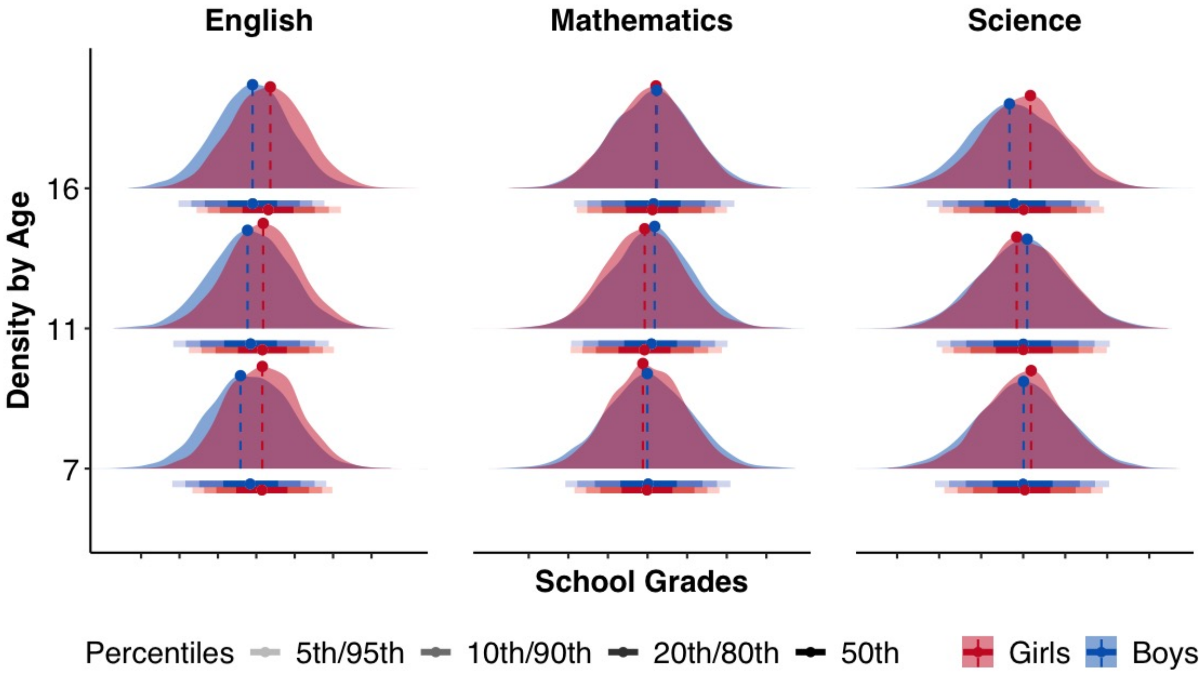
Note. Sex differences are nested between boys and girls means, standard deviations and percentage representation at each achievement level: Cohen's d [95% CI], Levene's (σ^2_k) test p-value, and Pearson chi-square (χ^2) test p-value for each achievement level. P indicates achieved power (Champely, 2020). Data are excluded (–) where the observed count of boys or girls <10. Subject groups: 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, and highest achievement levels, and which grades each level encompasses, are detailed in A.5.

3.4.2 Longitudinal Changes in Grade Variability

Boys' school grades were always significantly more variable than girls' in English, reading, and writing (see Figures 3.1, 3.2 and Tables 3.2-3.5). Science grades were

Figure 3.2

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.

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 Figures 3.1, 3.2 and Tables 3.2-3.5). Sex differences in grade variability were smaller in STEM 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 age 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 Tables 3.6, 3.7 and Appendix A, Tables A.4-A.5). In summary, we find that grade variability tended to change with age, but changes were sex and subject-specific. Girls' grades tended to become more variable with age, and boys' grades became less variable with age, so that sex differences in grade variability reduced, or were no longer significant. There were subject differences in how quickly grade variability converged.

The percentage of cohort members at the upper and lower tails differed between subjects at age 16 (see Appendix A, Tables A.2 – A.3). 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 grade distributions of the optional subjects often substantially differed, likely indicating that the profiles of students who chose each subject differed between subjects, and sometimes between girls and boys.

The percentage of students assessed at the lowest achievement levels in science at ages 7 (0.4%) and 11 (0.8%) was substantially lower than at age 16 (5.0%). The reliance on teacher-assessed grades for science assessments in primary school contributes to this difference. Teachers' 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 (<10). 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 (age 7: d = -0.63, age

11: *d* = -0.37, age 16: *d* = -0.77) and girls in the upper tail at each age (age 7: *d* = 0.23, age 11: *d* = 0.29, age 16: *d* = 0.38), indicating stronger female performance in this subject throughout (see Figure 3.2). 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 in 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. This pattern may indicate some improvement for boys at the lower achievement levels in English between ages 7 and 11, which is not subsequently maintained between ages 11 and 16.

In mathematics, boys were significantly overrepresented in the upper (*d* = -0.22) and lower tails (*d* = -0.22) at age 7 only. At age 11, boys were significantly overrepresented among the upper 13% (*d* = -0.26) and 5% (*d* = -0.31), girls in the lower 13% (*d* = 0.12), 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 (*d* = -0.15), 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%. These changes suggest that more girls achieved lower achievement levels in the age 11 mathematics tests than had been assessed at that level at age 7, resulting in a substantial increase in the proportion of girls in the lower 13% of the mathematics grade distribution. At age 16, though, more boys were observed in the lower achievement levels, so that girls were no longer overrepresented here. At the same time, fewer boys were observed at the upper achievement levels at age 16 than at age 11.

In science, at age 7, boys were significantly overrepresented at the upper (*d* = -0.11) and lower tails (*d* = -0.27). 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 (*d* = -0.02) and lower grades (*d* = -0.03) 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 (*d* = -0.22). At age 16, in science and biology, we find a very small, significant female overrepresentation in the upper tail (upper 13% in science (*d* = 0.18) and biology (*d* = 0.15)). 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% (*d* = -0.16) and 5% (*d* = -0.15).

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.

3.5 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.

3.5.1 Longitudinal Changes of Sex Differences in Mean School Grades

Girls achieved significantly higher school grades in English, reading, and writing at each age. The difference was of small magnitude during primary education and moderate during secondary school. In mathematics, we found a very small 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 very small female advantage in science, biology, and chemistry. The English and science results fully support our hypotheses H1a and H1c, which state that there will be a significant female advantage in language subjects at all ages, that this will be wider at age 16, and that a female advantage will emerge in science at age 16 despite parity in primary education. Our Hypothesis H1b is partially supported, as the significant male advantage at age 11 was unexpected, but the expectation of parity in mathematics at ages 7 and 16 was supported.

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 with age in English and reading, as sex differences reduced at the lower achievement levels. In contrast, science and writing remained stable 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 standardised 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 Appendix A, Table A.8), 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-Cimpian et al., 2014). 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, such as ethnic group membership (Hsieh et al., 2021; Strand, 2010). Also, Strand (2010) reported that boys made more progress than girls between ages 7 and 11 in the UK context; male progress in mathematics, along with the reduced overrepresentation of boys at the lower achievement levels, may underlie this finding.

As the age 11 male advantage is eliminated by age 16, and overrepresentation of either sex at the tails is reduced, 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). The increased importance and orientation toward peers may influence adolescents' endorsement of gender stereotypes, as adolescents may become increasingly influenced by the observed achievement of their peers, and self-appraisals of ability in comparison to their peers, rather than the stereotypes (Kurtz-Costes et al., 2014; Marsh & Hau, 2003). Additionally, future educational or career plans develop and crystallise 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.

3.5.2 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 was, therefore, partially supported as we expected that boys' grades would be more variable than girls' grades at all ages. In line with prior reports and our Hypothesis H3a, sex differences were smaller in science and mathematics than in English, reading, and writing. Hypothesis H3b was partially supported. Girls' grades were more variable in mathematics and science than in English at age 11, but girls' STEM grades were not consistently more variable than their non-STEM grades at age 7. Boys' grades were more variable in English and reading at age 11 than in mathematics, whereas we expected boys' grade variability to be similar across subjects. 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 were 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 Appendix A, pg. 148–153).

3.5.3 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 negatively influence girls' outcomes in mathematics and science. 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 favouring 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). However, the age 7 results may also be influenced by teacher grading biases, which have been reported to align with gender, ethnicity, SEN, and socioeconomic stereotypes (Burgess & Greaves, 2013; Carlana, 2019; Lavy & Megalokonomou, 2019; Lavy & Sand, 2018; Terrier, 2020). The age 7 grades are wholly teacher-assessed, whereas the age 11 and age 16 grades are the results of anonymised, standardised, curriculum tests that are marked externally. It is therefore plausible that stereotypes influence the grades reported by teachers at age 7. Prior reports from this cohort have reported teacher bias effects when comparing the cohort members' scores on cognitive ability tests with teachers' subjective judgements of their ability (Campbell, 2015). Teacher biases have been reported in teacher-assessed grades when compared with standardised curriculum test grades in other samples (Burgess & Greaves, 2013; Lavy &

Megalokonomou, 2019; Lavy & Sand, 2018). In Appendix A, Section A.3, we compared the teacher-assessed grades with the standardised curriculum test grades at age 11, concluding that while there may be some evidence of teacher bias, effects were small, the evidence was not conclusive, and that the conclusions drawn from teacher-assessed grades did not substantively differ from those drawn from the test grades. However, given prior evidence, we cannot discount that teacher grading biases may influence the age 7 grades reported here.

The fact 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 have 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 (Gallagher et al., 2020; 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 to 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-16 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 prioritise 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-prioritisation 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., 2017, 2022; 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., 2017, 2022). Evidence in this area is sparse and requires further investigation to establish how puberty may be linked to educational outcomes.

3.5.4 School Grades vs. International Assessments

The discrepancies between sex differences in school grades, TIMSS, and PISA scores in mathematics and science tests and other standardised 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 standardised tests (Wu, 2010). There are also notable differences in the balance of content between different standardised 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, 2013). 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 centres around sex differences in learning styles (Kenney-Benson et al., 2006). Girls prioritise 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 standardised aptitude tests, contributing to the male advantage (Kenney-Benson et al., 2006). Finally, we must consider whether PISA and other international standardised tests exhibit similar female underprediction effects as observed in other standardised aptitude tests (Kling et al., 2013; Mattern & Patterson, 2014). PISA has been shown to predict school exam grades, those at risk of dropping out of school, and future earnings (Fischbach et al., 2013; Pulkkinen & Rautopuro, 2022). PISA scores have also been shown to have similar predictive ability as the SATs, even when controlling for the socioeconomic circumstances (SES) of individuals (Fischbach et al., 2013; Mattern & Patterson, 2014). However, when predicting outcomes using PISA scores, SES, and ethnic background, girls are more likely to achieve the top grades in mathematics than boys, highlighting that PISA data may also underpredict female achievement (Matějů & Smith, 2015; Pulkkinen & Rautopuro, 2022). If this finding holds for UK PISA scores when compared to school grades, then the inconsistencies reported here may, in part, reflect a female underprediction effect. Facets of the Big Five personality trait conscientiousness (Costa et al., 2001) are reported to contribute to the female underprediction effect (e.g., self-discipline: Duckworth & Seligman, 2006). However, some facet-level sex differences in conscientiousness do not appear until emerging adulthood (e.g., self-discipline Soto et al., 2011), which is after the achievement of these school grades. In contrast, analyses of observer-reported personality traits find small female advantages in trait-level conscientiousness at ages 15 and 16 years old, which is driven by the facets order, dutifulness at both ages, and self-discipline and deliberation at age 15 only (De Bolle et al., 2015).

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 reported 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 socialisation, and greater male variability contribute to the overrepresentation of boys at the highest levels of mathematical reasoning ability (Benbow, 1988). 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. In mathematics and science at age 16, boys' grades are not significantly more variable than girls' grades. 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).

3.5.5 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 generalise 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 categorisation 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 the ages of 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 3.2 (see also Appendix A, Tables A.9-A.10) demonstrate that at age 7 teacher-assessed grades are less reliable than standardised 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 standardised school test outcomes not reflecting performance and achievement in class. In supplementary analyses (see Appendix A, Tables A.9-A.10, pg. 150-151), we compared age 11 teacher-assessed grades and standardised 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 standardised 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 undersampled. 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.

3.6 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.

4 Puberty and Age Associations with School Engagement.

4.1 Abstract

Previous work has identified puberty as a risk factor for academic underachievement. The mechanisms through which puberty and achievement are linked are not well understood. Building on two previous studies linking puberty to changes in achievement motivation and school engagement, this study examined the longitudinal association between parent and self-reported pubertal status with emotional and behavioural engagement and disengagement. This secondary data analysis examined how changes in perceived pubertal maturation, self-concepts, and educational expectations predicted school engagement for 5,617 Millennium Cohort Study cohort members (2,752 boys and 2,865 girls) between early and middle adolescence. Previous results indicated that biological puberty was more strongly associated with disengagement, especially in boys. We found that advanced perceived puberty was associated with decreased behavioural engagement in boys and increased emotional and behavioural disengagement in girls. High mathematics and language self-concepts were protective against losses in behavioural and emotional engagement and, to a lesser extent, against increases in behavioural and emotional disengagement. Higher educational expectations were also protective. Correlational analyses indicated that behavioural disengagement was more strongly associated with later achievement in mathematics and language school grades. SEN and low-income-specific associations with school engagement were also observed. For highly motivated individuals, small puberty associations may be outweighed by high self-concepts and educational expectations. Perceived puberty associations may indicate a socioemotional risk factor for students with low self-concepts,

special educational needs (SEN) or from low-income backgrounds during adolescence.

4.2 Introduction

Sex differences in academic achievement change during the adolescent years (Voyer & Voyer, 2014). Although girls continue to achieve higher grades on combined measures, such as grade point average (GPA) and language subjects, they also achieve higher grades, on average, in the sciences, despite no differences in primary school (Oakley et al., 2024; Voyer & Voyer, 2014). Voyer and Voyer's (2014) meta-analysis reports a similar trend in mathematics as in science. Our Study 1 results indicate a male advantage in mathematics at the end of primary education, but no differences at age 16 (Oakley et al., 2024). Despite the varying findings for primary school mathematics, the direction and magnitude of change in mathematics and science are similar. Together, these reports suggest an increasing female advantage in educational outcomes from the end of primary school to mid-secondary school. Currently, it is unclear why this occurs, particularly in subjects such as mathematics and science where boys typically are more confident, have more positive emotions, and achieve better results in standardised tests of mathematics and science skills (Baye & Monseur, 2016; Frenzel et al., 2007).

Girls are generally reported to be more engaged and conscientious learners, and these attributes each predict higher achievement (De Fruyt et al., 2008; Lam et al., 2012; Poropat, 2009; Van Houtte, 2023). These sex differences in engagement and conscientiousness are apparent throughout later childhood and into adulthood, although the differences between boys and girls may reduce during middle adolescence (De Bolle et al., 2015; Lam et al., 2012; Soto et al., 2011). Some evidence points to links between pubertal maturation and educational outcomes, indicating that early-maturing girls and late-maturing boys are at risk of lower achievement. Few studies have identified the mechanism(s) through which pubertal maturation may influence educational achievement. This study builds on two recent reports linking puberty with achievement motivation and school engagement, using longitudinal panel study data collected in early and mid-adolescence (Martin et al., 2017, 2022). This study examines the longitudinal association between adolescent development and school engagement, asking whether puberty and school engagement are related and whether this association differs between boys and girls. Academic self-concepts and educational goals are theorised as predecessors of school engagement. Therefore, this study asks how changes in mathematics and language self-concept influence changes in school engagement. Finally, this study also asks whether the association between puberty and school engagement contributes to the widening female advantage observed during secondary education.

Early pubertal development in girls has been linked with academic underachievement and fewer years in education (Gill et al., 2017; Mendle et al., 2007; Stattin & Magnusson, 1990). Some researchers suggest that this may be due to a connection between maturation and problem behaviour or psychosocial mechanisms linked to socioeconomic factors (Cavanagh et al., 2007; Goering et al., 2023; Koivusilta & Rimpelä, 2004; Martin et al., 2017; Mendle et al., 2007; Simmons & Blyth, 1987; Stattin & Magnusson, 1990). Early maturing girls report less interest in academic subjects and are less likely to continue in education after compulsory schooling ends (Cavanagh et al., 2007; Gill et al., 2017; Stattin & Magnusson, 1990). However, the association between early maturation and low achievement is not consistently found across studies (Dubas et al., 1991; Graber et al., 1997; Senia et al., 2018; Stattin & Magnusson, 1990). Other studies have linked later maturation in girls with higher grades, suggesting that this may be explained by late-maturing girls achieving higher results in spatial reasoning tasks (Dubas et al., 1991; Suutela et al., 2022; Torvik et al., 2021). Koerselman and Pekkarinen (2018) associated later puberty with the slower growth of cognitive skills in boys and girls, but Paus et al. (2017) concluded that, in general, cognitive differences in boys and girls matured but remained stable during adolescence. While there is some disagreement between some studies for girls, most point to lower achievement for early maturing girls.

There is less evidence for boys and less agreement between studies.

The relationship between male puberty and achievement is not as well documented, partly due to measurement difficulties; the current evidence is inconclusive (Goering et al., 2023; Mendle & Ferrero, 2012; Suutela et al., 2022; Torvik et al., 2021). Some studies find that late-maturing boys are at risk of underachievement (Dubas et al., 1991; Koivusilta & Rimpelä, 2004; Mendle & Ferrero, 2012; Torvik et al., 2021). In contrast, other research finds no association (Graber et al., 2004; Graber et al., 1997; Senia et al., 2018; Simmons & Blyth, 1987; Suutela et al., 2022) or reports lower achievement among early and late maturing boys (Goering et al., 2023). While the findings linking advanced puberty and achievement are somewhat mixed for boys, and girls to a lesser extent, there is substantial evidence linking early puberty to adverse outcomes that include substance use, psychological distress, and delinquency (Ge et al., 1996, 2001; Goering & Mrug, 2022; Mendle & Ferrero, 2012; Mendle et al., 2007; Senia et al., 2018), which may be exacerbated for those with pre-existing problems before reaching puberty (Caspi & Moffitt, 1991; Torvik et al., 2021). Many of these negative outcomes are themselves associated with poor educational achievement (e.g., Bugbee et al., 2019; Gottfried, 2009; Hoffmann, 2020; van Tetering et al., 2022), therefore it is likely that the association between puberty and academic achievement may be indirect rather than direct.

Few studies have attempted to explain why puberty and school achievement may be related more generally, beyond the explanations offered by puberty links to poor mental health, substance use, and problem behaviours. Puberty is one aspect of the physical maturation process that takes place during adolescence, which is itself a set of interrelated and interacting processes encompassing physical, cognitive, and socioemotional change (Peper & Dahl, 2013; Sisk & Foster, 2004). Adolescence is an important, adaptive developmental period during which changes in social behaviour and developing self-identity influence decision-making and self-appraisals (Crone & Dahl, 2012). Adolescents are especially oriented towards their peers and influenced by others' perceptions of them. Adolescence is also an important period of brain reorganisation that does not complete until several years into adulthood. During this period, reward regions of the brain are more highly activated, while cognitive control systems develop linearly with age (Crone & Dahl, 2012; Spear, 2000; Spear, 2013). Dual systems theories of adolescence highlight a gap between the development of cognitive control systems and a tendency toward increased sensation seeking - the tendency to seek out and take risks to try new experiences - during early to mid-adolescence (Shulman et al., 2014; Shulman et al., 2016; Steinberg et al., 2018; Zuckerman, 1994). As a result, early maturers are at an increased risk of engaging in risky behaviours, with risky behaviours more likely to take place in the presence of peers.

While dual systems theories explain a propensity towards risky behaviour during adolescence, they do not account for processes related to self-identity development - an important task of adolescence - and how these may influence and motivate behaviour (Klimstra et al., 2010; Pfeifer & Berkman, 2018). Pfeifer and Berkman (2018) propose that, in addition to the developmental changes resulting in increased sensitivity to rewards and the social context, evolving self-identity also drives motivated behaviour. Pfeifer and Berkman (2018) describes adolescent decision-making as taking a value-based approach, in which gains and costs are weighed to make choices, whereby behaviours related to self-identity are more highly valued (Identity-Value Model: Berkman et al., 2017). Pfeifer and Berkman (2018) illustrates this mechanism using the example of an adolescent with a strong academic identity, who may increase the value of studying, vs. the adolescent who more highly values their peer group identity, who may increase the value of socialising when making decisions. Supporting this proposal, self-referencing information is remembered more accurately than other information due to increased encoding of information related to the self in memory during adolescence. Further, neural activity associated with social self-evaluations may increase with both age and perceived puberty status (Dégeilh et al., 2015; Pfeifer et al., 2013).

Recent evidence finds that advanced puberty is associated with lower achievement motivation and higher school disengagement (Martin et al., 2017, 2022). In addition, Martin et al. (2017) reported that the link between puberty and academic achievement was indirect via achievement motivation (academic self-efficacy and valuing). Educational achievement motivation is an important predecessor to school engagement, as described by the self-system model of motivational development (Green et al., 2012; Schnitzler et al., 2021; Skinner, 2023; Skinner et al., 2008, 2009; Veiga et al., 2015). The self-systems model emphasises that school engagement, which directly predicts educational outcomes, results from a set of perceptions and cognitions about the self, which are formed through socialisation processes (Skinner et al., 2008, 2009). These self-appraisals include evaluations of competency (self-concepts), school connectedness, perceptions of autonomy, valuing of school, and educational or career aspirations/goals (Green et al., 2012; Skinner & Raine, 2022; Skinner et al., 2008, 2009). During adolescence, self-appraisals are influenced by increased social and peer awareness (Huguet et al., 2009; Jansen et al., 2022; Marsh et al., 2005; van Tetering et al., 2020). Emotions and behaviours are, therefore, influenced during adolescence as a result of socioemotional changes, maturing self-identity and self-appraisals, heightened sensitivity to rewards, and the heightened risk-taking behaviours (Crone & Dahl, 2012; Galván, 2013; Mendle, 2014; Nelson et al., 2016; Pfeifer & Berkman, 2018; Shulman et al., 2016). Neuroscience evidence has linked perceived puberty status with social self-evaluation-related neural activity, and hormone levels with increased neural activity when envisaging social emotion-inducing scenarios (Klapwijk et al., 2013; Pfeifer et al., 2013). Given this evidence, it is plausible that socially-influenced self-appraisals, which include self-concepts, educational goals, and perceptions of school connectedness and autonomy, may each be predicted by puberty in addition to age, leading to direct and indirect associations between puberty and school engagement (Marsh et al., 2005; Skinner & Raine, 2022; van Tetering et al., 2020).

From the current literature, we know that school engagement declines with age, particularly close to school transitions in early adolescence (Wang et al., 2015). According to stage environment fit theory, closely related to the self-system model of motivational development, this decline is due to a mismatch between a student's developing needs and reduced opportunities to support these needs within the school environment (Eccles & Roeser, 2013; Eccles et al., 1993; Reeve, 2012; Skinner et al., 2009). Adolescents' increasing need for autonomy, involvement in decision-making, orientation toward their peers (and away from the home environment), and ongoing self-identity development often conflict with the increasing emphasis on academic performance. How students perceive their interactions with teachers and other students and their ability to succeed provides adolescents with insight into how connected, accepted, competent, and autonomous they perceive themselves to be in the school setting. In combination with a school environment featuring fewer individualistic teaching practices, opportunities for autonomy, and more interactions with more teachers who know students less well, these self-appraisals lead to patterns of engagement or disengagement (Eccles et al., 1993; Reeve, 2012; Skinner et al., 2008, 2009).

Originally conceptualised as three dimensions - behavioural, emotional, and cognitive - the structure of school engagement remains under discussion in the literature. In the original conceptualisation, school engagement encompassed a student's active participation in learning or related activities and reflected their emotions toward school (Fredricks et al., 2004). Since then, agentic and social dimensions have been added, and emotional engagement has been critiqued (Eccles & Wang, 2012; Fredricks et al., 2004; Pekrun & Linnenbrink-Garcia, 2012; Reeve, 2012; Wang et al., 2011, 2017). Behavioural engagement refers to participation in school-based learning and activities, and emotional engagement refers to positive and negative emotions related to school, teachers, and peers (Fredricks et al., 2004). Cognitive engagement is defined as the willingness to invest time and effort to understand or master concepts or skills (Fredricks et al., 2004). Agentic engagement measures an individual's contribution to their learning, while social engagement assesses social behaviours and participation within and outside the classroom (Reeve, 2012; Wang et al., 2017). More recent theoretical perspectives separate emotional responses from the core tenets of school engagement, as they represent emotional reactions to experiences rather than the action of being invested and engaged at school, and are therefore motivational (Eccles & Wang, 2012; Pekrun & Linnenbrink-Garcia, 2012). Others have argued that the emotional evaluations of school encompassed within emotional engagement should be retained, as they represent a level of investment in school (Symonds & Hargreaves, 2016).

More recent work has concluded that engagement and disengagement are different constructs within each dimension of school engagement (Fredricks et al., 2019; Skinner et al., 2008, 2009; Wang et al., 2017). Despite this, it is less common to find engagement and disengagement measured as different constructs in the literature, except for work that dichotomises school engagement/disengagement without distinguishing between dimensions (e.g., Martin et al., 2022). Most studies examine engagement/disengagement on a single continuum with lower engagement scores indicating disengagement (Fredricks et al., 2019; Wang & Degol, 2014). Engagement, defined as active, sustained, and directed action toward learning, is associated with higher achievement and better psychological adjustment. Conversely, disengagement, described as a withdrawal from learning, is linked with truancy, problem behaviour, and dropping out of education (Li & Lerner, 2011; Skinner et al., 2009; Wang & Peck, 2013). When engagement and disengagement are examined as separate constructs, engagement is more strongly associated with achievement than disengagement (Wang et al., 2017). However, few studies have examined changes in school engagement during secondary education after the initial transition from primary school (Fredricks et al., 2019).

There is some discussion on the relative importance of each dimension of school engagement to school achievement. Behavioural engagement is reported to be more strongly associated with achievement in secondary education than emotional engagement, despite larger declines in emotional engagement during adolescence (Wang et al., 2011; Wong et al., 2024). The relative importance of each dimension may also change depending on the age of the students, but there is less evidence in primary age samples (Bae & DeBusk-Lane, 2019). There are also group differences in behavioural and emotional engagement between girls and boys, ethnic groups, high and low income groups and students with identified special educational needs (SEN) (O'Donnell & Reschly, 2020; Tomaszewski et al., 2020; Wang et al., 2011). The relative importance of each dimension also differs between Eastern and Western samples (Lei et al., 2018; Wang & Peck, 2013). Profiles and trajectories of school engagement can vary between individuals, while school engagement and educational achievement are reciprocally related (Chase et al., 2014; Li & Lerner, 2011).

4.2.1 Present Study

This study builds on prior work that associated adolescent development with changes in achievement motivation and school engagement (Martin et al., 2017, 2022). By examining how parent and self-reported pubertal development influences school engagement during early to middle adolescence, this study aims to clarify the association between perceived puberty and school engagement. As prior work has reported different associations between puberty measures and school achievement, motivation and engagement, the puberty-school engagement associations are examined separately for girls and boys (Martin et al., 2017, 2022). Using the self-systems model of motivational development as a theoretical foundation, we examine the influence of self-appraisals on school engagement, here in the form of self-concepts at ages 11 and 14 and educational expectations at age 14.

Prior research has demonstrated that self-concepts can explain additional variance, beyond motivational factors, in predicting school engagement (Green et al., 2012). Sex differences in mathematics and language self-concepts are established within the literature, with girls reporting higher language self-concepts and boys reporting higher mathematics self-concepts (Huang, 2013). Associations between self-concepts and achievement are gender invariant, and self-concepts are influenced by prior achievement (Huang, 2011; Kriegbaum et al., 2018). A student's educational expectations, as opposed to educational aspirations, represent an assessed likelihood of achieving an educational goal or outcome (Pinquart & Ebeling, 2020). Educational expectations are influenced by self-concepts of ability and their parents' expectations of them (Buchmann et al., 2022). Educational expectations are sometimes reported to differ between boys and girls - SEN status and parental occupations can contribute to these differences (Dockery et al., 2022; Koshy et al., 2019; Ortiz-Gervasi, 2020; Rampino & Taylor, 2013). To date, the literature is divided on whether educational expectations are more influential for boys or girls (Flouri & Hawkes, 2008; Pinquart & Ebeling, 2020; Zhang et al., 2011). Parental educational expectations influence their child's educational expectations, can tend towards being slightly higher than their child's educational expectations, and are more strongly associated in families from higher socioeconomic groups (Pinquart & Ebeling, 2020). Parents from lower socioeconomic groups often report lower educational expectations of their children (e.g., Pingault et al., 2015). This study examines the expectation of going to university in the future, which will be a motivational factor for some, but not all, students. For other students, it will represent a discarded option due to lower self-appraisal of the ability to succeed or due to social or financial barriers (e.g., Fagence & Hansom, 2018).

As there is limited evidence to date linking school engagement with pubertal maturation, this study tests the association between perceived puberty and school engagement against the prior findings for puberty hormones, hypothesising that similar results will be found while bearing in mind that perceived puberty and biological puberty measures (e.g., hormone levels) may elicit different results (Martin et al., 2017, 2022; Torvik et al., 2021). The survey measures used in the MCS also predate much of the debate on the structure and composition of school engagement. Necessarily, the study continues with the original conceptualisation of school engagement, bearing in mind that engagement and disengagement are more recently examined as separate constructs (Fredricks et al., 2004, 2019; Wang et al., 2017). This study, therefore, examines the influence of pubertal development on behavioural and emotional engagement and disengagement between ages 11 and 14 years old, hypothesising that:

> H1. Perceived puberty is associated with greater decreases in engagement and greater increases in disengagement, over and above the association with age.
>
> H2. Perceived puberty status is a stronger predictor of school disengagement than engagement.
>
> H3. There will be a stronger association between puberty and school disengagement for boys than for girls.
>
> H4. Higher mathematics and language self-concepts, as well as educational expectations, will positively influence school engagement and disengagement.
>
> H5. School engagement will be negatively associated with a student's SEN status and positively predicted by higher income, parental educational level, and prior academic achievement.

4.3 Methods

This study uses data collected in two survey waves of the Millennium Cohort Study (MCS) combined with linked data from the National Pupil Database (NPD: Department for Education, 2021). The NPD contains detailed information on the school achievement of all pupils in England, Wales and Northern Ireland throughout their formal education. The extract used here is for students educated in England only (Department for Education, 2021). The MCS data analysed here is survey data collected at eleven years old (Age 11, MCS5: UCL Centre for Longitudinal Studies, 2021a) and fourteen years old (Age 14, MCS6: UCL Centre for Longitudinal Studies, 2021b), referred to as the Age 11 and Age 14 waves for the remainder of this paper.

Data files from the Age 11 and 14 MCS survey waves were merged to produce a single, wide-format row per cohort member for each age group: Age 11 and Age 14. Similarly, NPD-linked data files were merged to create a long-format data set with one row for each cohort member, subject and assessment type (Key Stage 2 - Age 11, GCSE - Age 16). Finally, survey and educational data were merged, matching Age 11 achievement with Age 11 survey data and combining Age 14 survey data with Age 16 educational outcomes data. Chapter 2 describes each dataset further and provides details of the data merging and cleaning processes that took place prior to commencing data analysis.

Data cleaning, transformation, and analyses were performed using R version 4.3.6 with RStudio 2024.04.1 for Windows (R Core Team, 2021).

4.3.1 Participants

The sample consisted of 5,617 members of the Millennium cohort study (2,752 male, 2,865 female). As intended in the survey design, the sample is more ethnically diverse than the UK population was when the cohort members were born, as reported in Chapter 3 (Table 3.1). Due to a combination of the sampling age of the cohort members and the logistics of the survey, the cohort member age range when interviewed varied by up to 22 months (Age 11: $M = 11.2$, $SD = 0.33$; Age 14: $M = 14.3$, $SD = 0.34$). Similarly, parent-reported school year (grade) also varied. Most cohort members were in school year 6 (n = 5,446, Age 11 wave) or year 9 (n = 5036, Age 14 wave) when surveyed however some cohort members were in years 7 (n = 137) at the time of the Age 11 wave and year 10 (n = 518) when surveyed for the Age 14 wave. Age was centred at ten years old for the analyses.

4.3.2 Measures

Puberty Score. Pubertal development was reported by the parents of cohort members in the Age 11 wave and by the cohort members in the Age 14 wave, using the Pubertal Development Scale (PDS: Carskadon & Acebo, 1993; Petersen et al., 1988). The scale

Table 4.1

Distribution of Puberty Stage by Sex for each Survey Wave

Wave	Sex	Pre	Early	Mid	Late	Post	NR
Age 11	M	1647 (59.8%)	599 (21.8%)	381 (13.8%)	20(0.7%)		105 (3.8%)
	F	458 (16.0%)	481 (16.8%)	1469 (51.3%)	288 (10.1%)		169 (5.9%)
Age 14	M	27 (1.0%)	230 (8.4%)	1440 (52.3%)	900 (32.7%)	20 (0.7%)	135 (4.9%)
	F	12 (0.4%)	13 (0.5%)	244 (8.5%)	2261 (79.9%)	219 (7.6%)	116 (4.0%)

Note. Each individual's puberty stage was calculated as described in Petersen et al. (1988). Individuals are categorised as pre-pubertal, early pubertal, mid-pubertal, late pubertal, or post-pubertal. NR = nil response, or categorisation could not be calculated from the responses given.

consists of five items, which differ for boys and girls. Three items are common to boys' and girls' pubertal development: pubic hair, skin changes, and the growth spurt. The remaining two items are sex-specific: facial hair and voice change for boys, breast growth and the start of menstruation for girls. In the Age 11 survey, parents were asked whether these features had started to change, with responses recorded using a 1-3 ordinal scale (1 - not started, 2 - barely started, 3 - definitely started). Cohort members were asked for their responses using a 1-4 ordinal scale in the Age 14 survey (1 - not yet begun, 2 - barely started, 3 - definitely started, 4 - changes seem completed). For girls, menstruation is reported as a yes / no question (No = 1, Yes = 4). The puberty (PDS) score was calculated by averaging the responses to the items. The original PDS scale was reported to have good reliability (parent: α = 0.68 - 0.78; student: α = 0.67 - 0.70 Carskadon & Acebo, 1993) and has been evaluated as a valid measure of physical development when more objective measures are unavailable, that correlates with hormone levels (Brooks-Gunn et al., 1987; Schmitz et al., 2004; Shirtcliff et al., 2009). Large mean sex differences in pubertal status were reported in each survey wave, with boys' pubertal status being less variable in the Age 11 survey and more variable in the Age 14 survey, when compared to girls (Age 11: boys' *M* = 1.45, *SD* = 0.42, girls' *M* = 1.93, *SD* = 0.64; Age 14: boys' *M* = 2.56, *SD* = 0.53, girls' *M* = 3.09, *SD* = 0.45). For an alternate comparison, the distribution by puberty stage is provided in table 4.1. Each cohort member's puberty stage was calculated as described by Carskadon and Acebo (1993) using the method originally proposed by Crockett (1988). These categories were not used in the analyses. The use of continuous measures of pubertal status is preferred, as statistical models can more accurately capture individual differences in (perceptions of) maturation tempo over time (Mendle & Koch, 2019). Their calculation is described in Appendix D, Section C.1.

Table 4.2

School Engagement Measures following Confirmatory Factor Analysis

Measure How much do you?	Dimension
Try your best at school	Behavioural engagement
Misbehave in class	Behavioural disengagement
Find school interesting	Emotional engagement
Feel school is a waste of time	Emotional disengagement
Feel unhappy at school	Emotional disengagement
Feel tired at school	Emotional disengagement

Note. Factor dimensions were identified using Wang et al. (2017).

School Engagement. Six measures of school engagement were available across the two MCS waves. The set of questions did not match a complete published scale. The school engagement questions were compared to published school engagement scales to identify which school engagement dimensions were measured (see Table 4.2). From comparing the MCS questions to those used in other studies, it was evident that not all dimensions were measured and that three of the six questions measured emotional disengagement. With the aim of avoiding the use of single-item measures, confirmatory factor analysis (CFA) was used to clarify how best to structure the school engagement measures in the analyses. The CFA is described further in Appendix C, Section C.1. As the CFA results indicated that the measures could not be combined reliably into either a dual-factor model representing behavioural and emotional dimensions, or engagement and disengagement, the six measures were therefore retained. Behavioural engagement, behavioural disengagement, and emotional engagement were operationalised as single-item measures. Emotional disengagement was operationalised as three items, which were analysed as correlated, multivariate outcomes in the analyses. Cohort members were asked to respond to each question using an ordinal 1-4 scale: 1 - All of the time, 2 - Most of the time, 3 - Some of the time, 4 - Never. The measures were reversed so that a higher numeric value indicated higher engagement or disengagement. Descriptive statistics for the six school engagement measures are provided in Table 4.3.

Table 4.3

Descriptive Statistics for Boys and Girls: School Engagement.

Measure	Wave	Sex	*n*	*M*	*SD*
Emotional Engagement	11	M	2752	2.78	0.68
		F	2865	2.93	0.66
	14	M	2752	2.50	0.67
		F	2865	2.47	0.67
Emotional Disengagement: Tired at School	11	M	2752	2.07	0.79
		F	2865	1.97	0.72
	14	M	2752	2.33	0.78
		F	2865	2.55	0.82
Emotional Disengagement: Unhappy at School	11	M	2752	1.81	0.65
		F	2865	1.76	0.63
	14	M	2752	1.72	0.68
		F	2865	1.90	0.73
Emotional Disengagement: Waste of Time	11	M	2752	1.69	0.81
		F	2865	1.41	0.66
	14	M	2751	1.75	0.77
		F	2865	1.76	0.79
Behavioural Engagement	11	M	2752	3.47	0.59
		F	2865	3.63	0.53
	14	M	2752	3.17	0.59
		F	2865	3.28	0.62
Behavioural Disengagement	11	M	2752	1.66	0.63
		F	2862	1.34	0.52
	14	M	2688	1.69	0.62
		F	2867	1.54	0.63

Note. Responses were recorded using an ordinal 1-4 scale.

Academic Self-Concept. At ages 11 and 14, cohort members were asked to rate their ability (“I am good at”) in mathematics and language (English) using a 1-4 Likert-type scale: 1 - Strongly disagree, 2 - Disagree, 3 - Agree, 4 - Strongly agree.

Educational expectations. In the age 11 survey, parents of cohort members were asked to indicate the likelihood that their child will attend university in the future. The answers were given on an ordinal scale (1 - 4): 1 - Very likely, 2 - Fairly likely, 3 - Not very likely, and 4 - Not at all likely. The parental response was reverse-coded. In the age 14 survey, the cohort members indicated how likely they were to attend university in the future as a percentage. The cohort member percentage likelihood was converted to a 1-4 scale for consistency between measures - responses were divided by $33\frac{1}{3}$, adding 1 to each resulting score.

Descriptive statistics for academic self-concepts and educational expectations are provided in Table 4.6.

Covariates. As sex, socioeconomic, and SEN status group differences in school engagement, self-concepts, and educational expectations have been reported, and achievement and self-concepts are reciprocally related, this study examines the influence of these variables - with family income and parental educational qualifications representing socioeconomic status - on changes in school engagement over time (Huang, 2011; Martin et al., 2017).

Prior Achievement. Prior achievement was calculated as the average of mathematics and language (English) fine grades in national, standardised curriculum tests taken at the end of primary education (boys' *M* = 4.86, *SD = 0.68*, girls' *M* = 4.89, *SD = 0.64*, *d* = 0.05[0.00, 0.10]). Raw test scores were standardised (*M* = 100, *SD* = 10). From these standardised scores, grade boundaries are set (range 0 - 6). For those scores within the boundaries of grades 3, 4, or 5, a fine grade is calculated as the grade plus a proportion of the standardised score within the upper the lower grade boundaries. For scores at grade 2 or below, teacher-assessed grades are used to set the fine grade. If the teacher-assessed grade is greater than 2, then the fine grade is calculated as grade 3 less a proportion of the standardised score within the upper and lower grade 3 boundaries. Further details of this calculation can be accessed in the UK's Digital Education Resource Archive (Department for Education, 2016).

Table 4.4

Sample Distribution by Parent Educational Qualifications

NVQ Level	*n*	Proportion	Description
1	158	2.8%	3-4 GCSE grades D-G
2	893	15.9%	4-5 GCSE grades A*-C
3	648	11.5%	2 A Levels
4	1972	35.1%	3+ A Levels, BTEC, Higher Educational Certificate
5	862	15.3%	Foundation Degree or higher
	1084	19.3%	Not provided or other overseas qualification

Parent Educational Qualifications. Parental educational qualifications were provided for both the main and partner-parent; therefore, the highest of these was calculated. The main parent, who was usually the mother of the cohort member, was the parent who completed two detailed surveys on their personal and family circumstances and a survey about the cohort member in each MCS wave. The partner parent completed a shorter survey that was limited to their personal circumstances, although often the partner parent survey was completed by the main parent by proxy. Parental qualifications were derived from detailed parent responses in the parent interviews and categorised by derived national vocational qualification (NVQ) levels 1 - 5, where 5 is the highest level. Parent responses of "none of these" and overseas qualifications that could not be matched back to an NVQ equivalent level were coded separately, but not used in these analyses.

Table 4.5

Sample Distribution by Disposable Family Income

OECD UK Quintile	*n*	Proportion of Sample
1	969	17.3%
2	984	17.5%
3	1107	19.7%
4	1257	22.4%
5	1282	22.8%
	18	0.3%

Family Income. Detailed family income data was reported by the main parent in each survey wave, from which the data owner calculated a derived OECD UK Income Quintile (ordinal 1-5, where 5 is the highest) for each family for each wave. The calculation was adjusted for the number of family members supported by the total family income. There was a statistically significant, but negligible, improvement in family disposable income between survey waves (MCS5: $M = 3.16$, $SD = 0.66$; MCS6: $M = 3.24$, $SD = 0.66$, $d = 0.06[0.02, 0.09]$). Despite MCS oversampling from lower socioeconomic geographic areas, due to attrition between waves and separate parental permission required to access and match educational records, the resulting sample has slightly more cohort members from the upper two income quartiles (see Table 4.5.

Special Educational Needs. These data include a relatively high proportion of students with identified special educational needs being taught in mainstream education (boys: 30.4%, $n = 836$; girls: 18.2%, $n = 520$), when compared to national statistics for the population sample (14.1% Department for Education, 2018). Due to this oversampling of SEN students and that SEN students are at risk of underachievement, we control for SEN status in these analyses and report how SEN status is associated with school engagement and achievement.

Age 16 Achievement Mathematics and language achievement in national, standardised, school achievement tests in which students were graded on a 1 - 9 scale, where 9 is the highest grade.

Table 4.6

Descriptive Statistics for Boys and Girls: Mathematics and Language Self-Concepts and Educational Expectations.

Measure	Wave	Sex	*n*	*M*	*SD*
Language Self-Concept	11	M	2680	3.05	0.68
		F	2815	3.20	0.67
	14	M	2691	2.97	0.71
		F	2823	3.07	0.69
Mathematics Self-Concept	11	M	2674	3.44	0.68
		F	2812	3.17	0.76
	14	M	2689	3.16	0.76
		F	2823	2.90	0.79
Educational Expectations	11	M	2791	3.05	0.85
		F	2814	3.21	0.79
	14	M	2566	2.97	0.86
		F	2892	3.18	0.81

Note. Self-concepts were reported using an ordinal 1-4 scale. Educational expectations are reported by the cohort member's parent in the Age 11 wave and by the cohort member in the Age 14 wave. Parents reported educational expectations of their child on an ordinal 1-4 scale. Cohort members reported educational expectations as a percentage likelihood, which was rescaled as described in the educational expectations paragraph. The rescaled value is reported here.

4.3.3 Analyses

Confirmatory factor analyses preceded the main analyses to clarify the factor structure of the available school engagement measures. These initial analyses were implemented using the lavaan R package and are described further in Appendix C, Section C.1 (Rosseel, 2012).

The main analyses consisted of three univariate and one multivariate, multilevel ordinal logistic growth models implemented under the Bayesian framework, clustered by individuals.

The models implemented were based on multilevel linear growth models implemented using the nlme R package as outlined by Grimm et al. (2016), but constructed instead using the brms R package (Bürkner, 2017; Bürkner & Vuorre, 2019). The multivariate outcome was fit using a Bayesian item response theory (IRT) growth model (Bürkner, 2021; Kurz, 2021). Growth modelling refers to statistical methods that aim to separate between and within-person change over time (Curran et al., 2010; Grimm et al., 2016). Individuals may have different trajectories; trajectories may be flat (no change over time), or systematically increasing or decreasing. There are two main approaches to fitting growth models to data: using a multilevel framework, as implemented here, or using structural equation modelling (Curran et al., 2010; Duncan et al., 2013; Grimm et al., 2016; McArdle & Epstein, 1987). The SEM package in R allows for fitting of growth models, but assumes that time points between surveys are equally spaced (Rosseel, 2012). This option is available in Mplus, as an alternative to lavaan in R. However, Mplus is not currently available in the secure environment where these analyses were implemented (Muthén & Muthén, 1998). As the constraint for equal time points between survey waves is not met in these data, a multilevel approach is appropriate. The choice to take a Bayesian approach through using the brms R package was made for pragmatic reasons. The brms package allows for a wide variation of models to be specified using the same framework, given the ordinal measures of school engagement with skewed response distributions (Bürkner, 2017, 2018, 2021; Bürkner & Vuorre, 2019). Multiple packages with different constraints would need to have been used to implement the multivariate and univariate ordinal regressions reported here under the frequentist framework within R. The Bayesian priors used in the analyses are described in Appendix D, Section D.2.

After clarifying puberty and age associations in unconditional models (see Appendix D, Section D.3), this study examines puberty and age associations with each measure of school engagement using two multilevel, longitudinal regression models. Model 1 examined the influence of puberty and age on school engagement when controlling for time-invariant SEN status, parental educational qualifications, and prior achievement, alongside time-varying family income. Model 2 examined how time-varying mathematics and language self-concepts and educational expectations were associated with the trajectories of school engagement from early to middle adolescence.

Separate models were implemented for boys and girls to examine whether the associations between puberty and school engagement measures differed between the sexes. There is a relatively small fraction of missingness (λ) among the variables used here across the two waves, which was lowest for the school engagement and achievement motivation variables ($\lambda \leq 0.01$) but was slightly higher for the puberty variables ($\lambda \leq 0.03$). As shown in Table 4.4, there was a substantial proportion of missingness in the parental qualifications variable ($\lambda \leq 0.1$). Imputed datasets were created using two-level multivariate imputation by chained equations (MICE) r package, with m = 10 iterations for the univariate models (mice: van Buuren & Groothhuis-Oudshoorn, 2011). The number of iterations needed to optimise both point estimates and standard errors has been discussed in the methodological literature. Previously, a common guideline was that m = 2 - 10 imputed datasets should be sufficient to optimise point estimates, although m may need to be increased when λ is very large (Bodner, 2008; von Hippel, 2020). Since then, Bodner (2008) examined how increasing λ influenced imputation variability - variance in confidence intervals half widths between independent imputed datasets - reporting that increasing m could counteract this effect. Bodner (2008) proposed a linear rule to calculate how many imputations would be required. Using the reported table specifying numbers of imputations required given a max λ = 0.1 among these data items, then the recommendation from this work is that m = 6 - 9 imputations should be sufficient (Table 4: Bodner, 2008). More recently, von Hippel (2020) countered that a quadratic rule is more appropriate, so that at $\lambda \leq 0.5$, the linear equation overestimates m, whereas at $\lambda \geq 0.5$, the linear equation underestimates m. Given the low level of missingness in these data ($\lambda < 0.1$ for all data items), m = 10 should be more than sufficient to ensure both point estimates and standard errors are optimised.

The male and female data were imputed separately, and the puberty development score was calculated after imputation. The MICE procedure imputes *m* versions of each missing data item based on the observed values for the individual and other participants in the dataset (Azur et al., 2011). Standard errors among the imputations will be smaller in samples where the observed data are strongly correlated with the missing values (Greenland & Finkle, 1995). While procedures can be used to prevent imputation of missing data where the data are not appropriate in some circumstances (for example, some puberty measures are sex-specific), as the literature predicts some sex-specific associations between some of the variables of interest (Martin et al., 2022), and the intention to analyse girls and boys data separately to examine sex-specific associations, the choice was made to impute boys and girls data separately.

An IRT model was used to analyse the predictors of emotional disengagement (Bürkner, 2021). These multivariate models required the data to be in a three-level structure, as the three emotional disengagement items are nested within each survey wave. Due to the higher computational processing cost for both the imputation and running the subsequent models, m = 5 iterations were imputed for the emotional disengagement imputed dataset using the ml.lmer extension to the mice package, enabling three-level imputation (miceadds: Robitzsch & Grund, 2022). As Bodner (2008) suggested m = 6 - 9 imputations for $\lambda \leq 0.1$, and von Hippel (2020) concluded that this was an overestimate, m = 5 imputations appears to be a reasonable compromise given the large increase in computing cost when running three-level Bayesian growth models on large datasets. Three-level emotional disengagement imputations were successfully carried out and tested for the boys. However, a technical issue arose before the imputations could be generated for girls, and subsequent attempts to impute the female data for emotional disengagement or repeat the imputation of the male data proved unsuccessful. The results reported below are, therefore, from the imputed datasets for emotional engagement, behavioural engagement, and behavioural disengagement, as well as the missing data dataset for emotional disengagement. The results from the datasets with missing data to coincide with the imputed alternatives are reported in Appendix D, Tables D.7-D.8.

Parental educational qualifications did not converge in the imputed models. As parental qualifications did not make a significant contribution to behavioural engagement, behavioural disengagement and emotional disengagement in the analyses with missing data, parental qualifications were excluded from the imputed models, and the models were re-run. Parental educational qualifications were a significant predictor of emotional engagement for girls only in the covariates-only model. However, the cohort members' parents' educational level was no longer a significant predictor in Model 2, which included educational expectations and domain-specific self-concepts. This may simply reflect the positive relation between parents' level of education and their income, and as such, parental educational level was less influential (Erola et al., 2016). Full results that include parental educational expectations are reported in the main text and Table C.4 for emotional disengagement, and in Appendix C, Tables D.7-D.8 for the remaining school engagement variables.

The Bayesian leave-one-out cross-validation (LOO) estimate of the log point-wise predictive density (ELPD) and pareto-smoothed importance sampling (PSIS) criteria are used for model fit assessment and comparison, as these assessments are more robust when priors are weak or in the case of influential observations (Vehtari et al., 2017). ELPD differences between models of less than 4 are considered small, and therefore, model predictions similar (McLatchie & Vehtari, 2023; Sivula et al., 2020). When ELPD differences between models are greater than 4, then the difference is compared to the standard error of the ELPD difference, which is assessed using pseudo-BMA (Bayesian Model Averaging) weights of the ELPD of each model (McLatchie & Vehtari, 2023). The ELPD difference 95% confidence interval is calculated to indicate the range of values for the difference between the ELPD value of each model (Sivula et al., 2020).

The Sequential Effect eXistence and sIgnificance Testing (SEXIT) guidelines are followed for describing Bayesian statistical results (Makowski et al., 2019). The SEXIT framework describes three criteria for evaluating regression model results. The medians of the posterior distributions of each fitted model were examined to understand the contribution of each predictor. A 95% HDPI (highest density probability interval) was calculated for each predictor, indicating the smallest interval in which there is a 95% probability that the population coefficient falls within this interval; therefore, it is a measure of uncertainty of an effect's true value. The probability of direction (*pd*) and significance are used to assess whether an effect is of sufficient size to be considered further, i.e., the effect is not negligible (Makowski et al., 2019). Ranging from 50 - 100%, the *pd* indicates the certainty of the direction of an effect as the proportion of the distribution that is in the same direction as the median. Significance indicates the probability that an effect is greater than the threshold of 0.05, below which an effect is considered negligible. Heuristics for the probability of direction are not yet fully evidenced. Therefore, the suggested 95%, 97%, and 99% reference points are used here (Makowski et al., 2019). As curvilinear associations with ordinal categorical predictor variables are automatically reported, to simplify results tables, curvilinear associations are only reported where they represent a non-negligible contribution to the outcome in at least one of the models for either boys or girls, as far as practical. The convergence and stability of Bayesian sampling were evaluated using $\widehat{R}$, which should be less than 1.01, and effective sample size (ESS), which should be greater than 1000 (Bürkner, 2017; Vehtari et al., 2021). Post hoc conditional/marginal effects were calculated and plotted to visualise and compare the associations of puberty and age on school engagement (see Figures 4.1 and 4.2) and the associations of mathematics and language self-concepts, and educational expectations, on school engagement (see Figures 4.3 and 4.4) (Bürkner, 2017).

Bivariate correlations are also reported between Age 14 school engagement, self-concepts, and educational expectations with subsequent achievement at age 16 in standardised mathematics and language school grades (Makowski et al., 2022). Concurrent Age 11 correlations are also briefly reported in the main text below, with reference to Appendix C. Similarly, transition effects are examined in simple post-hoc analyses reported below in Appendix C.

4.4 Results

4.4.1 Unconditional Model Testing

Initial unconditional multilevel growth models tested whether linear or quadratic functions of age or puberty best explained changes in school engagement. These models were tested on a random sub-sample of n = 1500 cohort members (735 boys, 766 girls) in unconditional models predicting emotional engagement, as larger differences in emotional engagement are reported during secondary education (Wang et al., 2011). The ELPD difference between the quadratic and linear models for age was lower than the threshold of 4, indicating that the quadratic model did not significantly improve model fit over the simpler linear models. For puberty, the ELPD difference was larger than the threshold, and the 95% CI did not cross zero, indicating that the model with the quadratic function for puberty provided a better fit for the data.

Given the strong correlations between age and puberty and the quadratic association between puberty and emotional engagement, it was hypothesised that the curvilinear association may result from an interaction between puberty and age. Four models were tested with random intercepts for cohort members as follows:

- Model A: Main effect of Age with Puberty/Age Interaction.
- Model B: Main effect of Puberty with Puberty/Age Interaction.
- Model C: Main effects of Puberty and Age with Puberty/Age Interaction.
- Model D: Main effects of Puberty and Age.

Comparisons of these four models using ELPD LOO indicated that there were significant differences between the models for the disengagement measures but not for the engagement measures (see Appendix D, Tables D.1 and D.2. For emotional disengagement, there was a clear preference for Model C. For behavioural engagement, the best model fit resulted from Model B. However, the fit was not significantly better than the simpler Model D. For emotional engagement, the preferred model was Model A. The fit for Model A and the other three models for emotional engagement did not significantly differ. Model C was discarded for emotional engagement as it resulted in no significant associations between age, puberty, or the puberty/age interaction. Behavioural engagement preferred Model C, but the Model C fit was not significantly better than the other alternatives.

Model C was chosen for emotional disengagement. Model B was initially selected for behavioural disengagement. However, when the models were fit separately for boys and girls, neither puberty nor its interaction with age made a significant contribution to changes in behavioural engagement. Given that the Model D fit did not significantly differ from the fit for Model B, this study reverted to the simpler Model D for behavioural engagement. Overall, the unconditional models indicated a stronger role for puberty in both disengagement measures and a stronger role for age in both engagement measures. In sex-specific models, these initial conclusions did not always persist.

4.4.2 Puberty, Age, and School Engagement

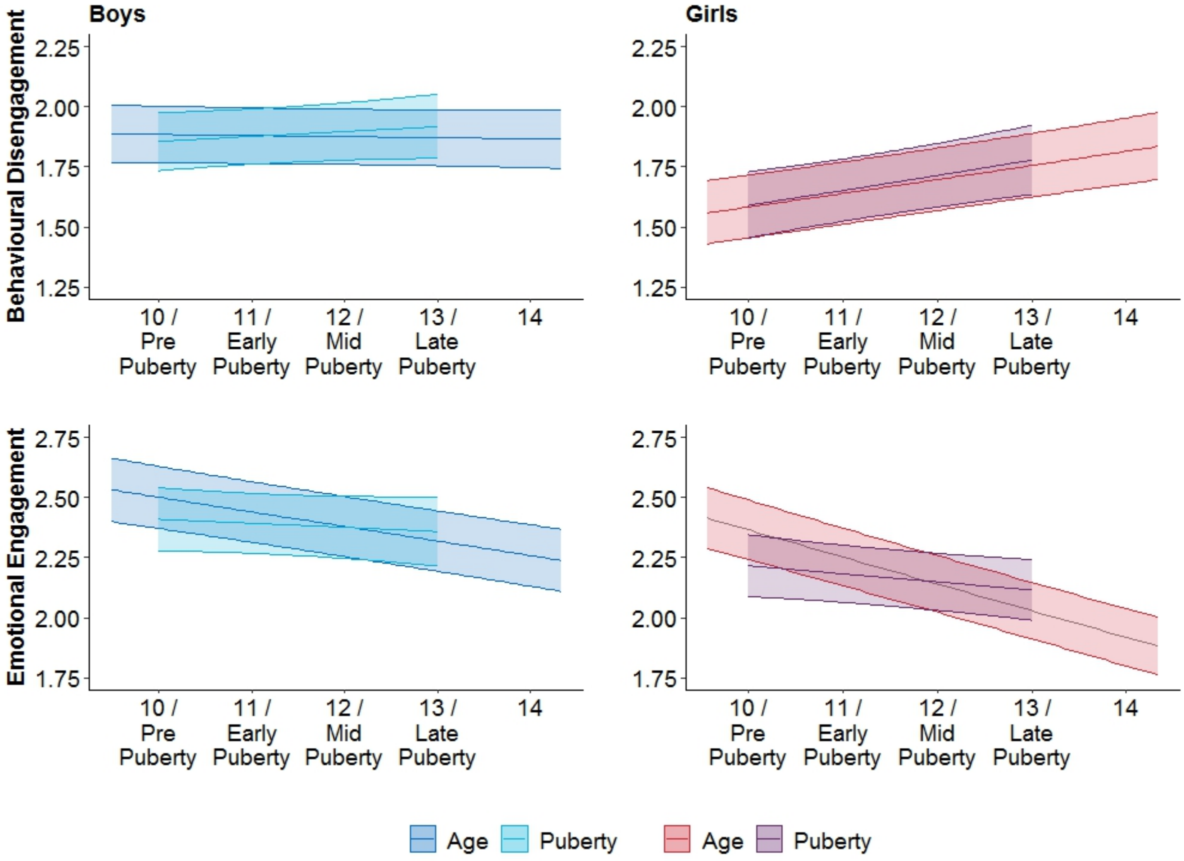
When examining the results of the two conditional models (Model 1 - Covariates only and Model 2 - Covariates + Self-Concepts + Educational Expectations; see Appendix C, Tables C.1 - C.4) for all four measures of school engagement, separately for boys and girls, there were different associations between puberty score and school engagement between the sexes. In partial support of the first hypothesis (H1) and full support of H2, girls' perceived puberty status was associated with increases in behavioural (Model 2, *Mdn*: 0.13 [0.05, 0.21]) and emotional disengagement where puberty and age interacted (Model 2, *Mdn*: 0.08 [0.05, 0.11]). Perceived puberty did not significantly influence girls' emotional or behavioural engagement. Increasing age predicted higher emotional (Model 2, *Mdn*: -0.23 [-0.26, -0.19]) and behavioural engagement (Model 2, *Mdn*: -0.12 [-0.22, -0.01]). Age did not significantly predict girls' emotional disengagement, and age and puberty contributed equally to girls' behavioural disengagement (Model 2, Age, *Mdn*: 0.12 [0.08, 0.16]). For behavioural disengagement, the association with puberty, beyond age, was additive, whereas the associations of puberty and age were multiplicative for emotional disengagement. As girls aged, puberty was associated with increasingly higher emotional disengagement from school. Increasing age was associated with declines in behavioural disengagement, and the association between age and emotional engagement was minimal.

For boys, the contribution of perceived puberty was practically significant for behavioural engagement only, where puberty interacted with age (Model 2, *Mdn*: -0.06 [-0.11, -0.02]). H1 was, therefore, partially supported, and H2 was unsupported for boys (see Appendix C, Tables C.1 - C.4 and D.7, and Figure 4.2 below, for details). There was an association between puberty and behavioural disengagement in the models with missing data (see Appendix D, Table D.7), but not in the results from the imputed datasets reported here. The association of puberty and behavioural engagement varied with age (see Figure 4.1). In the Age 11 survey, the association between puberty and behavioural engagement was minimal, with a tendency towards higher behavioural engagement for more mature boys with higher puberty scores. In the Age 14 survey, being less mature was linked to declines in behavioural engagement. There were significantly stronger declines for the older, more mature boys. Puberty was not a significant predictor of boys' emotional disengagement. Like girls, boys' emotional engagement decreased with age, but to a lesser extent (Model 2, *Mdn*: -0.12 [-0.16, -0.09]). However, age did not predict boys' behavioural engagement or school disengagement.

In sum, across the models for boys and girls, H1 was partially supported. Perceived

Figure 4.1

Conditional Main Effects of Puberty and Age on the Behavioural Disengagement and Emotional Engagement of Boys and Girls.

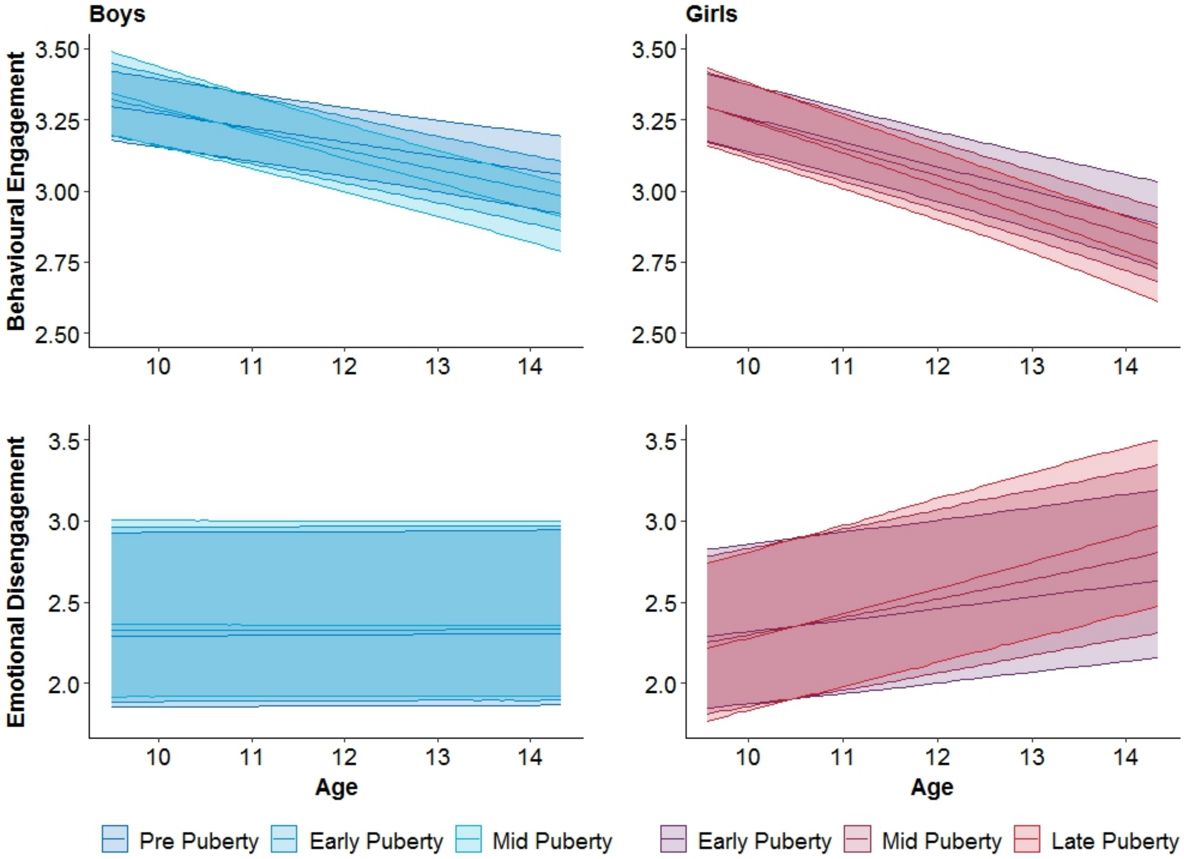


Note. Visualises conditional main effects extracted from fitted Model 2 using imputed datasets are shown on the same plot to allow for easier comparison. The x-axis indicates chronological age for age associations and puberty status category for puberty associations.

puberty was associated with greater increases in disengagement for girls, but puberty status did not significantly predict disengagement for boys. Perceived puberty was associated with greater declines in behavioural engagement (only) for boys, especially in mid-adolescence, but puberty status did not predict girls' school engagement. Perceived puberty was a stronger negative predictor of girls' disengagement only. Therefore, H2 was fully supported for girls and unsupported for boys. As perceived puberty did not predict boys' disengagement, but did predict girls' disengagement, H3 was not supported.

Figure 4.2

Conditional Interaction Effects of Puberty and Age on the Behavioural Engagement and Emotional Disengagement of Boys and Girls.



Note. Visualises conditional interaction effects extracted from fitted Model 2 using imputed datasets are shown on the same plot to allow for easier comparison. The x-axis indicates chronological age for age associations and puberty status category for puberty associations

4.4.3 Academic Self-Concepts, Educational Expectations, and School Engagement

Higher mathematics and language self-concepts, and educational expectations, were protective against declines in emotional and behavioural engagement and increases in emotional and behavioural disengagement for boys and girls, in full support of H4 (see Appendix C, Tables C.1 - C.4 and Figures 4.3, 4.4 below). Associations between self-concepts and school engagement were often curvilinear.

For boys, the associations between mathematics and language self-concepts with emotional and behavioural disengagement were similar and curvilinear, except for the association between mathematics self-concept and emotional disengagement, which was linear. For example, behavioural engagement: Model 2, mathematics self-concept, quadratic *Mdn*: 0.34 [0.21, 0.48]; language self-concept, quadratic *Mdn*: 0.23 [0.10, 0.35]; educational expectations: 0.16 [0.11, 0.21]). The curvilinear associations indicated that while there was a tendency for boys with higher self-concepts to report higher engagement, boys with the highest self-concepts did not always report higher behavioural or emotional engagement than boys with the lowest self-concepts, except for the association between language self-concept and emotional engagement (see Figures 4.3 and 4.4). There was even less evidence of incrementally lower school disengagement with higher self-concepts. Boys' educational expectations were more strongly related to their reported emotional engagement (*Mdn*: 0.30 [0.25, 0.34]) than their reported emotional disengagement (*Mdn*: -0.19 [-0.22, 0.15]), behavioural engagement (*Mdn*: 0.16 [0.11, 0.21]) or behavioural disengagement (*Mdn*: -0.13 [-0.19, -0.08]).

For girls, the associations between self-concepts and engagement measures were stronger than the association between educational expectations and engagement measures, particularly for emotional engagement. For example, for emotional engagement: mathematics self-concept, linear *Mdn*: 0.67 [0.56, 0.79]; language self-concept, linear *Mdn*: 0.63 [0.49, 0.77]; educational expectations *Mdn*: 0.29 [0.24, 0.34]). Girls' educational expectations were more strongly related to their reported emotional engagement than their reported emotional disengagement (*Mdn*: -0.15 [-0.19, 0.12]) or behavioural disengagement (*Mdn*: -0.16 [-0.22, -0.11]). Similar to the findings for boys, curvilinear associations were often found between girls' school engagement and self-concepts. This meant that while self-concepts were negatively related to emotional and behavioural disengagement, there was little evidence that the emotional or behavioural disengagement of girls with low self-concepts differed significantly from girls with high self-concepts. In contrast, for both engagement measures, the post hoc comparisons (see Figures 4.3 and 4.4) revealed that where the associations with self-concepts were stronger, girls with high self-concepts reported higher emotional and behavioural engagement than girls with low self-concepts.

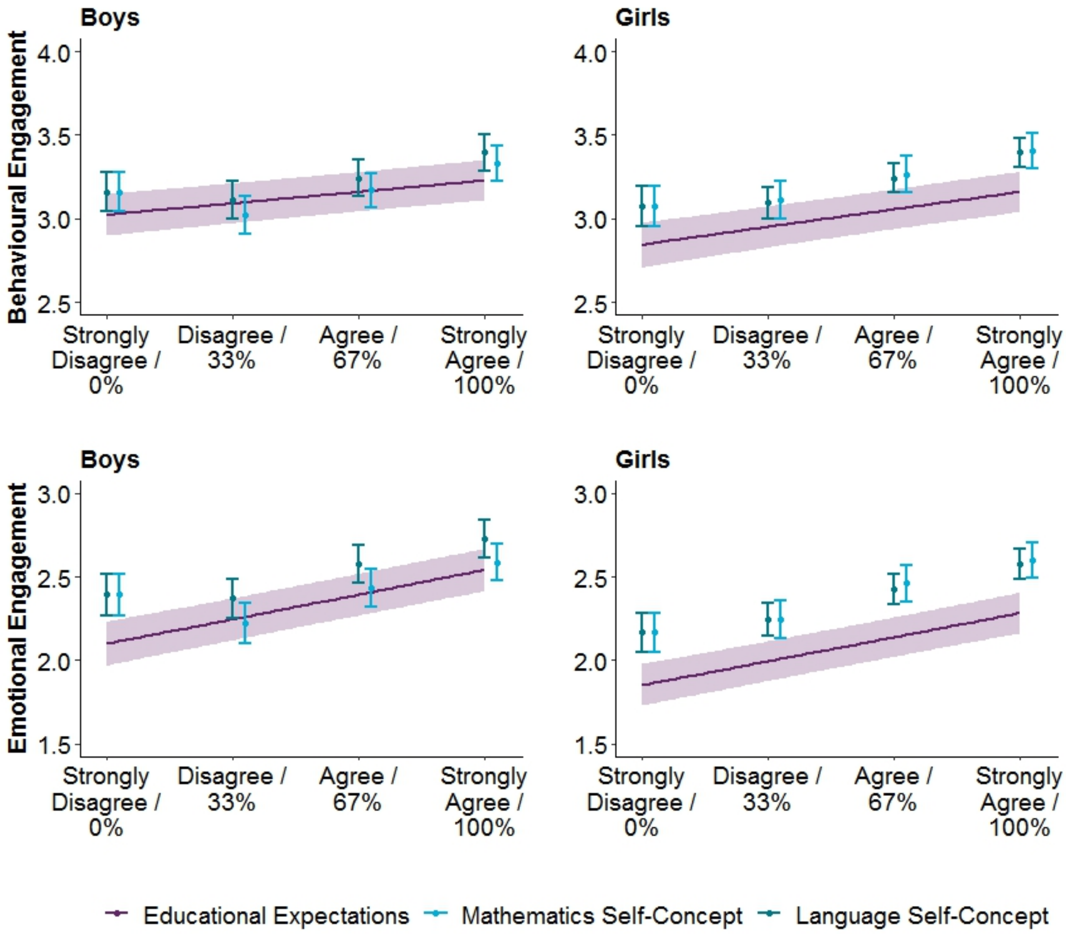
In sum, the positive association between mathematics and language self-concepts and emotional engagement were stronger for girls than for boys, evidenced by stronger median estimates from the models and post hoc comparisons indicating that both self-concepts were more strongly associated with emotional engagement than educational expectations. However, the stronger associations meant that while girls with the highest self-concepts reported similar levels of emotional and behavioural engagement as boys with the highest self-concepts, girls with the lowest self-concepts reported lower emotional engagement than boys with the lowest self-concepts. It was not the case that boys with the lowest self-concepts reported the lowest emotional or behavioural engagement with school, but for girls only, having the lowest self-concepts meant significantly lower reported emotional and behavioural engagement than girls with the highest self-concepts, and most boys in the sample had higher emotional engagement with school. The negative associations between self-concepts and educational expectations with school disengagement were similar for boys and girls.

4.4.4 Covariates Predicting School Engagement

The covariate inconsistently predicted measures of school engagement, so H5 was partially supported. Boys with identified SEN reported higher decreases in emotional disengagement compared to typically developing (TD) boys (Model 2, *Mdn*: 0.15[0.07, 0.23]), in partial support of H5 for SEN status. SEN status appeared to also be related to emotional engagement, but this was only evident in Model 1 (*Mdn*: -0.14[-0.24, -0.05]), suggesting that low educational expectations and/or self-concepts may explain this association. Otherwise, there was no association between SEN status and behavioural engagement or disengagement for boys and SEN status was not related to girls’ school engagement.

Figure 4.3

Conditional Mean Effects of Self-Concepts and Educational Expectations on the Behavioural and Emotional Engagement of Boys and Girls.

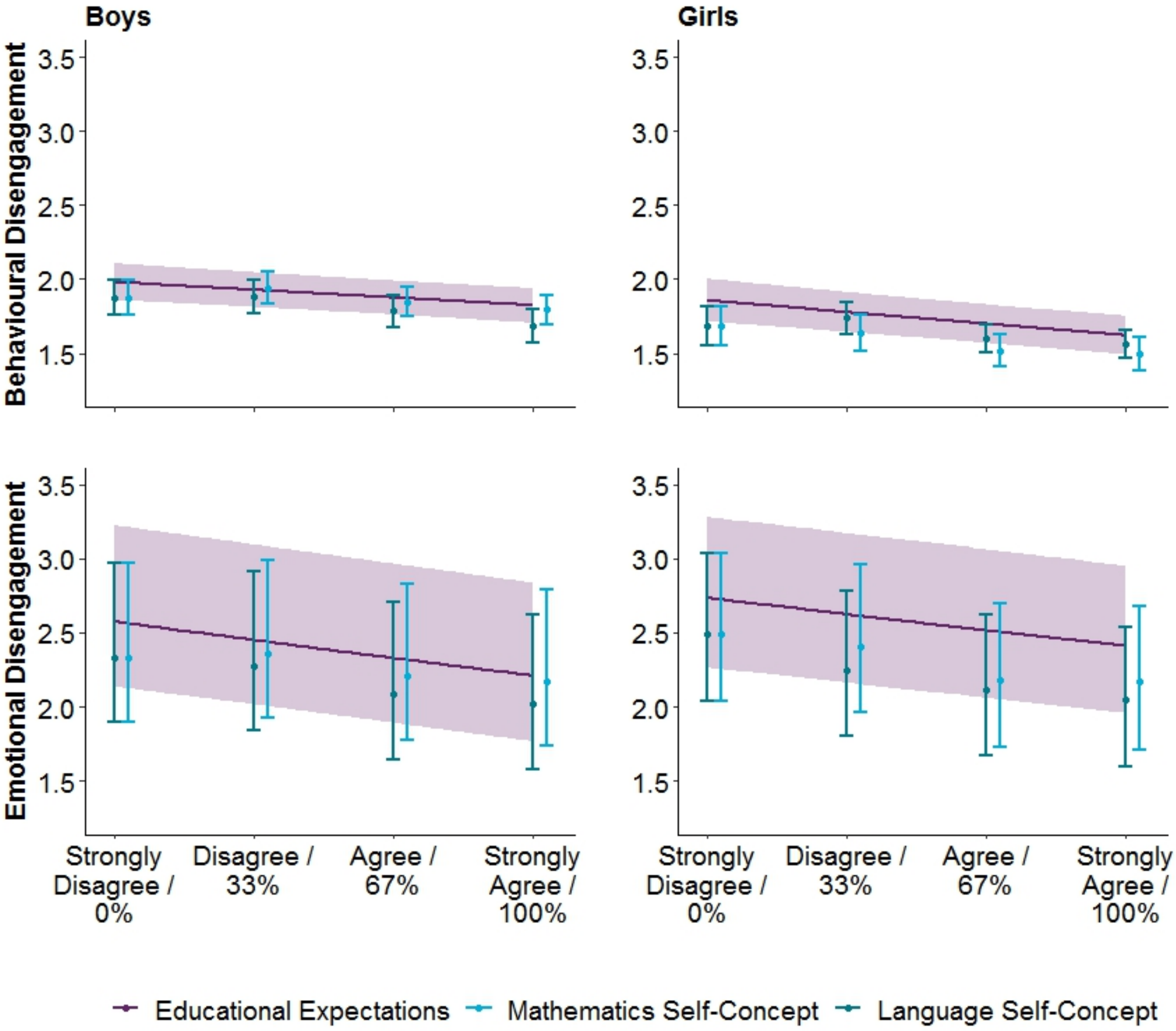


Note. Conditional effects extracted from the Model 2 regression models. Presents the average response and standard error bars for each reported level of mathematics and language self-concept, while other predictors are held at mean values. For educational expectations, the plot presents an average continuous response and standard error ribbon for the percentage likelihood of going to university in the future. Educational expectations and self-concepts are represented together on the x-axis.

Higher family income was associated with higher behavioural and emotional engagement and lower behavioural disengagement for girls and boys. There was no association with emotional disengagement. The association between girls' and boys' behavioural engagement and disengagement were similar. For example, the median associations in Model 2 for behavioural disengagement were (*Mdn*: -0.24[-0.35, -0.14]) for boys and (*Mdn*: -0.30[-0.40, -0.19]) for girls. Parental educational level was removed from the imputed data models but is

Figure 4.4

Conditional Mean Effects of Self-Concepts and Educational Expectations on the Behavioural and Emotional Disengagement of Boys and Girls.



Note. Conditional effects extracted from the Model 2 regression models. Presents the average response and standard error bars for each reported level of mathematics and language self-concept, while other predictors are held at mean values. For educational expectations, the plot presents an average continuous response and standard error ribbon for the percentage likelihood of going to university in the future. Educational expectations and self-concepts are represented together on the x-axis.

reported in full in the results for emotional disengagement (Appendix C, Table C.4, and in the Appendix D, Tables D.6 - D.8) In general, across the results from the missing datasets, there was little evidence of an association with parental educational qualifications. H5 was therefore partially supported for socioeconomic status, but only for family income.

Associations between prior achievement and school engagement were inconsistent, and either not significant or reversed in Model 2 when compared with Model 1. H5 was, therefore, unsupported for prior achievement. There was a protective association between prior achievement and school engagement observed in Model 1 for behavioural disengagement (boys, *Mdn*: -0.12[-0.20, -0.04]; girls, *Mdn*: -0.22[-0.30, -0.14]), emotional disengagement (boys, *Mdn*: -0.12[-0.18, -0.06]; girls, *Mdn*: -0.16[-0.22, -0.09]) and girls only for emotional engagement (0.19[0.12, 0.26]), but no association with behavioural engagement. In Model 2, which included self-concepts and educational expectations, there was a negative association with prior achievement that was the same for boys' and girls' behavioural engagement (*Mdn*: -0.19[-0.27, -0.12]) and differed for emotional engagement (boys, (*Mdn*: -0.24[-0.31, -0.17]); girls, *Mdn*: -0.11[-0.18, -0.04]). Prior achievement and emotional and behavioural disengagement were not related in Model 2. This may indicate an inconsistency or change in self-concepts or educational expectations occurs either resulting from the Age 11 achievement tests or subsequent to these tests, perhaps as a result of the transition to secondary school (Coelho & Romão, 2017). Also, educational expectations, once set, can be persistent, especially among high-SES students and are less influenced by prior achievement (Bernardi & Valdés, 2021).

In sum, H5 was partially supported for the deleterious association between SEN status and emotional disengagement for boys only, and a protective association between family income and behavioural engagement and disengagement, and emotional engagement for girls. Prior achievement was an inconsistent positive predictor of school engagement, and the association was negative once changes in self-concepts and educational expectations were accounted for in the Bayesian models. Parental educational qualifications were not related to school engagement.

4.4.5 Correlational Analyses

The longitudinal correlational analyses between Age 14 measures and mathematics and language achievement at age 16 (see Figure 4.5) revealed moderate correlations between self-concepts and educational expectations with grades for boys and girls (Gignac & Szodorai, 2016). Correlations were stronger for concordant domain-specific self-concepts and grades than for self-concepts from the discordant domain. For example, for boys, the correlation between mathematics self-concept and language grade (r = 0.21) was weaker than the correlation between mathematics self-concept and mathematics grade (r = 0.41). Educational expectations were similarly related to mathematics and language grades. For example, for boys, the correlation between educational expectations and language grades (r = 0.39) was almost identical to the correlation with mathematics grades (r = 0.40).

Longitudinal correlations between age 14 puberty status and age 16 achievement were negligible. Longitudinal correlations between school engagement measures and grades were generally small or very small, although moderate for behavioural disengagement measures. The correlations between behavioural disengagement measures and grades were similar between subjects and between girls and boys (boys, mathematics: r = -0.22, language: r = -0.19; girls, mathematics: r = -0.24, language: r = -0.21). Small correlations were found between age 16 language and mathematics grades and emotional engagement (boys, language: r = 0.12; mathematics: r = 0.13; girls, language: r = 0.16; mathematics: r = 0.17) and feeling that school is a waste of time (boys, language: r = -0.14; mathematics: r = 0.12; girls, language: r = -0.16; mathematics: r = - 0.17). There was also a small correlation between girls' grades and behavioural engagement (language: r = 0.10; mathematics: r = 0.12). All other correlations with school engagement measures were negligible.

Correlations between age and puberty score with self-concepts, educational expectations, and grades were also examined. For boys, only language self-concept correlated with age (r =

0.04). Boys' puberty scores negatively correlated with mathematics self-concept (*r* = -0.04), and positively correlated with language grades (*r* = 0.04). For girls, there were no correlations of age with self-concepts or grade. Girls' puberty score positively correlated with language self-concept (*r* = 0.04), mathematics grade (*r* = 0.08), and language grade (*r* = 0.10). Self-reported educational expectations did not significantly correlate with puberty or age for boys or girls.

For completeness, correlations between Age 11 measures and concurrent Age 11 school grades are reported in Appendix C, Figure C.1. For girls, both age and puberty were positively correlated with age 11 school grades but not self-concept (language grade, age: *r* = 0.08; puberty: *r* = 0.06; mathematics grade, age: *r* = 0.09; mathematics: *r* = 0.05). For boys, age 11 school grades and language self-concepts were positively correlated with age and age 11 grades were negatively correlated with puberty scores while language self-concept was positive correlated (language grade, age: *r* = 0.06; puberty: *r* = -0.04; language self-concept, puberty: *r* = 0.04; mathematics grade, age: *r* = 0.05; puberty: *r* = -0.10). Parent-reported educational expectations were positively correlated with their child's age for boys (*r* = 0.08) and girls (*r* = 0.05).

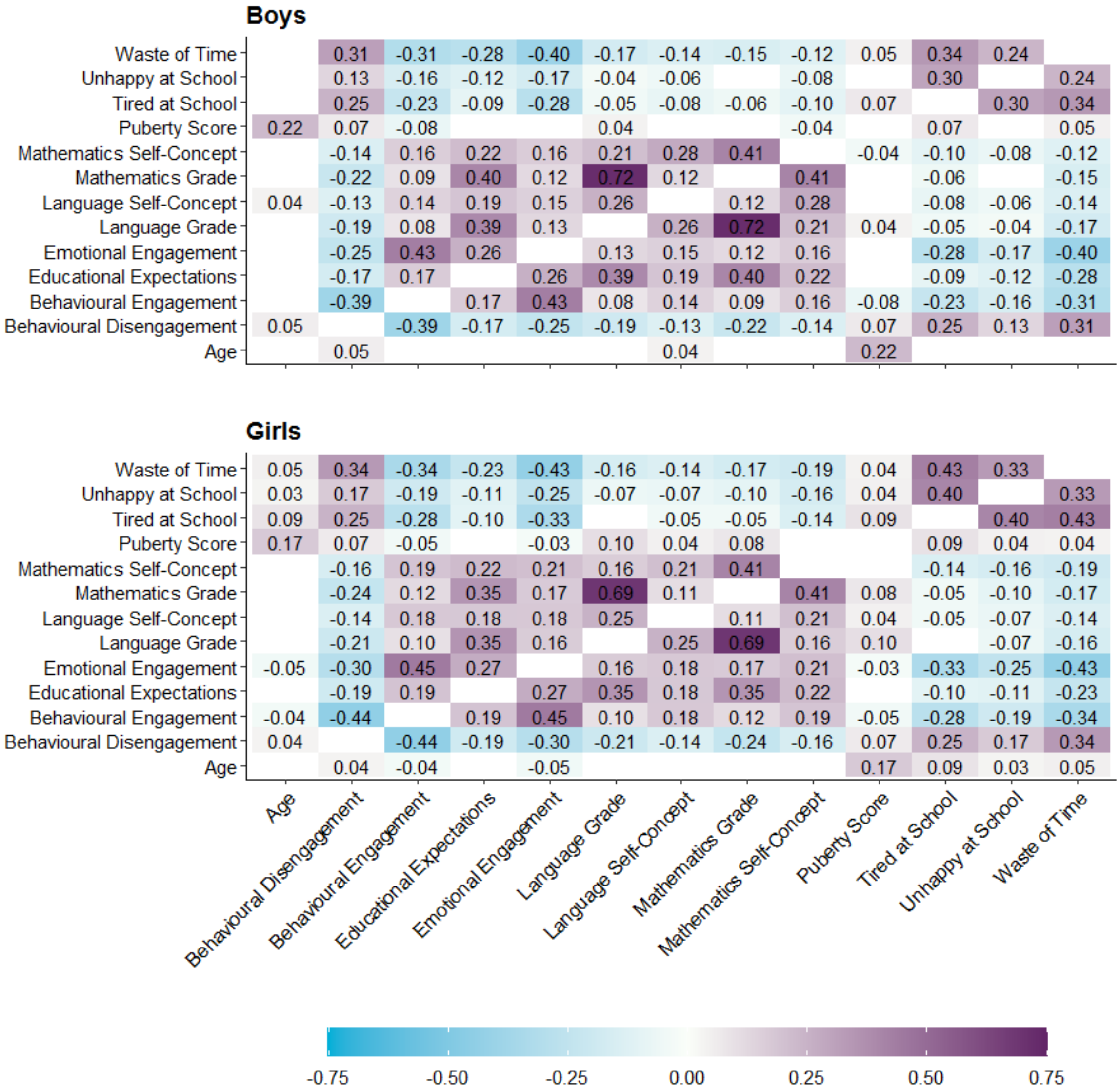
4.4.6 Primary to Secondary School Transition

Comparisons of year 6 and year 7 students find that cohort members who were interviewed after the transition to secondary education reported higher school engagement than those in primary school (see Appendix C, Table C.6). In general, in the Age 11 wave, year 7 secondary-school boys reported lower behavioural disengagement than the year 6 primary-school boys (Year 6: *M* = 1.66, *SD* = 0.63; Year 7: *M* = 1.48, *SD* = 0.60, *d* = -0.29 [-0.54, -0.03]). The boys who had transitioned to secondary school also reported significantly lower mathematics self-concept than those who had not transitioned, but this association was not significant (Year 6: *M* = 3.44; Year 7: *M* = 3.32, *SD* = 0.71, *SD* = 0.68, *d* = -0.17 [-0.43, 0.08]).

Year 7 girls reported higher emotional engagement (Year 6: *M* = 2.93, *SD* = 0.66; Year 7: *M* = 3.12, *SD* = 0.70, *d* = 0.29 [0.06, 0.52]) and were less unhappy at school than year 6 girls (Year 6: *M* = 1.77, *SD* = 0.63; Year 7: *M* = 1.52, *SD* = 0.60, *d* = -0.39 [-0.63, -0.16]). Otherwise, reported behavioural engagement and disengagement, mathematics and language self-concepts, and feeling tired at school and that school is a waste of time, were similar for year 6 and year 7 girls in the Age 11 wave.

Figure 4.5

Bayesian Correlations for Boys and Girls of Age 14 Measures and Age 16 Achievement



Boys

	Age	Behavioural Disengagement	Behavioural Engagement	Educational Expectations	Emotional Engagement	Language Grade	Language Self-Concept	Mathematics Grade	Mathematics Self-Concept	Puberty Score	Tired at School	Unhappy at School	Waste of Time
Waste of Time		0.31	-0.31	-0.28	-0.40	-0.17	-0.14	-0.15	-0.12	0.05	0.34	0.24	
Unhappy at School		0.13	-0.16	-0.12	-0.17	-0.04	-0.06		-0.08		0.30		0.24
Tired at School		0.25	-0.23	-0.09	-0.28	-0.05	-0.08	-0.06	-0.10	0.07		0.30	0.34
Puberty Score	0.22	0.07	-0.08			0.04			-0.04		0.07		0.05
Mathematics Self-Concept		-0.14	0.16	0.22	0.16	0.21	0.28	0.41		-0.04	-0.10	-0.08	-0.12
Mathematics Grade		-0.22	0.09	0.40	0.12	0.72	0.12		0.41		-0.06		-0.15
Language Self-Concept	0.04	-0.13	0.14	0.19	0.15	0.26		0.12	0.28		-0.08	-0.06	-0.14
Language Grade		-0.19	0.08	0.39	0.13		0.26	0.72	0.21	0.04	-0.05	-0.04	-0.17
Emotional Engagement		-0.25	0.43	0.26		0.13	0.15	0.12	0.16		-0.28	-0.17	-0.40
Educational Expectations		-0.17	0.17		0.26	0.39	0.19	0.40	0.22		-0.09	-0.12	-0.28
Behavioural Engagement		-0.39		0.17	0.43	0.08	0.14	0.09	0.16	-0.08	-0.23	-0.16	-0.31
Behavioural Disengagement	0.05		-0.39	-0.17	-0.25	-0.19	-0.13	-0.22	-0.14	0.07	0.25	0.13	0.31
Age		0.05					0.04			0.22			

Girls

	Age	Behavioural Disengagement	Behavioural Engagement	Educational Expectations	Emotional Engagement	Language Grade	Language Self-Concept	Mathematics Grade	Mathematics Self-Concept	Puberty Score	Tired at School	Unhappy at School	Waste of Time
Waste of Time	0.05	0.34	-0.34	-0.23	-0.43	-0.16	-0.14	-0.17	-0.19	0.04	0.43	0.33	
Unhappy at School	0.03	0.17	-0.19	-0.11	-0.25	-0.07	-0.07	-0.10	-0.16	0.04	0.40		0.33
Tired at School	0.09	0.25	-0.28	-0.10	-0.33		-0.05	-0.05	-0.14	0.09		0.40	0.43
Puberty Score	0.17	0.07	-0.05		-0.03	0.10	0.04	0.08			0.09	0.04	0.04
Mathematics Self-Concept		-0.16	0.19	0.22	0.21	0.16	0.21	0.41			-0.14	-0.16	-0.19
Mathematics Grade		-0.24	0.12	0.35	0.17	0.69	0.11		0.41	0.08	-0.05	-0.10	-0.17
Language Self-Concept		-0.14	0.18	0.18	0.18	0.25		0.11	0.21	0.04	-0.05	-0.07	-0.14
Language Grade		-0.21	0.10	0.35	0.16		0.25	0.69	0.16	0.10		-0.07	-0.16
Emotional Engagement	-0.05	-0.30	0.45	0.27		0.16	0.18	0.17	0.21	-0.03	-0.33	-0.25	-0.43
Educational Expectations		-0.19	0.19		0.27	0.35	0.18	0.35	0.22		-0.10	-0.11	-0.23
Behavioural Engagement	-0.04	-0.44		0.19	0.45	0.10	0.18	0.12	0.19	-0.05	-0.28	-0.19	-0.34
Behavioural Disengagement	0.04		-0.44	-0.19	-0.30	-0.21	-0.14	-0.24	-0.16	0.07	0.25	0.17	0.34
Age		0.04	-0.04		-0.05					0.17	0.09	0.03	0.05

Note. Correlations that are not practically significant (*pd* < 0.95) are excluded.

4.5 Discussion

In partial support of hypothesis H1, perceived puberty predicted the school disengagement of girls and the school engagement of boys. For boys, advanced puberty was associated with lower behavioural engagement. For girls, advanced puberty was associated with increased emotional and behavioural disengagement. Perceived female puberty was not related to girls' behavioural and emotional engagement. Perceived male puberty was not related to the emotional engagement or emotional and behavioural disengagement of boys. Hypothesis H2 was supported for girls only due to the association between puberty and emotional and behavioural disengagement. The association between perceived puberty and emotional disengagement increased in older girls. The effect of puberty on male behavioural engagement was also dependent on the age group being examined. Advanced perceived puberty negatively predicted male behavioural engagement at ages 14-15 years old. These results also suggested that being more mature may tend towards being protective at earlier ages. As H2 is only supported for girls, H3 was unsupported, as there was no association between boys' perceived puberty and school disengagement. The results fully supported hypothesis H4. Higher mathematics and language self-concepts were associated with higher behavioural and emotional engagement and lower behavioural and emotional disengagement. For girls, the strength of the association between mathematics and language self-concepts and emotional engagement would be considered of large magnitude (Gignac & Szodorai, 2016; Schäfer & Schwarz, 2019). The association between mathematics self-concept and girls' emotional engagement was stronger than for boys, and language self-concept was more strongly associated with boys' emotional engagement with school than mathematics self-concept. Finally, H5 was partially supported. SEN status increased emotional disengagement for boys only. Parental educational qualifications were not associated with changes in school engagement, and the positive associations between prior achievement and school engagement were reversed when self-concepts and educational expectations were included in the statistical models. Cohort members who came from higher-income families reported less decline in school engagement compared to lower-income cohort members.

This study supports recent reports of an association between puberty and school disengagement; although this association was found for girls only, despite hormone levels predicting puberty status (Martin et al., 2017, 2022). Contrary to Martin et al. (2022), we find that perceived male puberty is more strongly associated with declines in behavioural engagement, especially in middle adolescence. Different relationships between biological measures and self-reported, perceived measures of maturation have been reported previously (Martin et al., 2017; Torvik et al., 2021). For example, Martin et al. (2017) found that perceived puberty predicted achievement motivation, but puberty hormones did not. Additionally, Torvik et al. (2021) reported that an earlier growth spurt was associated with higher academic achievement, whereas an earlier onset of menstruation was linked to lower academic achievement. Both of these physical maturation events are associated with rising hormone levels (Mendle & Koch, 2019). The underlying mechanisms that explain why different results are elicited depending on the measures of maturation are unclear and require further evidence. Advanced pubertal maturation has been linked to improvements in IQ, cognitive skills growth, and executive function gains, which indicates that biological cognitive change may underlie some findings (Chaku & Hoyt, 2019; Koerselman & Pekkarinen, 2018; Newcombe, 1989). The maturation measures may also represent a social reaction to biological change. The growth spurt is a change that is externally salient and has been linked to wellbeing (Rees et al., 2009). For example, there are additional psychological benefits from being taller as an adolescent for boys, although whether these benefits contribute to achievement has been questioned (Rees et al., 2009). Self-reported puberty measures may also encompass social comparison processes, as individuals respond based on their perceptions of their development in comparison to their peers. Therefore, self-reported perceived measures capture biological, cognitive, and social change (Mendle & Koch, 2019). Individuals tend to socially compare themselves upwards, which can result in negative affect and lower self-appraisals (Dijkstra et al., 2008; Jansen et al., 2022).

The lower educational achievement of more mature, or earlier maturing girls, is commonly reported (Cavanagh et al., 2007; Dubas et al., 1991; Goering et al., 2023; Koivusilta & Rimpelä, 2004; Mendle et al., 2007; Simmons & Blyth, 1987; Suutela et al., 2022). Although other research does not support this conclusion (Graber et al., 2004; Graber et al., 1997; Senia et al., 2018; Stattin & Magnusson, 1990). In line with the majority, we found that more advanced female puberty was associated with increased girls' emotional and behavioural disengagement, and these measures are negatively correlated with mathematics and language grades at age 16. The current literature on the association between male puberty and educational achievement remains inconclusive (Goering et al., 2023; Mendle & Ferrero, 2012; Senia et al., 2018). This study finds that advanced male perceived puberty was associated with declines in behavioural engagement, with behavioural engagement making a very small positive contribution to mathematics grades and language. Together, these associations would predict very small negative associations with later achievement for boys, which we find in correlational analyses for mathematics, but not for language grades, which were positively associated with puberty scores. These changes would provide some support for findings that more mature boys are at risk of lower academic achievement, but these very small associations do not provide enough evidence to substantively link puberty to the lower achievement of boys during early to middle adolescence through its association with school engagement (Dubas et al., 1991; Goering et al., 2023; Torvik et al., 2021). Neither would the negative association between advanced puberty and school disengagement explain the general female advantage during secondary education (Oakley et al., 2024; Voyer & Voyer, 2014). Indeed, while at age 14-15, girls significantly advanced through their maturation period compared to boys, the negligible association with age 16 does not suggest that this sex difference substantially influences outcomes.

Different maturation measures elicit different associations with achievement and motivation and therefore account for some differences between studies (Martin et al., 2017; Torvik et al., 2021). Where direct associations between puberty and achievement are examined, the achievement measures used may also contribute. For example, some studies that reported no association with puberty used self-reported grades, which tend to be overestimated, more often by boys, and are less reliable in less-able students (Crockett et al., 1987; Dubas et al., 1991; Graber et al., 1997; Kuncel et al., 2005; Senia et al., 2018). Some studies have examined combined measures of achievement, such as grade point average or averages across all UK GCSE subjects, and others have focused on individual subjects and found different associations between school subjects and between the sexes (Simmons & Blyth, 1987; Suutela et al., 2022; Torvik et al., 2021). This study found consistent subject-specific correlations with school engagement measures. However, the correlational analyses revealed some sex-specific associations between age and perceived puberty with age 14 self-concepts and age 16 grades. For boys, being older was correlated with higher language self-concept, whereas the association was with perceived maturation, not age, for girls. However, being more mature at age 14 was positively associated with later language grades for girls and boys, and mathematics grades for girls only. These correlational results partially align with Suutela et al. (2022), who reported that earlier age of peak height velocity in girls was positively associated with mathematics grades at age 14/15 years old, but did not find an association with language grades for either sex. Like Torvik et al. (2021), there was a very small correlation between perceived puberty and school grades, although Torvik et al. (2021) examined concurrent teacher-assessed grades rather than longitudinal correlations, and they used a combined measure of achievement. To my knowledge, puberty correlations with domain-specific self-concepts have not been reported previously.

The correlations of emotional engagement (small) and emotional disengagement (negligible) with Age 16 mathematics and language grades mostly align with those reported by

Wang et al. (2019), except that feeling school is a waste of time had a stronger negative association with school grades. The correlation between behavioural engagement and disengagement with Age 16 achievement is stronger than that previously reported. The correlations for behavioural engagement and disengagement reported here at age 11 with age 11 prior achievement are smaller/negligible and align with those reported by (Wang et al., 2019). The stronger correlations with achievement in the older age group may indicate that the association between behavioural engagement and disengagement with school achievement changes with age, and support reports of a lesser role of the emotional dimension of school engagement in secondary education (Chang et al., 2016; Lei et al., 2018; Wang et al., 2017). Interestingly, the correlations suggest that only the role of emotional disengagement here is consistent across both age groups.

The brief supplementary exploration of pre- and post-school-transition cohort members revealed some positive and negative associations. Boys who had transitioned to secondary school were more behaviourally engaged than those boys in the last year of primary education, and the difference was of moderate magnitude. However, the secondary school boys also reported lower mathematics self-concepts of small magnitude. This finding was not significant, but may be influenced by the comparatively very small sample of post-transition boys. The results were mixed for post-transition girls, who were more emotionally engaged and were also less unhappy at school than pre-transition girls. The positive changes in school engagement findings align with reports that transition effects are not always consistently negative and that there are some perceived benefits to transition (Jindal-Snape et al., 2020; Symonds & Galton, 2014). It is plausible that some of the older pupils at the end of primary school may be more developmentally ready than others for transition, and primary school may no longer meet their developmental needs, for example. The high-stakes exams at the end of UK primary education may also influence school engagement in year 6. Symonds and Galton (2014) highlighted that students might appreciate new subjects, lessons, better equipment, and a wider range of friendships associated with secondary school, but that teacher-relatedness can decrease. The negative association between boys' mathematics self-concept and the transition to secondary school may result from the end of primary school tests, although boys' comparative performance in mathematics was, on average, good (Oakley et al., 2024). Instead, this change may reflect a change specifically associated with the transition, possibly as a result of social comparison, for example, the big-fish-little-pond effect (Marsh & Hau, 2003). Secondary schools are generally much larger than primary schools, with multiple primary schools feeding into a single secondary school. Mathematics self-concepts may therefore be adjusted when faced with a much larger comparison group, or alternatively, the lower self-concepts may result from the academic streaming of pupils, which has been shown to be negatively linked in (at least) the short-term (e.g., Liu et al., 2005).

4.5.1 Self-Concepts and Educational Expectations

We found that self-concepts and educational expectations opposed declines in school engagement, in line with H4, and these associations were of stronger magnitude than perceived puberty associations, similar to the prior findings for achievement motivation (valuing and self-efficacy: Martin et al., 2022). Mathematics and language self-concepts were each positively associated with engagement and negatively associated with disengagement to a lesser degree. Educational expectations were also protective against declines in engagement and increases in disengagement. The strength of the associations between self-concepts and emotional engagement and disengagement sometimes differed between boys and girls. The associations between educational expectations and school engagement were gender-invariant.

The associations between self-concepts and changes in school engagement were often curvilinear. Mathematics self-concept was more protective of girls' emotional engagement and behavioural engagement than boys'. Boys with the lowest self-concepts did not report lower school engagement than boys with higher self-concepts. This finding that low self-concepts were not necessarily detrimental for school engagement may align with reports of identity-based decisionmaking during adolescence (Berkman et al., 2017; Pfeifer & Berkman, 2018). Under this proposal, decision-making is identity-related and, therefore, influenced by group affiliations and positive subjective value. For example, effort and engagement in school are often associated with a feminine gender identity. Boys may gain socially from being seen to be less engaged and put less effort into school; having low self-concepts may therefore not be detrimental to self-identity (Berkman et al., 2017; Kessels et al., 2014; Yu et al., 2020). Nevertheless, Martin et al. (2017) proposed that achievement motivation was an enabling or protective factor through which school engagement could be improved. This study aligns with this conclusion, finding similar associations for educational expectations and domain-specific self-concepts. These results also align with a prior study demonstrating that academic self-concepts are predictive of school engagement and support the role of self-concepts as important components of the self in an educational context, as described by the self-systems model of motivational development (Green et al., 2012; Skinner, 2023; Skinner et al., 2008, 2009).

4.5.2 Family Income and Special Educational Needs

Higher family income was associated with higher behavioural and emotional engagement and lower behavioural disengagement, and so opposed the negative age and puberty associations. The association of family income with behavioural engagement was the strongest of these. School engagement has been previously identified as a mediator of the association between socioeconomic status and school achievement (Tomaszewski et al., 2020). Tomaszewski et al. (2020) found lower school engagement in lower-SES students, particularly behavioural engagement, and highlighted that school engagement profiles differed in students from different backgrounds. Socioeconomic associations with school engagement are less well studied and are an area for future research (Reschly & Christenson, 2012; Tomaszewski et al., 2020). This study provides further evidence that family income predicts school engagement, resulting in higher behavioural disengagement and lower emotional and behavioural engagement for students from low-income backgrounds.

SEN status was associated with increases in emotional disengagement. SEN status was associated with declines in boys' emotional engagement, although this effect was no longer significant in the model accounting for changes in self-concepts and educational expectations. The inconsistency across both models for emotional engagement may indicate that the SEN association with male emotional engagement is linked to lower self-concepts and educational expectations, which have been reported elsewhere (Frawley et al., 2014). However, the association with emotional disengagement was persistent, and the descriptive statistics indicated that while many of the differences between SEN and TD students decreased between the surveys, feeling tired at school was the only difference that increased (see Appendix C, Table C.5). The lower achievement of students with SEN in the UK education system and the higher prevalence of SEN diagnoses among boys are well-documented (Daniel & Wang, 2023; Department for Education, 2023). However, few studies have examined SEN status and school engagement, and to date, the results are mixed (O'Donnell & Reschly, 2020). A recent PhD thesis reported no sex or SEN differences in affective engagement towards the end of primary education, but that students with SEN were more likely to report feeling unsafe at school, that school rules were unfair, and that school was a waste of time (Veater, 2022). Given that we report emotional engagement and disengagement measures separately, these results align but in an older sample, indicating that at least some of these school emotions likely persist into secondary education. Similarly, Frawley et al. (2014) reported lower affective engagement for students from low-income circumstances and students with SEN, and this is related to lower self-appraisals and self-concepts (see also Murray & Greenberg, 2001, 2006). School engagement profiles may also differ between individual diagnoses or symptom severity (Moreira et al., 2015). Other evidence reports no association, particularly in the context of inclusive schools where SEN pupils are attended to more than typically developing students, indicating that these group differences can be reduced or gaps closed (Svetaz et al., 2000; Szumski et al., 2017; Wallace et al., 2002). How school engagement differs for SEN students is an important area for future research given the continuing under-performance of SEN students in education and the reported protective association of school connectedness against the increased risk of emotional distress in adolescents with SEN (Svetaz et al., 2000; Veater, 2022). However, given that male and SEN cohort members both exhibited similar lower school engagement profiles at the end of primary school, as well as distinguishing between types of diagnosis, it may also be important to differentiate between girls and boys with SEN.

4.5.3 Implications

In line with Martin et al. (2022), these results indicate a role for adolescent development in the observed changes in school engagement reported during early to middle adolescence. As this study uses a perceived measure of puberty, this likely points to the role of social processes related to maturity rather than the biological processes that underlie at least some of the changes in school engagement during adolescence, for example, as suggested by the identity-value model (Berkman et al., 2017; Mendle & Koch, 2019). The use of a self-reported puberty measure may therefore indicate that it is social processes related to maturity, rather than the underlying, linked biological processes, that underlie girls' increasing disengagement from school during early to middle adolescence. While these associations differ from the biological associations reported with disengagement, emotional and behavioural engagement and disengagement are not distinct, unrelated constructs (Martin et al., 2022). There is evidence of dynamic associations between each measure, such that declines in one construct will influence changes in the other three constructs, in a dynamic, reciprocal relationship over time (Green et al., 2012; Skinner et al., 2008, 2009).

In general, as reported elsewhere, while puberty predicted school engagement, the associations were small and likely only evident when comparing the least and most mature students within a school year group (Koerselman & Pekkarinen, 2018; Smith-Woolley et al., 2017). Teaching professionals should expect and account for maturity-related changes to occur during this period as a normative part of adolescent development. In particular, steeper declines in girls' school engagement during this early-middle adolescent period should be expected, and steps taken to minimise these declines. While the declines reported in school engagement, self-concepts, and self-esteem may be steeper for girls, advanced male puberty is also associated with declines in school engagement from their lower level on entry to secondary education (Martin et al., 2022; Schaffhuser et al., 2017). A focus on those students with lower self-appraisal may be appropriate, as high self-concepts and educational expectations support school engagement.

School engagement is susceptible to change and, therefore, is a promising avenue for practical interventions to improve outcomes for a range of students at risk of underachievement (e.g., Wallace et al., 2002; Wang & Degol, 2014). Targeting interventions at self-concepts may be one approach to achieve this through raising existing self-concepts and minimising age or maturity-related declines (O'Mara-Eves et al., 2006). Raising educational expectations directly is another approach, but is more complex due to the influence of social norms, financial barriers, and other influencing factors (Fumagalli, 2019; Kremer et al., 2018; Wilson et al., 2018). Raising self-appraisals may also have benefits for educational expectations, and vice versa (Buchmann et al., 2022; Pinquart & Ebeling, 2020). For boys and SEN students, a focus on primary education may be necessary to close engagement gaps prior to entering secondary education.

4.6 Limitations

While interpreting these results, there are some important limitations to address that provide avenues for future research. The use of single-item measures for emotional and behavioural engagement and behavioural disengagement may limit this study, as well as how well the measures represent each school engagement concept. As CFA indicated that these measures could not be reliably combined into either engagement and disengagement or otherwise behavioural and emotional dimensions, the use of single-item measures was unavoidable - although single-items are often an acceptable alternative (e.g., Castro et al., 2023). Notably, the correlations between each emotional disengagement measure indicate that some measures are more strongly associated with puberty (school is a waste of time) than others (unhappy at school and tired at school). As we have found sex-specific associations between puberty and engagement and disengagement within the dimensions of school engagement, this finding warrants further research to understand whether the sex-specific associations replicate in a sample with additional measures for each dimension and whether the associations are item-specific. For example, sex-specific associations between age and facets of personality dimensions during adolescence have been reported (De Bolle et al., 2015).

Puberty was measured in two MCS survey waves, and there was, on average, a three-year gap between measurements. Correspondingly, there are substantial changes to the distribution of pubertal status of boys and girls between the two occasions. In the Age 11 wave, there was more variability in girls' pubertal status than boys' pubertal status, which is unsurprising given the earlier start for female puberty. By the second Age 14 wave, boys' pubertal status was more variable and girls' less so. It is plausible, therefore, that girls by middle adolescence have reached the lowest point of the decline in self-concepts and school engagement and that each of these improves after this point. For boys, whose puberty lags behind girls, this lowest point may come later and closer to the age 16 outcomes examined here. Curvilinear changes have been reported previously for some measures, but other than age-related changes in the five-factor model of personality, competence beliefs, and self-esteem, separate trajectories for girls and boys are not often reported (De Bolle et al., 2015; Jacobs et al., 2002; Soto et al., 2011). For example, Soto et al. (2011) reported negative trends for girls' and boys' conscientiousness and self-discipline in early to middle adolescence when trajectories converged. After middle adolescence, the positive trend for girls was more pronounced, with girls reaching an adult profile earlier. However, this would likely benefit girls in older age groups, and after the educational outcomes that are measured here. De Bolle et al. (2015) reported the convergence of some male and female personality traits later in adolescence / towards adulthood, but sex differences in some personality facets changed with age. Similarly, Huang et al. (2022) reported age 14 as a turning point in self-esteem trajectories. As the MCS did not measure puberty after the Age 14 survey wave, this hypothesis cannot be tested. Prior analyses using puberty measured at age 16 have reported less variation in girls' puberty than boys' and less influence on girls' later outcomes, and a stronger association between age 16 pubertal status with achievement at age 16 (Koerselman & Pekkarinen, 2018; Torvik et al., 2021). The change between using a parent-reported and self-reported perceived puberty will also influence results (Dorn et al., 2003). Future research should aim to measure puberty more often and later into adolescence while maintaining consistency of raters across waves when perceived puberty measures are used, particularly as self-reported puberty likely encompasses social processes (Mendle & Koch, 2019).

Like the puberty measures, educational expectations were parent-reported at Age 11 and self-reported at Age 14. While parents' educational expectations of their children influence their child's educational expectations, there can be discrepancies, and children can tend to report lower educational expectations than their parents (Pinquart & Ebeling, 2020; Zhang et al., 2011). Any change or difference between the Age 11 and Age 14 surveys for individuals could represent a difference in opinion by parents and children or could represent a real change in expectations of the child based on their socially influenced, maturing self-appraisals, or a combination of these. For both sexes, cohort members' educational expectations were lower than parents' educational expectations, who were influenced by their child's age (see Table 4.6 for details). The difference between surveys was larger for boys than for girls. Future research should aim to clarify whether boys' educational expectations decline over adolescence or whether this represents a larger difference between parent and child expectations for boys.

4.7 Conclusion

Early to middle adolescence is an important period for school achievement. Outcomes during this period influence later subject choice and achievement in secondary school and the decision to continue in full-time education after compulsory schooling ends. As had been previously reported, boys ended primary education less engaged and more disengaged than girls across both dimensions of school engagement. The adolescent development period may serve to accentuate the already lower school engagement of some groups, which include boys, students with SEN, and low-income groups. This accentuation results in substantial declines in behavioural engagement and increases in behavioural and emotional disengagement. Advanced puberty increases the risk of lower school engagement for students who do not benefit from the protective factors of high self-concepts, especially language self-concepts for boys, and educational expectations. High self-concepts and educational expectations outweigh very small negative associations with puberty. While puberty, self-concepts and achievement are correlated, the associations were very small and are unlikely to contribute substantially to the widening female educational advantage reported during mid-adolescence.

,0

5 General Discussion

5.1 Overview and Aims of this Thesis

While the empirical chapters in this thesis addressed several research gaps, the predominant aims of this thesis were twofold. Firstly, to clarify how achievement gaps change with age in a single, longitudinal sample, re-examining the reported general female advantage in early to late-middle secondary education (Research Questions (RQs) 1-2). Secondly, this thesis took steps to examine associations between puberty and school engagement (RQs 7-8) and, therefore, the widening female advantage in school grades during this early to middle adolescent period (Cavaglia et al., 2020; Voyer & Voyer, 2014). In addition to these two main themes of this thesis, taking a data-driven approach, this thesis was also able to examine a set of additional research questions due to the size of the thesis sample and the available educational data.

The UK educational system features standardised curriculum tests at ages 11 and 16, and dual teacher-assessed and standardised grades at age 11. The thesis sample, with its linked educational outcomes data, is therefore well suited to an examination of whether the reported changes reflect actual achievement differences between boys and girls or, in part, reflect stereotype-driven biased (RQ4) opinion (e.g., Burgess & Greaves, 2013; Lavy & Sand, 2018; Terrier, 2020). The breadth of subjects tested in the UK system, particularly at age 16, allowed this thesis to examine how sex differences in STEM subjects evolve from primary school to mid-secondary school (RQ3). The large number of sample cohort members enabled this thesis to examine sex differences at the upper and lower achievement levels and, therefore, how greater male variability (GMV) contributes to school achievement (RQ5). This thesis also examined whether the representation of boys and girls at the upper and lower tails of grade distributions aligns with findings from international standardised tests (RQ6). The Millennium Cohort Study, with its linked educational data and a broad range of socioeconomic and attitudinal data, is one of a relatively few datasets in which several of these research questions can be investigated.

Several reports have highlighted that a substantial proportion of students who do well at the end of primary education fail to make expected progress during secondary education. These students are more likely to be male, of lower socioeconomic status, or have an identified special educational need (Montacute, 2018; Sutton et al., 2018). As the focus of Study 1 is on how sex differences in academic achievement change with age, the study contributes to our understanding of several broader issues, including the boy crisis in education, gender stereotypes, and the underrepresentation of women in STEM fields, as well as the underrepresentation of men in non-STEM fields. The boy crisis in education refers to the fact that boys earn lower grades, are more likely to drop out of school, and are less likely to attend university than girls (Cappon, 2011; Carroll, 2024; Cavaglia et al., 2020; Welmond & Gregory, 2021). Prior to the mid-1990s, women were underrepresented in tertiary education in the UK. Other OECD countries achieved enrolment parity earlier, while some achieved parity later; however, by the mid-2000s, most OECD countries had reversed male overrepresentation in tertiary education, and since then, female tertiary participation rates have continued to rise while male participation rates have fallen behind (Cappon, 2011; Welmond & Gregory, 2021). In addition to poorer academic performance, boys are more likely to be suspended or excluded from school, and are more likely to exhibit behavioural problems or learning disabilities (Daniel & Wang, 2023; Department for Education, 2022; Mead, 2006). The boy crisis is reported in many countries both within and outside of the OECD. For example, the relative underachievement of boys in education is reported in China, where the top male scorers in China's college entrance exam (the Gaokao) fell from over 65% to under 40% between 1999 and 2008 (Jing et al., 2021). Commonly, the low proportion of male teachers (especially in primary education), a lack of male role models at home, co-educational settings, the earlier maturation of girls (in early childhood and also pubertal maturation), cultural gendered roles, expectations and stereotypes, distractions that are more common in boys such as video gaming, gender differences in labour market opportunities, and the higher prevalence of SEN diagnoses are widely attributed as causes of, or as contributing to, the boy crisis (APPG, Men & Boys, 2023; Cappon, 2011; Jing et al., 2021; Stoet & Geary, 2018; Stoet & Yang, 2016). The following sections, therefore, discuss the results of Study 1 and Study 2 in the context of the theory and literature, as well as how this thesis contributes to these wider societal issues and discussions.

5.2 Study 1: Sex Differences of School Grades in Childhood and Adolescence

5.2.1 Contribution to Theory and the Literature

Study 1 examined the evolution of sex differences in mathematics, language, and science school grades from age 7 to age 16, by examining how mean sex differences and sex differences at the upper and lower tails of the achievement distribution changed with age. While numerous studies have reported sex differences in language and mathematics school grades, fewer studies have examined sex differences in science grades (O'Dea et al., 2018; Voyer & Voyer, 2014). Of the few studies that have examined sex differences in science achievement, a small number have reported that sex differences can vary between science disciplines, making it important to investigate sex differences in each available science discipline in addition to overall differences in science (Deary et al., 2007; Reilly et al., 2015; Stoet, 2015; Wirthwein et al., 2020). Large-scale analyses of sex differences in international and national assessments have examined sex differences at the mean and the upper and lower tails of achievement distributions in language, mathematics, and science skills (Baye & Monseur, 2016; Gray et al., 2019), and the O'Dea et al. (2018) meta-analysis examined sex differences in grade variability. To the best of my knowledge, this study is the first to directly explore sex differences in school grades at the upper and lower tails of the distribution, and it is also the first to investigate how sex differences evolve from primary to secondary education in curriculum assessments and test grades within a single longitudinal sample. This study also adds to the limited evidence on how sex differences differ between science disciplines.

The results for language achievement aligned well with Voyer and Voyer's (2014) school grades meta-analysis, prior reports from UK samples, and reports from large-scale international standardised assessments at the mean and the tails of the distribution (Baye & Monseur, 2016; Deary et al., 2007; Haworth et al., 2010; Stoet & Geary, 2013). A substantial overrepresentation of boys at the lower tail of the language and reading is highlighted, as also found in large-scale standardised reading and writing assessments at all ages (Baye & Monseur, 2016; Reilly et al., 2019; Stoet & Geary, 2015). Between the ages of 11 and 16, while the number of students observed at the lowest levels of achievement decreased, the proportion of boys to girls observed at these lower achievement levels increased. These results indicate that at the lower achievement levels, more girls than boys had maintained or improved their literacy achievement level between early and middle adolescence. Female advantages also widened at the highest levels of language achievement between ages 11 and 16. While fewer boys and girls were observed at these highest achievement levels at age 16 than at age 11, more boys failed to maintain their prior level of achievement in language.

Sex differences in writing are larger at the lower achievement level than they are in reading at age 11, despite being similar to reading at age 7. Much of the literature focuses on reading differences between boys and girls. However, Study 1 and earlier work by Reilly et al. (2019) suggest that sex differences in writing skills may be larger than reading differences. Interestingly, the sex differences reported at the lower tail in Study 1 are larger than those reported from the US analyses (Reilly et al., 2019). Sex differences in teacher-assessed writing in Study 1 are smaller at the upper achievement levels and are similar at lower achievement levels (Reilly et al., 2019). It is plausible that some differences between the US and these analyses may result from the UK writing assessments being teacher-assessed. We find that teachers make less use of the highest and lowest grades, for example. Still, there is little evidence in Study 1 that teacher-assessed grades significantly alter the proportions of males and females observed at different points in the distribution. Therefore, it does not appear that teacher-assessed grades significantly influence the conclusions drawn; however, they do underestimate within-group variability, a matter discussed further below. At age 11, sex differences in reading and writing were larger at the lower tail at age 11 than at age 7, and significantly larger at the lower tail in writing than in reading. Reilly et al. (2019) reported that sex differences in reading and writing are wider in secondary education and that sex differences in writing are substantial. As there are no separate assessments of writing in the UK education system after age 11, it remains an open question and an area for future research to examine how sex differences in writing skills change during adolescence. The predominant focus in the literature has been on sex differences in reading or, more broadly, language skills. If sex differences in writing widen during adolescence, these would likely contribute to wider sex differences in the age 16 examinations.

While standardised tests and school grades in language and reading skills largely agree, there are important differences between the conclusions drawn when comparing school grades and large-scale standardised assessments in mathematics and science, as well as between different standardised assessments in these subjects (Baye & Monseur, 2016; Stoet & Geary, 2018). Our findings for science, chemistry, biology, and mathematics align with those reported by Voyer and Voyer (2014) in their school grades meta-analysis, confirming that sex differences in mathematics are smaller outside of North America. Both our small female advantage in biology, chemistry, and no sex difference in physics at age 16, and Voyer and Voyer's (2014) results, contrast with Reilly et al. (2015), who reports no difference in life sciences, and a male advantage in physical science at grades 8 and 12 in a US. The disagreement between school grades and international large-scale assessments is most apparent in science. For example, at age 15, more girls were observed at the lower tail in science in PISA assessments, and more boys were observed at the upper tail (Baye & Monseur, 2016). In contrast, Study 1 reports the opposite when examining the tails of the age 16 science, chemistry, and biology school grades distributions. Only in physics was a non-significant tendency for more boys to be observed, and only in the upper 5% of achievers. Similarly, in mathematics, an overrepresentation of boys at the highest levels of mathematics achievement is reported in the PISA and SAT-M tests, which is larger than we found in school grades (Baye & Monseur, 2016; Wai et al., 2010). Study 1 reported an overrepresentation of boys in the upper 5% - the magnitude of the difference was very small. There was no evidence of a ceiling effect for the most able achievers in mathematics, and the study sample tended to be overrepresented in the upper income quintile. Applying the MCS sample weightings to the analyses reduced the overall percentage of cohort members observed at the highest grade in mathematics to 3.6%; however, the ratio of boys to girls remained unchanged (AQA Education, 2017; Department for Education, 2018). There is, therefore, little reason to suspect that this result is not representative of the proportion of boys to girls in the 2017 population cohort. Hence, this thesis highlights that the conclusions drawn when examining sex differences in science and mathematics differ between school grades and international standardised tests.

Beyond sex differences in language, mathematics, and science subjects, Study 1 highlighted a general female advantage in achievement at age 16 across most subjects, including many traditionally male-stereotyped subjects. Together, these results confirm the widening female educational advantage between early and middle adolescence in a single sample (Cavaglia et al., 2020; Deary et al., 2007; Voyer & Voyer, 2014). The increasing female advantage resulted from more girls than boys maintaining or improving on their prior achievement levels. As the grades at ages 11 and 16 are the results from standardised curriculum tests, not teacher-assessed grades, this study demonstrates that the widening female advantage in this age group is found in both teacher-assessed and anonymously marked grades, which indicates that the female advantage widens irrespective of teacher grading biases (Lavy & Megalokonomou, 2019; Lavy & Sand, 2018). Indeed, recent evidence from the COVID-19 pandemic indicates that when grades are wholly teacher-assessed based on observations of engagement in class and in-class assessments, then the female advantage is wider still (Lee, 2021). While the widening female advantage in language-related courses may be predicted by current educational theories, the closing of the prior male advantage in mathematics and opening of a small female advantage in the sciences challenge theories that predict boys' achievement in mathematics and science to increase during this period (e.g., Ceci & Williams, 2010; Halpern et al., 2007). Currently, educational theories do not fully explain or predict a female advantage in science or parity in mathematics in these age groups. In particular, these theories are often based on observations from standardised aptitude tests and international standardised assessments, which have been reported to underpredict female achievement (Halpern et al., 2007; Kling et al., 2013; Matějů & Smith, 2015; Mattern & Patterson, 2014; Pulkkinen & Rautopuro, 2022; Wai et al., 2010). Study 1, along with several other large-scale studies, makes observations that are not yet fully explained (Cavaglia et al., 2020; Deary et al., 2007; Voyer & Voyer, 2014).

Sex differences in international tests are often conflated as sex differences in school performance (e.g., Kollmayer et al., 2018), despite that school grades are better predictors of longitudinal outcomes, such as graduating and earnings, than standardised tests (e.g., Starr et al., 2024; van Hoogdalem & Bosman, 2024). Schools, in general, are focused on delivering their country's or other jurisdictions' curriculum. International standardised tests are taken by a sample of students within each country, with some parts of countries undersampled (e.g., Jerrim, 2021). Schools are unlikely to be focused on teaching their students how to do well in international tests, and PISA tests are not directly related to what is taught in school; therefore, describing results in PISA tests as school performance is incorrect. The relative performance in international standardised tests is more likely to identify strengths and weaknesses within education systems in teaching students to apply what they have learned in new contexts. How students apply what they have learned will ultimately depend on the teaching approach taken. Therefore, some education systems will do better than others at closing the gender gaps in international test performance (Wu, 2010). In the UK context, girls achieve results that match or better those of boys at age 16 in mathematics and science taught in the UK curriculum, and some male-dominated subjects such as physics and computing at age 18, but the societal narrative and perceptions remains closer to the results from the international standardised tests (Carroll, 2024).

Finally, Study 1 concluded that greater male variability was unlikely to substantially contribute to sex differences in the uptake of STEM subjects after age 16. In line with O'Dea et al. (2018) 's findings, Study 1 reports lower variance in STEM subjects than language, reading and writing. Study 1 also found that girls and boys were equally represented in the upper 5% in science subjects, and the upper 12.5% in mathematics and the sciences, except biology and computing, where there were more girls. From simulated distributions, O'Dea et al. (2018) predicted that girls and boys would be equally represented in the upper 10% of STEM achievement, which is similar to what we observe. However, our results are by subject rather than grouped, and we do not find this to be true for all STEM subjects. Although it cannot be discounted that more boys are observed at percentiles above the 95th, from the analyses in Study 1, it is difficult to support the argument that the increasing representation of men at higher levels of mathematics and science achievement is due to the greater availability of high-achieving boys in STEM subjects at school (Baye & Monseur, 2016). In this age group, there are few differences between boys and girls in the upper 5% of achievers to justify the argument that ability underlies the underrepresentation of girls in these subjects based on school achievement.

5.2.2 Contribution to Wider Debates and Societal Concerns

The Boy Crisis. Study 1 supported the general narrative of the underachievement of boys relative to girls in education and a widening female advantage in early to middle adolescence, highlighting that this finding holds irrespective of whether school grades are teacher-assessed or not (Lavy & Megalokonomou, 2019; Lavy & Sand, 2018; Voyer & Voyer, 2014). Boys' lower average results compared with girls are likely to lead to more boys choosing not to continue in full-time education after age 16 (Carroll, 2024; Cavaglia et al., 2020; Stoet, 2015; Voyer & Voyer, 2014). A female advantage was observed in most subjects at the age of 16. Only in mathematics, physics, and economics did boys achieve similar or better grades than girls (see also: Carroll, 2024; Deary et al., 2007). As the UK Age 16 test results reported here directly contribute to the decision to stay in full-time education, or to leave to take a vocational pathway, the general female advantage in this age group will contribute to the university enrolment gap (Cavaglia et al., 2020; Stoet, 2015). Study 1 also supports reports of a substantial proportion of high-achieving students at age 11 not maintaining that level of performance at age 16, and these students are more likely to be boys, have an identified SEN, or come from lower-income groups (Montacute, 2018). However, it should also be highlighted that after both the age 16 and age 18 examinations, approximately 2% more boys choose to take apprenticeship opportunities than girls, which can offer a well-paid alternative to continuing into tertiary education, and so avoid accumulating university-related debt (Carroll, 2024; Fagence & Hansom, 2018). Therefore, at least some of the university enrolment gaps result from boys choosing viable options for their future careers and earning potential. There may be fewer available options, or gender stereotypes hamper access to available options, leading to a stronger incentive to obtain educational qualifications for girls (Blossfeld et al., 2015; Cappon, 2011; Carroll, 2024).

Boys were overrepresented at the lowest levels of reading achievement in school grades as previously reported in PISA data, which should be of concern to all (Baye & Monseur, 2016;

Stoet, 2015; Stoet & Geary, 2013). The UK's GCSE grades (Age 16 achievement in this thesis) have been compared with PISA scores and with OECD reading proficiency levels (Jerrim, 2019; National Literacy Trust, 2011). From these comparisons, it is concluded that the lower levels of achievement reported in Study 1 align with PISA reading proficiency level 1a, at which readers may be unable to spot fake news or bias and may have difficulties carrying out everyday tasks that require reading skills, both of which have societal implications (Department for Business, Innovation and Skills, 2011). Low literacy and low educational achievement have been linked to criminality, a lack of cohesion in family structures and communities, and health outcomes (Welmond & Gregory, 2021). In general, reading proficiency is as important for achievement in the sciences and mathematics as it is for achievement in language and art subjects. It has been shown to contribute to the university entry gap that favours women (GL Assessment, 2020; Stoet & Geary, 2020a). There are wide benefits to improving literacy for all that go beyond the aims of improving school achievement and closing achievement and university entry gaps, as there are health and wealth-related benefits to spending more years in education (OECD, 2023). However, this is also not just an issue for boys, which at 4.1% at this lowest level of achievement would suggest > 21,000 boys from the 2017 population cohort left school with very low literacy. It is also of concern for the 1.1% (> 5,500) girls who also left school after eleven years of education with similarly low literacy. More needs to be done to address this for all students. Nevertheless, while there are small gaps in reading achievement between boys and girls, as we highlight in Study 1, writing gaps may be larger (Reilly et al., 2019).

Gender Roles and Stereotypes: Lack of Girls in STEM and Boys in HEAL. The results from Study 1 do not support the proposal that fewer girls than boys are capable of studying STEM subjects at age 18 and beyond. Study 1's conclusions align with previous reports from examination of standardised tests, which indicate that fewer girls enrol in STEM subjects than are capable of studying them (Stoet & Geary, 2018; Wang et al., 2013). Large sex differences in the uptake of STEM subjects persist after age 16 (Carroll, 2024; Stoet, 2015), and sex differences in the upper achievement levels in school grades at age 16 do not explain this phenomenon. While group-level sex differences in age 16 achievement do not provide support for the under-enrolment of girls in STEM subjects in older age groups, or at university, several authors have suggested that individual girls' and boys' abilities in language-related subjects, compared to their ability in mathematics/science subjects influences later subject choice (Stoet & Geary, 2018, 2020b; Valla & Ceci, 2014; Wang et al., 2013). These authors argue that for most girls, their abilities and self-concepts tend to be tilted toward non-STEM subjects. In contrast, for most boys, their abilities and self-concepts tend to be tilted toward STEM subjects (Stoet & Geary, 2018, 2020b; Valla & Ceci, 2014; Wang et al., 2013). That is, individually, more girls are more confident and more likely to achieve better results in language-related subjects when compared with their confidence and achievement in STEM subjects. Conversely, more boys are more confident and more likely to achieve better results in STEM subjects than in non-STEM subjects (Stoet & Geary, 2018, 2020b; Valla & Ceci, 2014; Wang et al., 2013). There may also be more girls than boys who have both high verbal and mathematics skills, and these girls may not choose to study STEM subjects, despite being capable of doing so (Ceci, 2017; Stoet & Geary, 2018; Wang & Degol, 2017; Wang et al., 2013). Wang et al. (2013) argue that girls with high mathematics and verbal ability have more choices, with a wider range of occupations open to them, than boys with high mathematics skills but moderate verbal abilities. Further, girls with balanced abilities may rely on their interests, values, and are guided by subject-stereotypes when both STEM and non-STEM choices are an option for them (Ceci, 2017; Stoet & Geary, 2018; Wang & Degol, 2017). Study 1 did not examine the patterns of individual achievement and self-beliefs across subjects, as has been done previously using PISA and SAT data (Stoet & Geary, 2018, 2020b; Wang et al., 2013). It may be here that smaller sex differences are observed at the upper tail in combined measures of achievement. Sex differences in the distribution of combined measures of achievement are an area where the current evidence could be extended to examine how mathematics, science, and language achievement, when combined, may be distributed between boys and girls, rather than focusing solely on sex differences in individual subjects. Additionally, once the Age 18 achievement data is made available for this cohort, then future research could examine how sex differences change among the more motivated students who continue in full-time education until 18 years old and how Age 16 achievement and prior self-beliefs influence subject choice at the end of secondary education.

Observing the grade distribution patterns at age 16 in optional subjects, it is plausible that the students who chose counterstereotypical subjects were, on average, more able than those who decided to study stereotypical optional subjects. Although Study 1 did not examine the reasons why fewer girls enrolled in optional STEM subjects or boys in non-STEM subjects, the male and female grade distributions in some subjects were very different. The disparity between the male and female distributions in optional subjects indicated that very different profiles of boys and girls chose those subjects. For example, despite computing being a male-stereotyped subject and that more than 3.5x more boys enrolled, 12 percentage points (pp.) more girls were observed at the upper and highest levels of achievement, and there were fewer than half as many girls as boys observed at the lower achievement levels. A similar observation is made when computing results are examined at the age of 18. Substantially more boys enrolled to study computing, but more girls have achieved the highest two grades over twelve years of results (Carroll, 2024). Interestingly, the opposite pattern is seen in language courses at age 18, in which more girls enrolled in the German, Spanish and French courses, but more boys achieved the top two grades (Carroll, 2024). In Study 1, at age 16, we found no sex differences at the upper achievement levels in German and no sex differences in the highest grade in French, which differed from the findings for most humanities and language-related subjects. However, at age 18, the male advantage in the top two grades of these modern foreign languages (German 3.4pp., Spanish 1.5pp., French, 1.6pp.), and small male advantages in the top two grades in chemistry (1.7pp.) and mathematics (0.5pp. - but not further mathematics), and other science (1.4pp.), are the only subjects in which boys achieved a slightly higher proportion of the highest grades at age 18 over the last twelve years (Carroll, 2024). From these observations, it is possible that those students who choose counterstereotypical subjects may be those who are more confident in their abilities than, perhaps, the average level of confidence of those students choosing stereotypical subjects. In general, students demonstrate more motivation in subjects where their gender identity and the subject-stereotypes match (Wirthwein et al., 2020). Therefore, it may be the highest-achieving and most confident students in a given subject who are choosing to study against the gender stereotypes. Choosing to defy gender stereotypes can come at a social cost, which may therefore prevent many students from choosing counterstereotypical subjects that they are very capable of doing well in, with some reports suggesting it may be more difficult for boys to challenge stereotypes (Mulvey & Killen, 2015; Rudman & Fairchild, 2004; Rudman et al., 2012; Skipper & Fox, 2022). Other research has found that counterstereotypical students can be perceived as more competent, particularly among higher-SES students, which would suggest a social gain from making counterstereotypical choices for some students (Meimoun et al., 2024). Again, this is an area where the current evidence could be extended further. If there is a higher social cost for low-SES students when challenging the stereotypes, this may contribute to less optimal decision-making when making subject choices. This social cost may be heavier for those students who are more susceptible to the influence of their peers or are more likely to align with the views of influential adults in their social network (Kollmayer et al., 2018; Skipper & Fox, 2022; Yu et al., 2020).

Gender Roles and Stereotypes: Teacher Bias. Some disparities between teacher-assessed and standardised curriculum test grades found in Study 1 could be interpreted as support for a teacher bias account. However, the evidence was not conclusive, and where differences were found, they were small. Teachers form expectations of students that are influenced by student demographics, behaviours and engagement, teacher beliefs, class and school factors, and the student-teacher relationship (Wang et al., 2018). Once formed, teachers interact with their students according to their expectations of them. There is consistent evidence pointing to teachers forming lower expectations of low-SES and SEN students, with the latter of these disproportionality influencing teachers' expectations of boys (Kashikar et al., 2023, 2024). Many other studies have found evidence for ethnicity and gender biases in teacher expectations, although not in all studies (Wang et al., 2018). Teacher grading bias was particularly highlighted during the recent COVID-19 pandemic, when many standardised curriculum tests were replaced with teacher grades as examinations were cancelled, and the female advantage substantially widened (Lee & Shi, 2021).

Teacher bias was not a primary focus of this empirical chapter or this thesis. Nevertheless, it is an important consideration within the academic field, and the linked educational data provides a suitable mechanism for examining teacher bias (Burgess & Greaves, 2013). Study 1 examined sex differences at the mean and in the upper and lower achievement levels in the teacher-assessed grades in supplementary analyses and compared the results with the outcomes from the standardised curriculum tests. When comparing sex differences, if teachers endorse gender stereotypes, one might expect to find that teachers generally award higher grades to girls or otherwise find that subject grades are biased in line with gender-based subject stereotypes (Campbell, 2015; Harlen, 2005; Malouff & Thorsteinsson, 2016; Terrier, 2020). Teacher endorsement of gender-based subject stereotypes when assessing students would predict teacher grades favouring boys in mathematics and girls in English. Teacher bias has been reported previously in this cohort, comparing cognitive ability test results with teacher survey responses using the Foundation Stage Profile — a measure of general school readiness at age 5 (Hansen & Jones, 2011), and comparing teacher subjective judgement of ability (e.g., ability in maths: above or below average) with cognitive ability test results (Campbell, 2015). However, intelligence and cognitive ability tests are criticised for being biased against some ethnic and lower-income groups, and neurodiversity can influence an individual's performance on some aspects of the tests (e.g., van Hoogdalem & Bosman, 2024; Woods & Patterson, 2024). The measurement of cognitive ability through tests is less stable in younger children, less stable in lower ability groups, and there is a period of change in verbal and non-verbal IQ during adolescence related to brain changes during that period (Breit et al., 2024; Ramsden et al., 2011). Comparing teacher and test-assessed grades that measure the same or similar learning is an alternative approach to examining teacher bias (Burgess & Greaves, 2013; Lavy & Megalokonomou, 2019; Lavy & Sand, 2018). Previous comparisons in the UK context have reported support for bias in teacher assessment for SEN students, some ethnic groups, and pupils from lower SES groups (Burgess & Greaves, 2013). In the Greek context, teacher bias is reported to vary between teachers, differing between more and less effective teachers. Teacher biases were reported to favour either boys or girls (Lavy & Megalokonomou, 2019; Lavy & Sand, 2018). However, comparing the teacher-assessed and standardised curriculum test grades has also been criticised as a teacher's performance is assessed based on their class's performance, and therefore it is in each teacher's interest to present a positive set of class results (Campbell, 2015).

Study 1 found that teachers assessed students less often in the lowest grades, which may reflect other motivations influencing teacher grades or differences between performance in tests vs. performance over time in class (Burgess & Greaves, 2013; Campbell, 2015). We observed that teacher-assessed grades were less granular than test grades, allowing teachers fewer opportunities to differentiate between students. The impact of less granularity in grade structures was particularly evident at age 7, such that the results at the highest achievement level represented a very broad indication of the differences in achievement between students. More students were observed in the lower grades in the mathematics and English tests than were assessed at that level by their teachers. Further, no students achieved the highest available grade in the English test, despite some cohort members being teacher-assessed at that grade, which would be more likely to support an account highlighting differences between test and class performance (Burgess & Greaves, 2013). More boys were teacher-assessed at the lower grades in mathematics, while more girls achieved these grades in the tests. This pattern might suggest lenient teacher assessment of girls in mathematics at the lower tail in line with prior reports of a bias towards girls (Malouff & Thorsteinsson, 2016; Terrier, 2020). Nevertheless, there may be other plausible explanations for why achievement in tests may be lower than teacher-assessed grades. These explanations may include the influence of text or mathematics anxiety, which is higher in girls but may be related to girls' lower perceived competence in mathematics (Devine et al., 2012; Goetz et al., 2013; Putwain & Wood, 2023). There is also the potential impact of teaching to the test, and some students may have received additional support or tutoring in preparation for the tests (Burgess & Greaves, 2013). Another possibility may be that observed differences between test and teacher-assessed grades result from other group biases rather than reflect gender biases. For example, teacher bias against SEN students - a group in which boys are overrepresented - or related to ethnic group membership (Burgess & Greaves, 2013; Daniel & Wang, 2023; Hansen & Jones, 2011; Perinetti Casoni & Barg, 2023). It is also possible that biases are minimised due to the use of detailed assessment criteria and moderation processes (Harlen, 2005; Testing and Standards Agency, 2018).

5.3 Study 2: Puberty and Age Associations with School Engagement

5.3.1 Contribution to Theory and the Literature

Study 2 examined whether the earlier maturation of girls may contribute to the widening female advantage between the ages of 11 and 16, as measured by standardised curriculum tests, by investigating the reported association between puberty and school engagement (Martin et al., 2022). The earlier maturation of girls is commonly highlighted as contributing to the female advantage in secondary education (APPG, Men & Boys, 2023; Torvik et al., 2021; Woolcock, 2024). Study 2 examined associations between puberty status and changes in school engagement between early and mid-adolescence. School engagement is predicted by academic self-concepts and, more broadly, achievement motivation, including educational expectations, leading to changes in achievement (Huang, 2011; Skinner & Raine, 2022). School engagement is theorised to be influenced by how well the school environment is perceived to meet the changing psychological needs of individual students as they develop (Eccles et al., 1993). Study 2 found that changes in perceived puberty status were associated with increased emotional and behavioural disengagement in girls and decreased behavioural engagement in boys. The results align with prior reports that perceived puberty influences achievement motivation (Martin et al., 2022), where achievement motivation is theorised as a predecessor of school engagement and disengagement (Skinner & Raine, 2022), but differed from Martin et al. (2022) as the association was stronger for girls. In general, Study 2's results would point to an association between more advanced perceived puberty status and decreasing school engagement, more so for girls.

Study 2 supports earlier findings of an association between puberty and school engagement during early to middle adolescence, and steeper declines for girls in language self-concept, which we also find for school engagement (Martin et al., 2022; Schaffhuser et al., 2017). To my knowledge, this is the first study to explore how puberty relates to both behavioural and emotional dimensions of school engagement. Among girls, increases in perceived pubertal status were associated with greater emotional and behavioural disengagement. Although a similar trend was observed for behavioural engagement, the additional effect of puberty, beyond the stronger association with age, was minimal. These findings align with previous research indicating that early-maturing girls are at greater risk of lower academic achievement (Goering et al., 2023; Mendle et al., 2007; Senia et al., 2018; Stattin & Magnusson, 1990). Sharper declines in school engagement may contribute to these findings, particularly as school engagement may not begin to increase until late adolescence, for girls only (Martin, 2012). However, some students will have chosen to leave full-time education before late adolescence (Carroll, 2024). Behavioural disengagement, in particular, showed a stronger correlation with later academic outcomes than other engagement measures for both sexes. For boys, perceived pubertal changes were linked to reduced behavioural engagement, though the connection between behavioural engagement and academic performance was weak. Interestingly, small positive correlations were found between behavioural engagement and both mathematics and language self-concepts. At the same time, puberty was associated with lower mathematics self-concepts but higher language self-concepts in boys. These patterns may reflect the growing influence of self-identity and social comparison as adolescents mature (Marsh, 1987). For instance, the transition from smaller primary schools to larger secondary schools in the UK often involves academic streaming based on prior achievement. The big-fish-little-pond effect (Marsh, 1987), for example, proposes that students' self-concepts are shaped by the average performance of their peers, positively when surrounded by lower-achieving peers, and negatively when surrounded by higher-achieving students (Liem et al., 2013; Marsh & Hau, 2003; Marsh et al., 2008). The transition from primary to secondary school may trigger a different understanding of an individual's relative achievement in comparison to a larger sample of peers.

Contrary to earlier findings, Study 2 found stronger associations between perceived puberty and disengagement in girls, whereas puberty hormones were more strongly linked to disengagement in boys (Martin et al., 2022). This discrepancy may stem from differences between self-reported and biological measures of puberty (e.g., Martin et al., 2017). While these measures are correlated, they may capture distinct social processes associated with how adolescents perceive their maturity relative to their peers (Balzer et al., 2019; Carter et al., 2017, 2018; Mendle, 2014; Torvik et al., 2021). Although the associations between puberty and school engagement were generally small, they contribute beyond age and changes in self-concept or educational expectations. Among boys, puberty and age had opposing effects on behavioural engagement: lower levels of maturation appeared protective in mid-adolescence, while increasing age was linked to declines. There was also a tentative suggestion that early maturation might be beneficial earlier in adolescence, though this was not conclusive. The use of parent-reported puberty measures in younger participants may have underestimated these effects, as such reports are likely to miss the social processes captured by self-reports (Dorn et al., 2003).

The steeper declines in school engagement and language self-concept coincided with a faster rate of average perceived pubertal change for girls than for boys. While there are inaccuracies in perceived measures, especially in the earlier pubertal stages, when compared with biological measures of pubertal development, the change between parent-reported puberty status at age 11 and self-reported puberty status at age 14 was larger for girls (Dorn et al., 2003). The speed of girls' perceived pubertal change and steeper declines in school engagement and language self-concept, when compared with boys, would provide support for proposals of the stressful change hypothesis for girls (Simmons & Blyth, 1987). While the puberty associations with school engagement were mostly negative for boys, they were sometimes positive and sometimes negative, and there was some support for opposing effects in different age groups. In general, there is less evidence examining boys' puberty and educational outcomes, and the results remain inconclusive at present, indicating a clear gap in our understanding of how male puberty may be associated with school achievement (Mendle & Ferrero, 2012; Mendle & Koch, 2019). Some studies examining the effects of male puberty find support for lower achievement in early maturers, some find support for lower achievement in later maturers, some report lower achievement for both early maturers and late maturers and some report no association (Goering et al., 2023; Mendle & Ferrero, 2012; Senia et al., 2018; Torvik et al., 2021). The mixed results, especially as there is limited support for an association between cognitive development and puberty, may point to psychosocial mechanisms underlying associations with achievement (Herlitz et al., 2013; Mendle & Koch, 2019; Paus et al., 2017). For example, late-maturing boys report lower self-esteem and being less popular than early-maturing boys, and can be subjected to victimisation and bullying (Jormanainen et al., 2014; Mendle & Ferrero, 2012; van Rijn et al., 2023). These associations would predict a negative effect on achievement in older age groups for late maturing boys, whereas here we find a negative association between achievement and early maturation. Links between early maturation and lower achievement support the proposal that there is a mismatch between physical and cognitive development during mid-adolescence that places some individuals at risk of poor outcomes through being exposed to situations that they are not developmentally ready to manage (Mendle & Ferrero, 2012). Associations between puberty and achievement may also be indirect (Martin et al., 2017; Stattin & Magnusson, 1990).

Study 2 provides evidence of an association between perceived puberty and declines in school engagement, especially for girls. There was, however, less evidence to suggest that within an age group, perceived puberty status would result in significant differences between the most mature and least mature girls. It is plausible, though, that there may be differences between the oldest, most mature girls and the youngest, least mature girls within a particular year group. One exception to this may be emotional disengagement, for which puberty was more influential than age. Within-year-group effects were not tested, and this remains an open question, particularly one that should focus on girls with lower self-concepts and educational expectations who do not benefit from the protective effects of higher self-appraisals and educational expectations, which I discuss further below. These results would suggest that higher school disengagement contributes to underachievement in more mature girls. Potentially, the higher disengagement of more mature girls contributes to findings that early-maturing girls spend less time in education, as enjoyment of school is needed to motivate individuals to continue in education, (Gill et al., 2017; Stattin & Magnusson, 1990; Wang & Eccles, 2012). Still, more mature girls do not necessarily do less well at school (Gill et al., 2017; Stattin & Magnusson, 1990). Perceived puberty associations for boys were limited to behavioural engagement, for which puberty associations were stronger than age associations. The results suggested that early maturers among younger boys and less mature, older boys reported higher school engagement than their peers. Again, the difference between boys of the same age was not substantial.

School Transitions. Students in early secondary school often reported similar or higher school engagement, which opposed the directional associations of puberty with school engagement. Study 2 briefly explored school engagement and self-concepts in the context of school transitions. Declines in engagement are often reported during school transitions, so this study examined the impact of moving from primary to secondary school on students' engagement and self-concepts (Wang et al., 2015). At the Age 11 survey wave, fewer than 2.5% of boys and girls had already transitioned to secondary school. Post-hoc exploratory analyses, therefore, compared those who had transitioned and those who had not, analysed separately by sex. Interestingly, the transition appeared to have a generally positive effect, in line with other reports that school transitions are not always negative (Jindal-Snape et al., 2020; Symonds & Galton, 2014). Boys in their first year of secondary school reported higher school engagement than boys still in primary school. Although their mathematics self-concept was lower, the difference was not statistically significant. For girls, self-concepts remained consistent across both groups. However, girls in secondary school reported significantly higher emotional engagement and were less unhappy at school than their primary school counterparts. These findings suggest that the effects of puberty on engagement, which were largely negative, are distinct from the changes associated with school transition. Due to the small number of students who had transitioned, combined models accounting for both puberty and school transition effects were not tested. However, it is plausible that puberty may intensify the impact of school transitions, particularly for students who are already at risk from negative school transitions, such as students with special educational needs (SEN), lower academic ability, or from disadvantaged socioeconomic or ethnic backgrounds (Bagnall et al., 2021; Schaffhuser et al., 2017; Sutton et al., 2018; West et al., 2010). This proposal would align with the accentuation hypothesis (Caspi & Moffitt, 1991), which proposes that stressful transitions amplify existing vulnerabilities. Future research should explore these interactions longitudinally to understand better how puberty and school transitions are differentially associated with school engagement.

Self-Concepts and Educational Expectations. Puberty associations with school engagement were less influential than those of self-concepts and educational expectations. The stronger association with these motivational variables supports previous findings on the importance of achievement motivation (Martin et al., 2022). Notably, a curvilinear pattern emerged for boys: those with the lowest self-concepts were not significantly less engaged than their peers, except for boys with the highest self-concepts, who reported markedly higher emotional and behavioural engagement. One plausible explanation is that boys with low self-concepts may receive targeted learning support, and positive relationships with support teachers could foster emotional engagement (Bakadorova et al., 2020). However, this pattern was not observed in girls, which undermines this as an explanation or may indicate that supportive relationships with teachers operate differently for girls. For example, students with ADHD, who may be overrepresented in lower ability groups, often have poorer relationships with teachers, and this is especially the case for girls with ADHD (Rogers et al., 2015). Alternatively, subgroup differences may underlie these findings. For instance, boys with low self-concepts but relatively stable emotional engagement may resemble the "cool guys" identified by Yu et al. (2020). Cool guys self-handicap, show low perseverance, and perform poorly academically, yet maintain high social status and visibility. This social standing may buffer their sense of school connectedness despite low academic motivation. Yu et al. (2020)

also identified "wild girls" and "modern girls," who made up half of their sample and showed similar patterns of low perseverance, high self-handicapping, and lower achievement in English. These groups, like the "cool guys," may prioritise peer approval over academic effort. While low academic self-concepts can hinder achievement, they may not reduce social connectedness, especially for students who align with gender-normative behaviours and popularity dynamics (Heyder et al., 2017; Yu, 2019). On the other hand, Yu et al. (2020) proposed that 'relational girls' with highly adaptive engagement profiles across mathematics and language may underlie the general female advantage found in secondary education. With the current data, it would be feasible to examine whether students who improved their academic performance between ages 11 and 16 shared similar engagement profiles to those identified by Yu et al. (2020).

Associations of Self-Concepts and Educational Expectations with School Engagement. Associations between self-concepts and school engagement differed for boys and girls. In general, self-concepts were often more predictive of engagement than disengagement, with notable differences between sexes. Girls' ability beliefs in both core subjects were more predictive of their school engagement than those of boys. Boys' school engagement was more strongly associated with language self-concept than mathematics self-concept, and their engagement was less influenced by ability beliefs overall. Nevertheless, educational expectations remained equally protective for both genders. Overall, these findings echo earlier research by Martin et al. (2022), which reported sex differences in the relationship between valuing school and disengagement, particularly for boys. However, direct comparisons are limited, as educational expectations in this study encompass broader considerations, such as the perceived feasibility of attending university, which in turn would place a higher or lower value on achievement in mathematics and language. Educational expectations are shaped not only by self-beliefs and prior achievement but also by socioeconomic and familial factors (Dockery et al., 2022; Koshy et al., 2019; Rampino & Taylor, 2013). The stronger positive association of ability-beliefs and school engagement, in the context of a generally declining trend, mirrors previous findings on the importance of self-efficacy in fostering school engagement (Martin et al., 2022).

Special Educational Needs. Students with an identified SEN reported lower school engagement than typically developing (TD) students at the end of primary school. Group differences in school engagement were smaller in mid-secondary school. Students with SEN also reported persistently lower mathematics self-concepts and language self-concepts across survey waves. Relatively few studies have directly examined school engagement in students with SEN (Moreira et al., 2015; O'Donnell & Reschly, 2020). Study 2, therefore, adds to the limited evidence to date. Similarly, relatively few studies have examined school transitions for students with SEN (Bagnall et al., 2021; Jindal-Snape et al., 2020). The MCS sample used for these secondary data analyses has a high proportion of students with an identified SEN, which enabled a brief exploration of SEN and TD differences in the four measures of school engagement and self-concepts. In line with other reports, substantially more boys in this sample have an identified SEN (e.g., Daniel & Wang, 2023; Department for Education, 2023; Mead, 2006). Students with SEN ended primary school reporting lower school engagement than TD students, aligning with reports of lower school connectedness and poorer relationships with teachers among primary age children with SEN (Murray & Greenberg, 2001). School engagement differences between students with SEN and TD students had declined by 14-15 years old, with the exception of feeling tired at school, where the difference had increased. This finding of similar school engagement for TD and SEN students in mid-adolescence may align with Moreira et al. (2015), who reported that the lower school engagement reported by ADHD students did not differ significantly from TD at age 13. However, I did not examine which diagnoses were prevalent among the students with SEN diagnoses in this sample. In general, differences in school engagement may depend on diagnosis or area of disability. For example, larger differences for students with visual impairments were reported (Moreira et al., 2015). Feeling tired at school increased between the end of primary school and mid-secondary school and likely underlies the persistent SEN association with emotional disengagement for boys reported in the main analyses (see Study 2 for details). For most of the school engagement measures, the smaller gaps resulted from steeper declines in TD school engagement trajectories. Given the higher prevalence of girls in the TD sub-sample, this observation may reflect that the TD trajectories were influenced more by decreases in girls' school engagement. In contrast, school engagement trajectories for students with SEN were influenced more by boys' school engagement, which declined less than girls. Hence, it is important to distinguish between male and female students with SEN in future studies, as trajectories differ between the sexes and between TD and SEN students. Some SEN diagnoses can present differently in girls and boys, for example, ADHD and Autism/ASD, and girls' and boys' school engagement profiles and trajectories differ (Mayes et al., 2020; Santos et al., 2022). Future studies should examine how students with SEN are influenced by transitions and puberty, perhaps using a transdiagnostic-style approach (e.g., McDougal et al., 2020), as there may be sex-specific effects within SEN groups, particularly during the early-middle adolescent period.

5.3.2 Contribution to Wider Debates and Societal Concerns

Puberty, Educational Achievement, and Widening Achievement Gaps. Study 2 does not provide support for the proposal that the early pubertal maturation of girls, as measured at ages 11 and 14, contributes to the wider sex achievement gaps observed at age 16. Correlations between age 14 perceived pubertal status and later achievement at age 16 were negligible, in line with other reports (Torvik et al., 2021), and school engagement and self-concepts declined during this period. However, other reports have associated puberty status measured at later ages (e.g., age 16: Torvik et al., 2021) with the female advantage in education at this age 16. It is plausible that developmental trajectories of some psychological measures are curvilinear (e.g., conscientiousness, self-discipline, school engagement: Martin, 2012; Soto et al., 2011). For example, despite the observed declines in early to middle adolescence, there may be some recovery or increases later, or the lowest point or recovery for boys' school engagement occurs later than for girls (e.g., De Fraine et al., 2007; Jacobs et al., 2002; Martin, 2012; Soto et al., 2011). Other psychological measures may offer alternate explanations for the association between puberty and achievement at age 16, and this remains an area for further research. For example, (De Bolle et al., 2015) reports a female advantage in conscientiousness between ages 12 and 17 years old, and that the gap between boys and girls is larger during early to middle adolescence. Some studies that examine girls' and boys' trajectories over time often report that mean scores for boys and girls are similar at ages 15-16 years old (e.g., conscientiousness, self-discipline: Soto et al., 2011). Conscientiousness has been shown to predict achievement beyond the influence of intelligence, and girls' higher self-discipline, a facet of the conscientiousness personality construct, has been associated with girls' higher achievement at school, and as an explanation for why standardised tests score under-predict girls' school achievement (Duckworth & Seligman, 2006; Poropat, 2009). Here, despite the observed declines in school engagement and language self-concept between the two survey waves, girls remained more behaviourally engaged and less behaviourally disengaged than boys, retaining a very small advantage in language self-concept. Correlations between school engagement measures and longitudinal mathematics and language grades did not significantly differ between boys and girls. The negative correlations between behavioural disengagement and school grades were slightly stronger than the other measures, in line with previous reports (Wang et al., 2017). Therefore, girls retained a small school engagement advantage despite the evident declines.

In addition to retaining a small school engagement advantage at age 14, girls also reported higher educational expectations that were moderately associated with later achievement at age

16, which likely contributes to the female advantage. Girls' and their parents reported higher educational expectations, of small magnitude, than boys and their parents, as also reported elsewhere (e.g., Dockery et al., 2022; Koshy et al., 2019). The moderate correlation between educational expectations and later achievement at age 16 aligns with the findings from the meta-analysis of parental educational expectations and achievement (Pinquart & Ebeling, 2020). The widening female advantage in secondary school may in part reflect the influence of differences in (perceived) labour market opportunity differences for boys and girls, and the lower availability of viable options with good earning potential for girls at lower achievement levels (Blossfeld et al., 2015; Cappon, 2011; Carroll, 2024). This account would suggest that the available options, or lack thereof, create a stronger incentive for girls to continue in education and gain further qualifications. The fact that boys are more likely to take apprenticeship opportunities may be indicative of this (Blossfeld et al., 2015). In addition, other reports find that fathers in trade occupations often report lower educational expectations of their sons, perhaps because these are seen as well-paying alternatives to continuing in education (Dockery et al., 2022; Koshy et al., 2019). Sons following in the footsteps of their fathers has a longstanding tradition in many countries, and is associated with a higher likelihood of finding a job (Lo Bello & Morchio, 2022). However, it is also associated with lower wages. Choosing your father's occupation may make finding a job easier due to access to parental networks and information, offering a comparative advantage over other choices. However, this may not necessarily be where the son's interests and talents lie (Lo Bello & Morchio, 2022). Pekkarinen (2008) observed that changing the timing of educational streaming between vocational and academic options from the end of primary (early-selection) to later in secondary education (late-selection) was associated with lower academic achievement for boys. The author went on to suggest that this may result from the influence of sex differentiated maturational processes during adolescence, meaning that under the late-selection, boys were less mature and this influenced their decision-making. However, these explanations are less likely to explain the underrepresentation of boys at the upper achievement levels, except, perhaps, less optimal subject choice or career decisions having a potential link with later achievement (e.g., Cuff, 2017; Lo Bello & Morchio, 2022).

The SEN Crisis. The underachievement of students with SEN within the UK's educational systems is longstanding. Despite political recognition and governmental enquiries - the most recent call for evidence was made in December 2024 (Education Select Committee, 2024) - little progress has been made. Education systems are observing increased demand for support, and therefore cost, as the proportion of students diagnosed with support needs has risen to around 18% of students within the school system (Roberts & Long, 2025). In the 2017 GCSE cohort from which this sample is taken, approximately 14% students had an identified SEN (Department for Education, 2018). In this context, the paucity of research into school engagement and school transitions for students with SEN is of concern. This thesis adds to this limited evidence, highlighting that future studies may need to differentiate between boys and girls with SEN in their analyses. It is widely recognised that students with SEN are at risk of lower achievement, poor psychosocial outcomes, and are at increased risk of finding school transitions more challenging (Bagnall et al., 2021; Hannah & Topping, 2013; Hannah & Topping, 2012; Hughes et al., 2013; Jindal-Snape et al., 2020). The evidence to date suggests that children with SEN can be at risk of peer problems, including victimisation and bullying, teacher alienation, and lower school connectedness, with notable individual differences and differences between diagnostic groups (Bear et al., 2015; Murray & Greenberg, 2001, 2006). Lower school connectedness is associated with perceptions of lower competence, and social problems are associated with poorer social, emotional, and behavioural adjustment (Murray & Greenberg, 2006). The higher proportions of boys diagnosed with SEN will therefore very likely contribute to sex differences in achievement and the boy crisis in education. Improving our understanding of which students with SEN are most at risk, and when, is crucial to understanding how best to support students with SEN throughout their education to enable them to meet their potential.

5.4 Limitations and Future Directions

The use of a large, longitudinal panel study has substantial benefits when examining achievement trajectories in education. They allow researchers to test theoretical predictions against the life trajectories of a large cohort over longer periods than is practical for most research (Gorard, 2003). Often, however, these panel studies are designed to cover a broad range of outcomes and, as such, can overly simplify the measurement of some psychological constructs. This oversimplification can limit the extent to which the measures used truly represent the psychological construct of interest (Allen et al., 2022). For example, the school engagement measures used here are biased towards emotional disengagement. Single items measured the other engagement dimensions and self-concepts, and these constructs may be more nuanced than the single items can truly capture, which may increase measurement error. Other research asserts that single-item measures are a good alternative (e.g., Castro et al., 2023). Additionally, the school engagement items do not capture the full breadth of the school engagement construct and are not closely linked to more recent discussions on this concept. Nevertheless, single-item measures are not uncommon in the literature. They have benefits including being accessible for a wider range of individuals, being faster to complete, and reducing data processing costs (Bergkvist & Rossiter, 2007; Castro et al., 2023; Wanous et al., 1997). Time to complete and data processing costs are important considerations in the context of large panel studies sampling widely from the population, and in which the surveys completed measure many constructs of interest. Shorter surveys can also increase participants' willingness to complete and submit their responses (e.g., Wanous et al., 1997). Despite these limitations, the trajectories of single items should broadly match those predicted by the psychological theories being examined, or at least identify some potential contradictions or inconsistencies, and have often been shown to be as reliable and valid as multi-item constructs (Ahmad et al., 2014; Allen et al., 2022; Ang & Eisend, 2018).

The technical challenges encountered with the multiple imputations in Chapter 4 will also need to be addressed. The missing imputation of the emotional disengagement data for girls, and therefore, reliance on listwise deletion, will likely limit the accuracy of the coefficients reported. Nonetheless, when comparing the regression models with and without imputed data for boys, the imputed result does not substantively alter the conclusions drawn. How best to resolve this is a matter for further consideration. The issues encountered pointed to an internal package coding error and, therefore, may require further investigation and potentially the support of the developer and a subsequent software release to be able to complete the imputations. An alternate approach would be to investigate using an alternate statistical software, such as MPlus (Muthén & Muthén, 1998). Using MPlus would avoid the need for multiple imputation as full information maximum likelihood (FIML) could be used instead (Muthén & Muthén, 1998). However, there can be convergence issues with FIML when data is skewed, which is a feature of these data (Lim & Cheung, 2022). Currently, though, Mplus is not available in the UK Data Service Secure Lab environment where these data were analysed. An alternate option is to modify the current Bayesian regression models to impute the data during model fitting (Bürkner, 2024; McElreath, 2020). Imputing missing data during model fitting would further increase the processing requirements.

Throughout this chapter, I have highlighted several potential areas in which this research could be further developed. From Study 1, once the Age 18 achievement data is available, the evidence could be extended to examine achievement trajectories among the more motivated students who continue in full-time education until 18 years old. Access to the Age 18 data would enable an investigation into how combinations of achievement and self-concepts contribute to subject choice at age 18 (e.g., Wang et al., 2013). The evidence could be further extended to examine sex differences in subject combinations at the upper tail and whether these contribute to sex differences in career or subject choice. The widening sex difference in writing achievement during adolescence, which is not measured separately in the UK education system, warrants further investigation to determine whether this phenomenon is observed in the UK system, and if it continues to grow as reported in the US analyses (Reilly et al., 2019). Ideally, future work here would clarify how writing gaps might influence overall academic outcomes in comparison to the reading gap. From Study 2, the influence of puberty on school achievement, especially among boys, remains underexplored, and studies using continuous measures of both biological and perceived puberty across diverse populations are needed to clarify how puberty is differentially associated with boys' and girls' achievement in school (Mendle & Koch, 2019). Clarifying exactly how or why adolescent development and achievement may be related opens the door for interventions to address risks to students achieving in line with their potential. Additionally, the social costs and motivations behind counterstereotypical subject choices, particularly among students from lower socioeconomic backgrounds, should be investigated to understand how peer and societal influences shape educational trajectories (e.g., Cuff, 2017; Skipper & Fox, 2022). The supplementary analyses on teacher bias could be examined more systematically at an individual level, ensuring that a broader set of student characteristics is included. Teacher subjective assessments, teacher grades that utilise detailed evaluation criteria, and outcomes in standardised tests could be compared, comparing differences to long-term outcomes. Finally, future work should disaggregate findings by sex within special educational needs (SEN) populations to better understand how transitions and engagement differ across diagnostic groups and genders. These directions will help clarify the complex mechanisms underlying sex and SEN differences in educational outcomes and inform more equitable educational policies and practices.

5.5 Thesis Summary

This thesis examined sex differences in educational outcomes in a single longitudinal sample. It went on to investigate whether the association between puberty and school engagement may support the claim that the earlier pubertal maturation of girls contributes to widening sex achievement gaps during adolescence. Study 1 contributed to the academic literature and societal discourse by offering a detailed, longitudinal analysis of sex differences in school achievement across core subjects from childhood to adolescence (RQ1). It confirmed that the widening female advantage is observed in most subjects during adolescence (RQ2). It demonstrated that during adolescence, sex differences in STEM subjects are either null or favour girls, despite earlier male advantages in mathematics (RQ3). This thesis, therefore, challenges assumptions that the male majority in STEM careers is due to the higher availability of boys at high levels of school achievement at age 16, and found limited support for greater male variability in school grades in older age groups (RQ5). Study 1 also highlights the different conclusions that may be drawn when assessing sex differences in STEM subjects using international standardised tests or school grades (RQ6). The study also underscores the societal implications of male underachievement, particularly in literacy, and the persistent gender imbalances in subject choices, which are likely shaped by stereotypes and self-concepts. While evidence of teacher bias was limited (RQ4), the findings call for further investigation into how teachers may influence educational outcomes and whether teacher bias is associated with other groups, such as having low expectations of SEN students. The study advocated for a more nuanced understanding and explanation of gendered educational trajectories in STEM subjects and supports the need for targeted interventions to promote equity and informed subject choice for all students. Having confirmed the widening female advantage during adolescence, Study 2 examined how puberty, self-concepts, and educational expectations interact to shape school engagement during adolescence (RQ7). While perceived puberty associations with engagement were generally modest, they were more pronounced for girls and added explanatory value beyond age, self-concepts, and educational expectations. Girls experienced steeper declines in school engagement and language self-concept, supporting the stressful change hypothesis, which did not support the proposal that girls' earlier maturation is beneficial, at least between ages 11 and 14 (RQ8). Study 2 was unable to test this proposal after age 14, so this remains an open question for future research. In contrast, during early to middle-adolescence, boys showed more variable patterns, with language self-concept playing a more protective role than mathematics self-concept. Students who had transitioned to secondary school reported higher school engagement on some measures, suggesting that school transition can be beneficial for some students. Ultimately, self-concepts and educational expectations emerged as stronger predictors of school engagement than puberty alone, underscoring the importance of fostering positive academic self-beliefs and educational aspirations to support students through developmental transitions and throughout secondary education.

5.6 Final Remarks

Recently, several researchers have proposed integrated motivational models incorporating several disparate branches within educational psychology into broader frameworks in an effort to consolidate and reduce complexity and fragmentation within the field (see Pekrun, 2024, for a review and commentary). As highlighted by Pekrun (2024), there are significant challenges to this endeavour. This thesis did not set out to test the predictions of these integrated theoretical models - some of these have been published late into this overall exercise, and this secondary data analysis is, necessarily, partly driven by the available data. Notably, stage environment fit theory was highlighted by Skinner (2023) as an important model through which an integrated motivational theory can encompass developmental change and, therefore, explain the trajectory changes that are observed as individuals develop. Adolescence is a period of substantive change as individuals refine their perceptions of the self, and their self-appraisals become increasingly influenced by reality and achievability (Marsh & Hau, 2003). The existing theories of adolescent development have been aimed at explaining the onset of mood and anxiety disorders and increased risk-taking and sensation-seeking during middle adolescence (Nelson et al., 2005; Nelson et al., 2016; Shulman et al., 2016). Many of the outcomes associated with these processes are associated with underachievement in education (Becherer et al., 2021; Dimler & Natsuaki, 2015; Hallfors et al., 2002; Ullsperger & Nikolas, 2017). These theories are, however, limited in their ability to explain the broader educational changes observed during adolescence and, as highlighted by Pfeifer and Berkman (2018), they do not account for self-identity formation, a key developmental task associated with adolescence. Pfeifer and Berkman (2018) addresses this by proposing that during adolescence, behaviour is motivated by value-based choices; however, choices are weighted according to each adolescent's priorities, and these priorities may differ from the adult perspective on what should be prioritised. Adolescent decision-making and prioritisation, and how these decisions differ between groups, are areas for future research in the context of sex differences in educational achievement.

A Sex Differences of School Grades in Childhood and Adolescence. Supplementary Materials

A.1 School Grades: Mean and Variability.

Table A.1

Observed vs. Expected Observations Between Boys and Girls. Odds Ratios and Cohen's d at the Upper and Lower Tails in Mathematics by Age Group

Subject	Age	Sex	Observed Frequencies (vs. Expected Frequencies)							
			Lowest Level		Lower Level		Upper Level		Highest Level	
				% Cohort		% Cohort		% Cohort		% Cohort
Maths	7	F	20 (25)		156 (190)		661 (766)			
	TA		1.56	0.9%	**1.49**	6.5%	**1.46**	26.0%		
			(*d* = -0.25)		**(*d* = -0.22)**		**(*d* = -0.22)**			
		M	30 (25)		**218 (184)**		**844 (739)**		–	
	11	F	192 (174)		**311 (284)**		334 (415)		96 (130)	
			0.85	6.2%	**0.81**	9.7%	**1.60**	14.1%	**1.76**	4.4%
			(*d* = 0.09)		**(*d* = 0.12)**		**(*d* = -0.26)**		**(*d* = -0.31)**	
		M	161 (174)		247 (274)		**482 (401)**		**159 (125)**	
	16	F	156 (161)		347 (358)		345 (366)		106 (121)	
			1.07	5.6%	1.07	12.5%	1.14	12.7%	**1.31**	4.2%
			(*d* = -0.04)		(*d* = -0.04)		(*d* = -0.07)		**(*d* = -0.15)**	
		M	160 (155)		355 (344)		372 (351)		**132 (117)**	

Note. The percentage of cohort members, odds ratios (*OR*), and Cohen's *d* (in parentheses) are nested between male and female observed and expected frequencies for each subject/age group. Statistically significant overrepresentation is indicated in bold typeface. *OR* > 1 and *d* < 0 indicates more males achieving or being assessed at that level. Grade groupings by age: Age 7: lowest – Below 1; lower – Below 1, 1; upper – 3, 4. Age 11: lowest – 1, 2, 3c, 3b; lower – 1, 2, 3c, 3b, 3a; upper – 5a, 6; highest: 6. Age 16: lowest – U (unclassified), 1; lower – U, 1, 2; upper – 8, 9; highest – 9. Data are excluded (–) where the observed count of boys or girls < 10. TA indicates teacher-assessed grades.

Table A.2

Observed vs. Expected Observations Between Boys and Girls. Odds Ratios and Cohen's d at the Upper and Lower Tails in Reading, Writing, and English by Age Group.

Subject	Age	Sex	Observed Frequencies (vs. Expected Frequencies)							
			Lowest Level		Lower Level		Upper Level		Highest Level	
				% Cohort		% Cohort		% Cohort		% Cohort
English	7	F	17(35)		308(409)		**1019(862)**		-	
	TA		**3.15**	1.2%	**1.81**	13.9%	**0.66**	30.3%		
			(d=-0.63)		**(d=-0.33)**		**(d=0.23)**			
		M	**51(33)**		**496(395)**		735(892)		–	
	11	F	121(173)		181(252)		**355(289)**		-	
			1.94	5.9%	**1.91**	8.6%	**0.59**	9.9%		
			(d=-0.37)		**(d=-0.36)**		**(d=0.29)**			
		M	**218(166)**		**315(244)**		212(278)		-	
	16	F	31(75)		153(262)		**338(261)**		**114(82)**	
			4.02	2.6%	**2.68**	9.1%	**0.50**	9.0%	**0.41**	2.8%
			(d=-0.77)		**(d=-0.54)**		**(d=0.38)**		**(d=0.49)**	
		M	**115(71)**		**359(250)**		171(248)		46(78)	
Reading	7	F	20(42)		232(316)		**1004(889)**		-	
	TA		**3.26**	1.4%	**1.85**	10.7%	**0.68**	30.1%		
			(d=-0.65)		**(d=-0.34)**		**(d=0.21)**			
		M	**62(40)**		**388(304)**		742(857)		-	
	11	F	125(173)		179 236)		**297(246)**		-	
			1.86	6.0%	**1.72**	8.1%	**0.62**	8.6%		
			(d=-0.34)		**(d=-0.29)**		**(d=-0.26)**			
		M	**216(168)**		**285(228)**		186(237)		-	
Writing	7	F	33(62)		306(414)		**523(432)**		-	
	TA		**2.85**	2.1%	**1.87**	14.0%	**0.60**	14.6%		
			(d=-0.58)		**(d=-0.35)**		**(d=0.28)**			
		M	**89(60)**		**507(399)**		325(416)		-	
	11	F	29(42)		261(373)		**1181(990)**		**51(39)**	
	TA		**1.94**	1.4%	**2.04**	12.7%	**0.55**	33.6%	**0.52**	1.3%
			(d=-0.37)		**(d=-0.39)**		**(d=0.33)**		**(d=0.36)**	
		M	**54(41)**		**472(360)**		766(957)		26(38)	

Note. The percentage of cohort members, odds ratios (*OR*), and Cohen's *d* (in parentheses) are nested between male and female observed and expected frequencies for each subject/age group. Statistically significant overrepresentation is indicated in bold typeface. *OR* > 1 and *d* < 0 indicates more males achieving or being assessed at that level. Grade groupings by age: Age 7: lowest – Below 1; lower – Below 1, 1; upper – 3, 4. Age 11: lowest – 1, 2, 3c, 3b; lower – 1, 2, 3c, 3b, 3a; upper – 5a, 6; highest: 6. Age 16: lowest – U (unclassified), 1; lower – U, 1, 2; upper – 8, 9; highest – 9. Data are excluded (–) where the observed count of boys or girls < 10. TA indicates teacher-assessed grades.

Table A.3

Observed vs. Expected Observations Between Boys and Girls. Odds Ratios and Cohen's d at the Upper and Lower Tails in Science by Age Group.

Subject	Age	Sex	Observed Frequencies (vs. Expected Frequencies)							
			Lowest Level		Lower Level		Upper Level		Highest Level	
				% Cohort		% Cohort		% Cohort		% Cohort
Science	7	F	12 (11)		171 (220)		722 (776)		–	
	TA		0.86	0.4%	**1.64**	7.5%	**1.21**	26.3%		
			d =-0.08)		***d* =-0.27)**		***d* =-0.11)**			
		M	10 (11)		**261 (212)**		**803 (749)**		–	
	11	F	18 (22)		224 (227)		1260 (1279)		–	
	TA		1.50	0.8%	1.03	7.7%	1.05	43.4%		
			d =-0.22)		*d* =-0.02)		*d* =-0.03)			
		M	26 (22)		222 (219)		1254 (1235)		–	
	16	F	125 (143)		330 (373)		**199 (173)**		24 (19)	
			1.32	5.0%	**1.33**	13.0%	**0.72**	6.0%	0.61	0.7%
			(d = -0.15)		***d* =-0.16)**		***d* =0.18)**		*d* =0.27)	
		M	**154 (136)**		**398 (355)**		140 (166)		14 (19)	
Additional Science	16	F	**91 (111)**		273 (308)		**261 (227)**		51 (43)	
			1.49	5.7%	**1.31**	15.8%	**0.66**	11.6%	0.62	2.2%
			***d* =-0.22)**		**(d =-0.15)**		***d* =0.23)**		*d* =0.26)	
		M	**123 (103)**		**321 (286)**		176 (210)		32 (40)	
Biology	16	F	–		–		**398 (370)**		139 (127)	
							0.75	12.8%	0.80	4.4%
							***d* =0.15)**		*d* =0.12)	
		M	–		–		321 (349)		108 (120)	
Chemistry	16	F	–		–		385 (362)		139 (134)	
							0.79	12.6%	0.91	4.7%
							d =0.13)		*d* =0.05)	
		M	–		–		325 (348)		124 (129)	
Physics	16	F	–		–		345 (356)		137 (144)	
							1.12	12.4%	1.12	5.0%
							d =-0.06)		*d* =-0.06)	
		M	–		–		349 (338)		143 (136)	
Average Science Grade	16	F	142 (171)		274 (312)		**454 (423)**		155 (147)	
			1.43	6.0%	**1.33**	11.0%	**0.84**	14.8%	0.89	5.2%
			***d* =-0.20)**		***d* =-0.16)**		***d* =0.10)**		*d* =0.06)	
		M	**191 (163)**		**336 (298)**		372 (403)		132 (140)	

Note. The percentage of cohort members, odds ratios (*OR*), and Cohen's *d* (in parentheses) are nested between male and female observed and expected frequencies for each subject/age group. 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. Statistically significant overrepresentation is indicated in bold typeface. *OR* > 1 and *d* < 0 indicates more males achieving or being assessed at that level. Grade groupings by age: Age 7: lowest – Below 1; lower – Below 1, 1; upper – 3, 4. Age 11: lowest – 1, 2, 3c, 3b; lower – 1, 2, 3c, 3b, 3a; upper – 5a, 6; highest: 6. Age 16: lowest - F, G, U (unclassified); lower - E, F, G, U; upper - A, A*, highest - A*. Data are excluded (–) where the observed count of boys or girls < 10. TA indicates teacher-assessed grades.

Table A.4

Observed vs. Expected Observations Between Boys and Girls. Odds Ratios and Cohen's d at the Upper and Lower Tails in Optional Subjects at Age 16. STEM, Humanities, and Modern Foreign Languages.

Subject	Sex	Observed Frequencies (vs. Expected Frequencies)							
		Lowest Level		Lower Level		Upper Level		Highest Level	
			% Cohort		% Cohort		% Cohort		% Cohort
Computing	F	11 (22)		16 (33)		**52 (37)**		**16 (9)**	
		2.43	13.6%	**2.79**	21.0%	**0.54**	23.5%	**0.51**	5.7%
		***d* =-0.48)**		***d* =-0.57)**		***d* =0.33)**		***d* =0.37)**	
	M	**90 (79)**		**140 (123)**		122 (137)		26 (33)	
Geography	F	77 (94)		159 (199)		**354 (303)**		**112 (97)**	
		1.48	7.9%	**1.60**	16.5%	**0.65**	25.2%	**0.73**	8.1%
		(d =-0.21)		***d* =-0.26)**		***d* =0.23)**		***d* =0.17)**	
	M	**120 (103)**		**256 (216)**		278 (329)		91 (106)	
Social Sciences	F	25 (29)		59 (65)		**120 (101)**		–	
		1.50	7.0%	1.37	15.7%	**0.42**	24.4%		
		d =0.22)		*d* =0.17)		***d* =0.48)**			
	M	17 (13)		36 (30)		28 (47)		–	
French	F	25 (33)		72 (90)		**239 (209)**		**102 (91)**	
		1.66	3.7%	**1.70**	10.0%	**0.63**	23.2%	**0.71**	10.1%
		***d* =0.28)**		***d* =0.29)**		***d* =0.25)**		***d* =0.19)**	
	M	**32 (23)**		**82 (64)**		118 (148)		53 (64)	
German	F	–		11 (19)		66 (61)		21 (19)	
				2.69	7.2%	0.79	23.5%	0.81	7.4%
				***d* =-0.54)**		*d* =0.12)		*d* =0.11)	
	M	–		**24 (16)**		48 (53)		15 (17)	
Spanish	F	16 (20)		38 (54)		**163 (142)**		**73 (61)**	
		1.71	3.9%	**2.17**	10.3%	**0.60**	27.3%	**0.56**	11.7%
		d =-0.30)		***d* =-0.43)**		***d* =0.28)**		***d* =0.32)**	
	M	19 (15)		**54 (38)**		80 (101)		31 (43)	
History	F	118 (140)		222 (258)		**465 (410)**		**165 (135)**	
		1.46	9.6%	**1.43**	17.7%	**0.67**	28.1%	**0.58**	9.3%
		***d* =-0.21)**		***d* =-0.20)**		***d* =0.22)**		***d* =0.30)**	
	M	**144 (122)**		**260 (224)**		301 (356)		87 (117)	
Religious Studies	F	100 (159)		207 (295)		**660 (537)**		**227 (178)**	
		2.55	9.0%	**2.22**	16.6%	0.49	30.2%	0.49	10.0%
		***d* =-0.52)**		***d* =-0.44)**		***d* =0.39)**		***d* =0.39)**	
	M	**198 (139)**		**345 (257)**		343 (466)		106 (155)	

Note. The percentage of cohort members, odds ratios (*OR*), and Cohen's *d* (in parentheses) are nested between male and female observed and expected frequencies for each subject/age group. Statistically significant overrepresentation is indicated in bold typeface. *OR* > 1 and *d* < 0 indicates more males achieving or being assessed at that level. Grade groupings: lowest - F, G, U (unclassified); lower - E, F, G, U; upper - A, A*, highest - A*. Data are excluded (–) where the observed count of boys or girls < 10. TA indicates teacher-assessed grades.

Table A.5

Observed vs. Expected Observations Between Boys and Girls. Odds Ratios and Cohen's d at the Upper and Lower Tails in Optional Subjects at Age 16. Applied Subjects, Sports, and the Arts.

Subject	Sex	Observed Frequencies (vs. Expected Frequencies)							
		Lowest Level		Lower Level		Upper Level		Highest Level	
			% Cohort		% Cohort		% Cohort		% Cohort
Design & Technology	F	27 (56)		68 (132)		**206 (132)**		**70 (42)**	
		3.54	7.7%	**1.89**	18.2%	**0.33**	18.1%	**0.32**	5.8%
		(*d* = -0.70)		**(*d* = -0.35)**		**(*d* = 0.61)**		**(*d* = 0.63)**	
	M	**103 (74)**		**239 (175)**		99 (173)		28 (56)	
Film, TV, Media & Office	F	28 (46)		55 (84)		**130 (102)**		**39 (26)**	
		2.33	8.7%	**2.18**	15.9%	**0.54**	19.1%	**0.37**	4.8%
		(*d* = 0.34)		**(*d* = -0.47)**		**(*d* = -0.43)**		**(*d* = 0.55)**	
	M	**77 (59)**		**136 (107)**		100 (128)		19 (32)	
Physical Education / Sports Studies	F	–		31 (44)		**124 (88)**		41 (33)	
				1.73	10.1%	**0.45**	20.3%	0.67	7.6%
				(*d* = -0.30)		**(*d* = 0.44)**		(*d* = 0.22)	
	M	–		**84 (71)**		**108 (144)**		46 (54)	
Art & Design	F	23 (35)		69 (98)		**312 (264)**		**130 (106)**	
		2.77	3.1%	**2.65**	8.8%	**0.41**	23.8%	**0.37**	9.6%
		(*d* = -0.56)		**(*d* = -0.54)**		**(*d* = 0.49)**		**(*d* = 0.55)**	
	M	**27 (15)**		**72 (43)**		68 (116)		23 (47)	
Performing Arts	F	19 (24)		57 (69)		227 (216)		**59 (54)**	
		1.72	3.0%	**1.62**	8.7%	0.84	27.3%	**0.75**	6.8%
		(*d* = 0.30)		**(*d* = -0.26)**		(*d* = 0.10)		**(*d* = 0.16)**	
	M	20 (15)		**55 (43)**		124 (135)		28 (33)	

Note. The percentage of cohort members, odds ratios (*OR*), and Cohen's *d* (in parentheses) are nested between male and female observed and expected frequencies for each subject/age group. Statistically significant overrepresentation is indicated in bold typeface. *OR* > 1 and *d* < 0 indicates more males achieving or being assessed at that level. Grade groupings: lowest - F, G, U (unclassified); lower - E, F, G, U; upper - A, A*, highest - A*. Data are excluded (–) where the observed count of boys or girls < 10. TA indicates teacher-assessed grades. Subject groups: 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.

Table A.6

Standardised Means, Standard Deviations, and Percentages of Boys and Girls Observed at the Upper and Lower Tails of the Grade Distributions in for Age 16 Subjects with Insufficient Statistical Power or Combined Due to Skewed Distributions.

Subject	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	
Other Mathematics	F	221	**0.06**		0.99	–	10.4%	33.5%	14.2%
			0.11 [-0.06, 0.28]	0.34	.518		.932	.575	.284
	M	299	-0.05		1.01	–	11.0%	30.7%	10.7%
Economics	F	27	-0.18		1.24	–	–	–	–
			-0.25 [-0.69, 0.20]	0.004	.150				
	M	69	0.07		0.89	–	–	–	–
Engineering	F	14	0.29		1.11	–	–	–	–
			0.35 [-0.23, 0.93]	0.37	.925				
	M	66	-0.06		0.97	–	–	–	–
Design & Technology (F)	F	451	**0.16**		0.95	3.5%	8.7%	–	–
			0.75 [0.55, 0.95]	1.00	.244	.013	< .001		
	M	128	-0.56		0.98	9.4%	29.7%	–	–
Design & Technology (M)	F	277	**0.44**		0.93	3.6%	8.7%	**27.7%**	**9.5%**
			0.61 [0.47, 0.75]	1.00	.140	.001	< .001	< .001	< .001
	M	831	-0.15		0.98	**9.4%**	**29.2%**	10.8%	2.3%

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. P indicates achieved power (Champely, 2020). Data are excluded (–) where the observed count of boys or girls < 10. Grade groupings: lowest - F, G, U (unclassified); lower - E, F, G, U; upper - A, A*, highest - A*. Data are excluded (–) where the observed count of boys or girls < 10. TA indicates teacher-assessed grades. Subject groups: Other Mathematics – Application of Mathematics, Mathematics (sat a year earlier than would be usual for these cohort members, under the prior A*-G grading system), Methods in Mathematics, Statistics. Design & Technology (F) - Food Technology, Textiles Technology. Design & Technology (M) - Electronic Products, Product Design, Resistant Materials, Systems & Control, Graphic Products, Design & Technology.

Table A.7

Observed vs. Expected Observations Between Boys and Girls. Odds Ratios and Cohen's d at the Upper and Lower Tails for Age 16 Subjects with Insufficient Statistical Power or Combined Due to Skewed Distributions.

Subject	Sex	Observed Frequencies (vs. Expected Frequencies)							
		Lowest Level		Lower Level		Upper Level		Highest Level	
			% Cohort		% Cohort		% Cohort		% Cohort
Other	F	–		23 (24)		74 (71)		31 (27)	
Mathematics				1.07	10.7%	0.88	31.9%	0.72	12.2%
				(*d* = -0.11)		(*d* = 0.08)		(*d* = 19.7)	
	M	–		33 (32)		92 (95)		32 (36)	
Design &	F	16 (22)		39 (60)		–		–	
Technology		**2.81**	4.8%	**4.46**	13.3%				
(F)		**(*d* = -0.57)**		**(*d* = -0.82)**					
	M	**12 (6)**		**38 (17)**		–		–	
Design &	F	11 (26)		29 (58)		**81 (43)**		**27 (13)**	
Technology		**2.97**	9.2%	**2.73**	20.8%	**0.29**	15.4%	**0.29**	4.7%
(M)		**(*d* = -0.60)**		**(*d* = -0.55)**		**(*d* = 0.68)**		**(*d* = 0.68)**	
	M	**91 (76)**		**201 (172)**		90 (128)		25 (39)	

Note. The percentage of cohort members, odds ratios (*OR*), and Cohen's *d* (in parentheses) are nested between male and female observed and expected frequencies for each subject/age group. Statistically significant overrepresentation is indicated in bold typeface. *OR* > 1 and $d < 0$ indicates more males achieving or being assessed at that level. Grade groupings: lowest - F, G, U (unclassified); lower - E, F, G, U; upper - A, A*, highest - A*. Data are excluded (–) where the observed count of boys or girls < 10. TA indicates teacher-assessed grades. Subject groups are detailed in Table A.6.

A.2 Alternate and Other Educational Outcomes

Table A.8

Percentage Point Achievement Gap Between Boys and Girls Meeting or Exceeding the Benchmark Standards at each Age: National Results vs. Sample Cohort Members.

Subject	Achieved Benchmark		Exceeded Benchmark	
	National	Sample	National	Sample
		Age 7		
English	9.5%	7.0%	NP	8.8%
Mathematics	2.9%	2.4%	NP	-7.3%
Science	3.2%	3.4%	NP	-3.7%
		Age 11		
English	9%	5.0%	10%	11.8%
Mathematics	0%	1.8%	6%	-7.9%
Science	1%	0.2%	NP	-1.3%
		Age 16		
English	16%	13.2%	NP	10.1%
Mathematics	1%	1.1%	NP	-1.4%

Note. NP – data not published. National assessment results at age 7, 11, and 16 years old (Department for Education, 2010, 2012, 2017). Exceeded benchmark at age 16: achieved grade 7, 8 or 9 in mathematics or English. Achieved the benchmark at age 16: achieved grades 4 – 9 in mathematics or English. All comparisons are the percentage of female members minus the percentage of male members of the national cohort or sample: negative values indicate a male advantage.

A.3 Comparing teacher-assessed grades and test outcomes.

The age 11 in mathematics and English, as analysed in the main text (see Table 2) and detailed in tables SM.A.1 - 2 above, are national, standardised tests, as are all the age 16 educational outcomes. Efforts have been made to minimise the impact of possible bias in these assessments through anonymised, electronic marking. Each paper is marked by multiple examiners; individual examiners only mark certain questions and do not see the entire paper. The age 7 educational assessments are teacher-assessed against a detailed national framework. Science and writing at age 11 are also teacher-assessed. Under England's education system, each cohort member has two reported outcomes in mathematics and English at age 11 – the national standard tests and teacher-assessed grades. This dual assessment offers the opportunity to examine teacher bias on educational outcomes. Prior research points to potential bias in teacher assessments in this cohort, comparing subjective teacher assessments of ability with outcomes from cognitive ability tests, and in the analysis of age 11 assessments of earlier cohorts (Burgess & Greaves, 2013; Campbell, 2015; Hansen & Jones, 2011). Burgess and Greaves (2013) reported evidence of bias in favour of and against students from ethnic minority groups in the age 11 school assessments we have analysed here. Campbell (2015) reported evidence of bias in the MCS, based on family income, sex, SEN status, and ethnicity at age 7, by comparing the British Ability Scale (BAS) test results and teacher subjective judgements. Hansen and Jones (2011) report larger gender gaps favouring females in teacher assessments, using the Foundation Stage Profile, a school-based assessment of development and learning, than in BAS tests at age 5. Teacher bias is also found in experimental research and other reports from other countries/ education systems (Harlen, 2005; Malouff & Thorsteinsson, 2016; Terrier, 2020). We examined how the teacher-assessed grades compared with test outcomes and therefore influence the results reported here (for details, Tables A.9, A.10).

We find that grade variability differs between age 11 teacher-assessed grades and test grades, and that mean sex differences do not significantly differ. There is less dispersion in teacher-assessed English grades - many more cohort members achieved the lowest test grades in English than were teacher-assessed at that level, and more students were assessed at the highest grades than achieved these grades in the test. Our findings of less dispersion in English grades are consistent with prior reports (Burgess & Greaves, 2013). There was slightly more dispersion in teacher-assessed mathematics grades, which is inconsistent with the prior analysis (Burgess & Greaves, 2013). Like English, more students achieved the lowest grades in mathematics than were assessed at that level by their teachers. However, there was closer

Table A.9

Standardised Mean, Standard Deviation, and Observed Percentages at the Upper and Lower Tails of the Grade Distributions for Age 11 Teacher-Assessed English and Mathematics Grades.

Subject	Sex	N	M	SD	Observed Percentages		
					Lowest Level	Lower Level	Highest Level
English	F	2946	**0.13**	0.96	0.9%	6.4%	1.6%
(TA)			0.27 [0.22, 0.32]	< .001	.020	< .001	.008
	M	2846	-0.14	**1.02**	**1.6%**	**11.1%**	0.8%
Mathematics	F	2948	-0.06	0.97	1.1%	9.1%	3.5%
(TA)			-0.12 [-0.17, -0.07]	< .001	.387	1.000	< .001
	M	2846	**0.06**	**1.03**	1.4%	9.1%	**6.0%**

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: lowest – 2, 1; lower – 3, 2, 1; highest - 6. *OR* > 1 and *d* < 0 indicates more males achieving or being assessed at that level. Data are excluded (–) where the observed count of boys or girls < 10.

alignment of teacher-assessed grades and test grades at the highest achievement level in mathematics. Changes in the grading structure implemented for this cohort - grade 6 achievement was not available at age 11 in the prior analysis – will account for some differences. Among the most able students in English, a small percentage of students were assessed at the highest achievement level, but fewer than 10 cohort members achieved this grade in the test, for example.

We compared the percentage of boys and girls assessed at each grade in the test vs. the percentage teacher-assessed at that grade (for details, see Supplementary Materials, Figure A.1). In general, the percentage of students whose teacher-assessed and test grades differed in mathematics was less than 1.5 percentage points (pp.) at each grade level, and most differences were less than 1 pp. The larger of these was 1.3 pp. of girls overassessed, and 1.4% pp. of boys underassessed, at grade 5 in mathematics. There were larger differences between teacher-assessed grades and test grades in English, particularly at grades 4 and 5. More boys and girls achieved a grade 5 (3 pp. each) in the test than were teacher-assessed at that level. For boys, most were likely underassessed at grade 4 (2.3 pp.) and the remainder overassessed

Table A.10

Observed vs. Expected Observations at the Upper and Lower Tails of the Grade Distributions for Age 11 Teacher-Assessed Grades.

Subject	Sex	Observed (vs. Expected) Frequencies					
		Lowest Level		Lower Level		Highest Level	
			% Cohort		% Cohort		% Cohort
English	F	25 (35)		189 (257)		**46 (35)**	
(TA)		**1.83**	1.1%	**1.82**	8.5%	**0.49**	1.2%
		***d* =-0.33)**		***d* =-0.33)**		***d* =0.39)**	
	M	**44 (34)**		**316 (248)**		22 (33)	
Mathematics	F	32 (36)		268 (268)		102 (138)	
(TA)		1.27	1.2%	1.00	9.1%	**1.76**	4.7%
		d =-0.13)		*d* =0.00)		***d* =0.32)**	
	M	39 (35)		258 (258)		**169 (133)**	

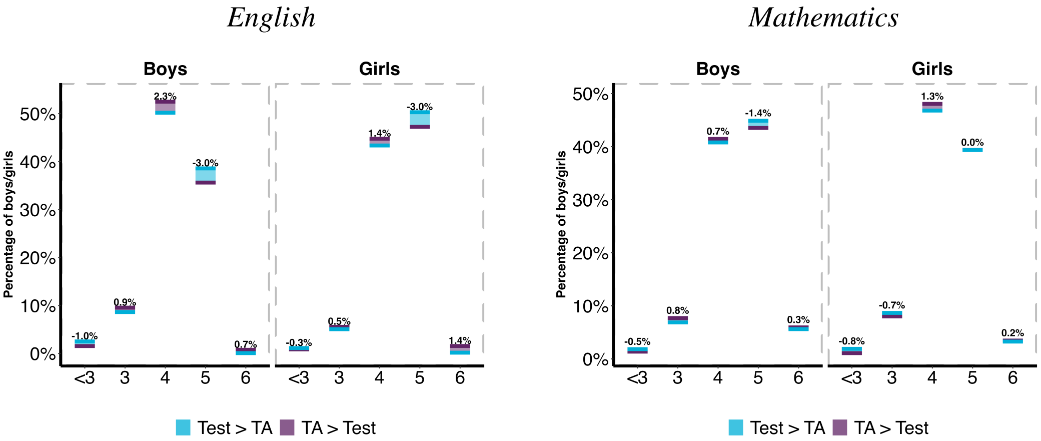
Note. The percentage of cohort members, odds ratios (*OR*), and Cohen's *d* (in parentheses) are nested between male and female observed and expected frequencies for each subject/age group. Statistically significant overrepresentation is indicated in bold typeface. *OR* > 1 and *d* < 1 indicates more males achieving or being assessed at that level. Data are excluded (–) where the observed count of boys or girls < 10.

at grade 6, although this is based on observation only; we have not tested by individual, only by group/sex. On the other hand, equal proportions of girls were under- or overassessed in grades 4 and 6. These patterns of greater teacher-assessed vs test grades differences in English, particularly more underassessment, more accuracy in mathematics with similar percentages of under and over-assessment, are in line with prior reports (Burgess & Greaves, 2013).

Our results may offer some support for a teacher gender bias account; however, the differences at a group level are very small. Teachers assessed more boys at grades 4 and below in mathematics, whereas more girls achieved these grades in the tests. More girls were over-assessed at grade 6 and more boys were underassessed at grade 4 in English. In mathematics, more girls were overassessed at grade 4, and more boys were underassessed at grade 5. Once tests were introduced at age 11, proportionally more girls were assessed at the lowest achievement level, and the effect size of male overrepresentation reduced in English and reading. Together, these could be interpreted as teacher grades favouring girls; however, in

Figure A.1

Difference in the Percentage of Boys and Girls Assessed at each Grade in English and Mathematics.



each of these, the proportion of students is very small (Gortazar et al., 2022; Hansen & Jones, 2011). There is more under and overassessment in English than in mathematics, and proportionally more girls achieved grade 5 in the test than were teacher-assessed at grade 5. However, test outcomes may naturally, through additional support or tutoring, or through test anxiety, be lower or higher than engagement or achievement in class as represented by teacher-assessed grades. For example, in writing (see Table 3.2), which is teacher-assessed in both age groups, there was a reduction in the percentage and effect size of male overrepresentation at the lowest achievement level, which may indicate a perception of improvement, more so for boys, which would not support a teacher bias account. It is also possible that test scores may be boosted by teaching to the test, or that the differences we find result from other stronger teacher biases that have been reported, such as SEN status, receiving free school meals, or ethnicity (Burgess & Greaves, 2013). Some research has questioned the impact of teacher bias on outcomes, and yet others have reported long-term impacts on school achievement and access to higher-paying career options (Burgess et al., 2022; Carlana, 2019; Lavy & Megalokonomou, 2019; Lavy & Sand, 2018; Terrier, 2020). Teacher bias effects may be more prevalent in less effective teachers and can benefit boys and girls depending on which gender is favoured by an individual teacher (Lavy & Megalokonomou, 2019). Further research is required to clarify when and how the discrepancies between teacher assessment and test grades influence outcomes.

A.4 Age 7 Science Assessments

In the main text, we have analysed sex differences in science at age 7, which is teacher-assessed. In addition to an overall science assessment, students are separately assessed on four strands of the national curriculum in this age group, and outcomes for these are available for analysis. Here we provide the same analyses as reported in the main text, for these four separate assessments – Life and Living Processes, Materials and their Properties, Physical Processes, and Scientific Enquiry - that underlie the overall science grade in this age group (Department for Education, 2015). The content of these strands of the science curriculum are:

Life and Living Processes. Identifying and naming a variety of common plants and animals, including humans, and being able to identify and describe the structure and basic parts of these. Understanding of differing habitats and how plants and animals obtain food and nutrients.

Materials and Their Properties. Distinguishing between objects and the materials they are made from. Being able to identify everyday materials and describe their physical properties and their uses.

Physical Processes. Describe and understand changes in the environment and understand the natural processes that cause these.

Scientific Enquiry. Practical scientific skills, methods and processes, including observation, using simple equipment, performing simple tests, gathering, and recording data.

In the main text, we report no sex differences in science at age 7. Underlying the science assessment, we also find no sex differences in Life and Living Processes, Materials and their Properties, and Scientific Enquiry. Our sample is large enough to find a negligible, significant difference in Physical Processes, favouring boys. These results partially align with prior reports in UK students, except for the significant differences found in the Physical Processes domain (Haworth et al., 2010). Like science, we find that boys are significantly overrepresented in the upper and lower tails of Materials and their Properties, Physical Processes, and Scientific Enquiry, and in the lower tail only in Life and Living Processes.

Table A.11

Standardised Mean, Standard Deviation, and the Percentage of Boys and Girls Observed at the Upper and Lower Tails of the Grade Distributions for Age 7 Assessments in Science components.

Subject	Sex	*N*	*M*	*SD*	Observed Percentages		
					Lowest Level	Lower Level	Highest Level
Life & Living Processes	F	2949	0.01	0.96	–	5.0%	28.5%
			0.03 [-0.02, 0.08]	.003	–	< .001	.356
	M	2846	-0.01	**1.04**	–	**7.7%**	29.7%
Materials & their Properties	F	2948	0.001	0.95	0.4%	6.4%	23.8%
			-0.002 [-0.05, 0.05]	< .001	.624	<.001	.016
	M	2845	-0.001	**1.05**	0.5%	**9.2%**	26.6%
Physical Processes	F	2948	-0.03	0.94	0.4%	7.4%	22.3%
			-0.05 [-0.11, -0.01]	< .001	.907	< .001	< .001
	M	2845	**0.03**	**1.05**	0.4%	**10.0%**	**28.0%**
Scientific Enquiry	F	2949	-0.01	0.95	0.5%	7.8%	22.2%
			0.008 [-0.04, 0.06]	< .001	1.000	< .001	.001
	M	2846	0.01	**1.05**	0.5%	**11.1%**	**26.0%**

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: lowest – Below 1; lower – Below 1, 1; upper – 3, 4. Data are excluded (–) where the observed count of boys or girls < 10.

Table A.12

Observed vs. Expected Frequencies of Boys and Girls Observed at the Upper and Lower Tails of the Grade Distributions for Individual Components of the Age 7 Science Assessments.

Subject	Sex	Observed (vs. Expected) Frequencies					
		Lowest Level		Lower Level		Highest Level	
			% Cohort		% Cohort		% Cohort
Life & Living Processes	F	–		147 (186)		840 (856)	
				1.58	6.3%	1.06	29.0%
				(*d* = -0.25)		(*d* = -0.03)	
	M	–		**218 (179)**		843 (827)	
Materials & their properties	F	11 (13)		189 (230)		702 (742)	
		1.32	0.4%	**1.46**	7.8%	**1.16**	25.2%
		(*d* = -0.15)		**(*d* = -0.21)**		**(*d* = -0.08)**	
	M	14 (12)		**262 (230)**		**757 (717)**	
Physical Processes	F	13 (12)		217 (235)		658 (740)	
		0.88	0.4%	**1.38**	8.6%	**1.35**	25.1%
		(*d* = 0.07)		**(*d* = -0.18)**		**(*d* = -0.17)**	
	M	11 (12)		**285 (247)**		**796 (714)**	
Scientific Enquiry	F	14 (14)		231 (278)		654 (708)	
		0.96	0.5%	**1.47**	9.4%	**1.23**	24.0%
		(*d* = 0.02)		**(*d* = -0.21)**		**(*d* = -0.11)**	
	M	13 (13)		**316 (269)**		**736 (684)**	

Note. The percentage of cohort members, odds ratios (OR), and Cohen's *d* (in parentheses) are nested between male and female observed and expected frequencies for each subject. Statistically significant overrepresentation is indicated in bold typeface. OR > 1 and $d < 0$ indicates more males achieving or being assessed at that level. Grade: lowest – Below 1; lower – Below 1, 1; upper – 3, 4. Data are excluded (–) where the observed count of boys or girls < 10.

B Sex Differences of School Grades in Childhood and Adolescence. Supplementary Materials 2

B.1 Testing the Millennium Cohort Study Sampling Weights

The Millennium Cohort Study (MCS) provides sampling weights to adjust for attrition between the current sample and the original sample of cohort members recruited at nine months old (see Chapter 2 for further discussion). The sampling weights also adjust for the sampling approach - clustered, random sampling by electoral wards, as described in the Methods section for this study. These weights aim to remove bias in analytical models, particularly when aiming to calculate population coefficients and make inferences about the population of interest (UCL Centre for Longitudinal Studies, 2020).

As discussed in Chapter 2, the sampling weights are not applied to the Study 1 analyses. This decision was made because the Study 1 sample is a sub-sample of the Age 14 sample, for whom educational achievement data were available. Sampling weights are provided for the Age 14 sample but not for the educational data. The educational data were extracted from the NPD based on parental permission given in the Age 7 survey, two waves before the Age 14 wave. As detailed in Chapter 2, parental permission was given for 60% of the Age 7 sample. The educational data was successfully matched for this thesis for 57% cohort members present in the Age 14 sample.

The demographics of the Age 7 and Age 14 samples will differ due to attrition between waves (UCL Centre for Longitudinal Studies, 2020). The educational data is more closely related to the Age 7 sample than the Age 14 sample. As demonstrated in Chapter 2, the Age 14 sampling weights provided when applied to these data do not achieve the intention of weighting the sample back to the demographic profile of the weighted full sample - those cohort members recruited and surveyed in the nine-month-old wave. Extending the demographics comparisons of Chapter 2, this thesis section re-examines the Study 1 results for English, mathematics, and average science grade at Age 16 and English and mathematics at Age 11 to understand whether applying the sampling weights alters the conclusions drawn in the main study.

B.2 Demographics of the Weighted Sample

As described in Chapter 2, the weighted sample represented n_w = 5529 cohort members (n_w = 2594 boys and n_w = 2635 girls). Of these 83.7% of the sample represented White ethnic groups, 8.8% South Asian or South-East Asian ethnic groups (Pakistani: 4.2%, Indian: 2.3%, Bangladeshi 1.5%, Other Asian including Chinese 0.8%), and 3.2% from Black ethnic groups (Black African: 1.7%, Black Caribbean & Other Black: 1.5%). The remaining sample represented 3.4% Mixed and 0.8% Other ethnic groups. The weighted sample was negatively skewed towards the upper-income quintiles (Quintile 1: 14.6%, Quintile 2: 16.9%, Quintile 3: 19.9%, Quintile 4: 23.7%, Quintile 5: 25.0%). Demographics by sex are reported in Chapter 2.

B.3 Weighted Sample Analyses

The MCS weights can be applied and used in complex survey analyses, with detailed documentation provided on how to use the survey weights in SPSS or STATA, provided for researchers (Jones & Ketende, 2010; Ketende & Jones, 2011). Complex survey functionality is also available in the R environment (Lumley, 2010, 2024). The functionality allows for independent samples t-tests to be calculated with the survey weights applied, and also enables the calculation of weighted sample size, grouped as needed. Here, I used these R functions to apply the sampling weights to the sample analysed in the main, published body of Study 1 to understand whether the results differ under weighted analyses.

Table B.1

Weighted Sample: Standardised Means, Standard Deviations, and Percentages of Boys and Girls Observed at the Upper and Lower Tails of the Grade Distributions in Mathematics, English, and Average Science Grade by Age Group

Age Group Subject	Sex	N_w	M_w [CI] (Cohen's d_w [CI])	*SD*	Weighted Percentage of Girls/Boys (χ^2_w *p*)			
					Lowest Level	Lower Level	Upper Level	Highest Level
Age 11								
English	F	2667	0.11[0.07, 0.15]	0.96	4.4%	6.7%	**11.4%**	–
			0.29[0.23, 0.35]		**<.001**	**<.001**	**<.001**	
	M	2632	-0.18[-0.22, -0.13]	1.04	**8.2%**	**11.7%**	6.8%	–
Mathematics	F	2677	-0.10[-0.14, -0.07]	0.97	6.7%	10.9%	10.8%	3.0%
			-0.15[-0.21, -0.09]		.474	.065	**<.001**	**<.001**
	M	2643	0.05[0.01, 0.09]	1.02	6.2%	9.4%	**16.2%**	**5.0%**
Age 16								
English	F	2634	0.24[0.21, 0.28]	0.97	1.4%	6.4%	**10.9%**	**3.5%**
			0.40[0.34, 0.46]		**<.001**	**<.001**	**<.001**	**<.001**
	M	2574	-0.16[-0.20, -0.11]	0.99	**4.5%**	**14.5%**	5.6%	1.5%
Mathematics	F	2624	0.08[0.04, 0.12]	0.96	5.6%	12.9%	10.7%	3.1%
			-0.01[-0.07, 0.04]		.434	.365	.214	.070
	M	2585	0.09[0.05, 0.13]	1.00	6.2%	13.8%	11.9%	4.1%
Average	F	2495	0.02[-0.02, 0.06]	0.97	3.4%	10.6%	**18.2%**	**6.1%**
Science			0.17[0.10, 0.23]		**.001**	**.001**	**.001**	**.044**
Grade	M	2399	-0.14[-0.19, -0.10]	1.03	**5.3%**	**13.9%**	14.7%	4.7%

Note. Sex differences are nested between boys' and girls' weighted means (M_w [95% CI]), standard deviations and percentage representation at each achievement level: Cohen's d_w [95% CI] and Pearson chi-square (χ^2_w) test *p*-value for each achievement level. As the mean values are calculated from complex survey data with weights applied, confidence intervals (CIs) for group means are also provided. 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 Appendix A, Tables A.1 - A.3.

Sex differences in English, mathematics, and average science grades at Age 16 and English and mathematics at Age 11 were calculated using survey design-based independent sample t-tests. Mathematics, English and science are studied by all students at Age 11 and 16, however the Age 11 science grades are teacher-assessed, and as discussed in Appendix A, we found these to be less reliable than the standardised curriculum tests. I therefore do not apply the sample weights to the Age 11 science outcomes. Other Age 16 subjects are increasingly smaller sub-samples of the Study 1 sample. Often, the choice of studying each of these subjects is influenced by factors such as general achievement level (e.g., individual science subjects) and the school the cohort member attends - not all Age 16 subjects are offered by all schools, for example. I also examined sex differences at the upper and lower tails of the distribution by extracting the number of cohort members achieving each grade, using Chi-squared tests to test the observed number of boys and girls at the upper and lower achievement levels.

B.4 Weighted Results

B.4.1 Age 11 Weighted Results

Weighted mean sex differences in mathematics (d_w = -0.15) and English (d_w = 0.29) at Age 11 do not differ from those reported in the main study chapter (Chapter 3: English *d* = 0.29, mathematics *d* = -0.16). Effects sizes and odds ratios are very similar at the upper and lower tails of the distribution (see Tables B.1, B.2). However, the very small (*d* = 0.12) overrepresentation of girls at the lower achievement level in mathematics is significant in the larger Study 1 results but not significant (d_w = 0.10) in the smaller weighted sample. In summary, the weighted analyses support and confirm the conclusions in the larger unweighted sample. Still, the loss of statistical power at the distribution tails in the weighted sample has influenced the results.

B.4.2 Age 16 Weighted Results

Weighted mean sex differences in English (d_w = 0.40), mathematics (d_w= -0.01) and average science grade (d_w = 0.17) at Age 16 do not differ from those reported in the main survey chapter (Chapter 3: English *d* = 0.41, mathematics *d* = -0.01, average science grade *d* = 0.14). For the most part, the odd ratios at the upper and lower tails of the achievement

Table B.2

Weighted Sample: Weighted Observations vs. Expected Frequencies Between Boys and Girls. Odds Ratios and Cohen's d at the Upper and Lower Tails in English, Mathematics and Average Science Grade by Age Group

Subject	Sex	Weighted Observations (vs. Expected Frequencies)							
		Lowest Level		Lower Level		Upper Level		Highest Level	
			% Cohort		% Cohort		% Cohort		% Cohort
Age 11									
English	F	118 (169)		178 (245)		**303 (242)**		–	
		1.94	6.3%	**1.86**	9.2%	**0.57**	9.8%		
		(*d* = -0.37)		**(*d* = -0.34)**		**(*d* = 0.31)**			
	M	**217 (166)**		**309 (242)**		178 (239)		–	
Maths	F	179 (172)		292 (271)		288 (360)		80 (106)	
		0.92	6.4%	0.84	10.1%	**1.60**	13.4%	**1.70**	4.0%
		(*d* = 0.05)		(*d* = 0.10)		**(*d* = -0.26)**		**(*d* = -0.29)**	
	M	163 (170)		247 (268)		**427 (355)**		**131 (105)**	
Age 16									
English	F	38 (78)		168 (274)		**288 (218)**		**92 (66)**	
		3.22	3.0%	**2.50**	10.4%	**0.48**	8.3%	**0.41**	2.5%
		(*d* = -0.64)		**(*d* = -0.51)**		**(*d* = 0.40)**		**(*d* = 0.49)**	
	M	**116 (76)**		**374 (268)**		143 (213)		38 (64)	
Maths	F	148 (155)		338 (350)		282 (297)		82 (95)	
		1.11	5.9%	1.08	13.3%	1.12	11.3%	1.33	3.6%
		(*d* = -0.06)		(*d* = -0.04)		(*d* = -0.06)		(*d* = -0.16)	
	M	160 (153)		356 (344)		307 (292)		106 (93)	
Average Science Grade	F	84 (108)		265 (305)		**455 (412)**		**151 (135)**	
		1.60	4.3%	**1.36**	12.2%	**0.77**	16.5%	**0.77**	5.4%
		(*d* = -0.26)		**(*d* = -0.17)**		**(*d* = 0.14)**		**(*d* = 0.14)**	
	M	**127 (103)**		**333 (293)**		353 (396)		113 (129)	

Note. The weighted percentage of cohort members, odds ratios (OR_w), and Cohen's d_w (in parentheses) are nested between male and female observed and expected weighted frequencies for each subject group. Statistically significant overrepresentation is indicated in bold typeface. $OR_w > 1$ and $d_w < 0$ indicates more males achieving at that level. Grade groupings English & mathematics: lowest – U (unclassified), 1; lower – U, 1, 2; upper – 8, 9; highest – 9. Achievement level boundaries were calculated for the average science grade, average grade encompassing the 5th, 10th, 90th and 95th percentile boundary scores. Science is analysed at Age 16 only, as Age 11 Science is teacher-assessed and less reliable (see Appendix A for a detailed discussion).

distributions differ little (see Tables B.1, B.2). In mathematics, the results for the weighted sample (OR_w = 1.33, *d* = -0.16) are the same as reported for the main study in Appendix A (*OR* = 1.31, *d* = -0.15), but this difference is significant in the main sample and not significant in the weighted sample. While slightly more boys are reported at the lowest, lower or upper levels of mathematics achievement, these differences are negligible and not significant. In English, the results at the lower, upper and highest achievement levels are the same in both samples. There are some differences at the lowest achievement level (Study 1: *OR* = 4.02, *d* = -0.77; weighted sample: OR_w = 3.22, d_w = -0.64) - the weighted sample reduces the overrepresentation of boys and the lowest level of language achievement.

There are more differences in the comparisons of the composite average science grade between Study 1 and the weighted sample. The overrepresentation of girls at the highest achievement level across science subjects is significant in the weighted sample but not in the Study 1 sample (Study 1: *OR* = 0.89, *d* = 0.06; weighted sample: OR_w = 0.77, d_w = 0.14). Applying the weightings decreased the weighted count of boys at the lowest, upper, and highest achievement levels in science, whereas the weightings reduced the count of girls at the lower and lowest achievement levels, thereby increasing the sex differences at the upper, highest and lower achievement levels, and the mean difference.

B.5 Weighted Sample Discussion

The conclusions from the weighted sample support those from the Study 1 sample, except for a very small, significant overrepresentation of girls at the highest achievement level in science found in the weighted sample but not in the unweighted sample. There was also a smaller overrepresentation of boys at the lowest achievement level in English, from 4x more boys to 3.2x more boys. Weighting the sample has a more substantive effect on the relative distribution of girls and boys at the upper and lower tails in science than it does in mathematics and English. Weighting the sample results in a higher weighted count of White cohort members and reduces the counts for other ethnic groups, particularly South Asian groups. White students are less likely to go to university compared to all other ethnic minority groups in the UK, and boys are less likely to go to university than girls (Cavaglia et al., 2020;

Ethnicity Facts and Figures Service, 2023a). Given that the individual sciences tend to be studied more often by students planning to go to university in the future, ethnic background and being a boy may alter the choice of which science subjects will be studied at Age 16 and sex differences at the upper tail of the distribution (Department for Education, 2019; Ethnicity Facts and Figures Service, 2023a). However, despite the distribution of the combined science measure being more influenced by ethnic background, the weighted analyses support the overall conclusions reported in Study 1 (Chapter 3).

C Puberty and Age Associations with School Engagement. Supplementary Materials 1

C.1 School Engagement Confirmatory Factor Analysis

The initial face assessment of the available school engagement measures collected by the MCS indicated a bias towards emotional disengagement among the measures. Confirmatory factor analysis (CFA) was used to determine the most appropriate factor structure to be used for the subsequent main analyses. Models were compared using goodness of fit measures, comparative fit index (CFI), and root mean square error of approximation (RMSEA). An adequate model fit is indicated by CFI $\geq$ 0.90 and RMSEA $\leq$ .08, and excellent fit by CFI $\geq$ 0.95 and RMSEA $\leq$ .05 (Kline, 2016). Two-factor models were proposed utilising the measures (see Table 4.2) as follows:

- Model 1: Two-factor model of engagement and disengagement comprising 2 measures of engagement and 4 measures of disengagement.
- Model 2: Two-factor model of behavioural and emotional (dis)engagement comprising 2 measures of behavioural and 4 measures of emotional dis(engagement), with the disengagement measures reversed.

The CFI values for each model were promising (Age 11. Model 1 CFI = .92, Model 2 CFI = .95; Age 14. Model 1 CFI = .96, Model 2 CFI = .91), although the best fit according to CFI was Model 2 at Age 11 and Model 1 at Age 14. However, the RMSEA values for each model were poor, with each being > .08 (Age 11. Model 1 RMSEA = .13[.12, .14], Model 2 RMSEA = .11[.10, .12], Model 1 RMSEA = .11[.10, .11], Model 2 RMSEA = .14[.13, .14]). As a result, the main analyses tested the effects of puberty and chronological age on behavioural engagement, emotional engagement, and emotional disengagement, measured using a single item at each wave. Emotional disengagement was measured using three items at each wave, as described in Chapter 4, Table 4.2.

C.2 Main Results, Imputed Datasets

Table C.1

Bayesian Ordinal Logistic Regression Results for Girls and Boys: Behavioural Engagement.

	Median [95% HDPI]			
	Model 1		Model 2	
Measure	Male	Female	Male	Female
Age	-0.05[-0.13, 0.04]	**-0.15[-0.26, -0.04]**	-0.03[-0.12, 0.05]	**-0.12[-0.22, -0.01]**
Puberty	0.15[-0.01, 0.31]	0.02[-0.10, 0.14]	0.10[-0.05, 0.26]	0.03[-0.01, 0.14]
Puberty*Age	**-0.08[-0.12, -0.03]**	-0.04[-0.08, 0.00]	**-0.06[-0.11, -0.02]**	-0.04[-0.08, 0.00]
Covariates				
SEN Status	-0.03[-0.13, 0.07]	0.03[-0.08, 0.15]	0.01[-0.09, 0.11]	0.07[-0.04, 0.18]
Income	0.06[-0.04, 0.16]	0.05[-0.05, 0.14]	**0.11[0.01, 0.20]**	**0.09[0.00, 0.18]**
Prior Ach	0.07[-0.02, 0.16]	0.06[-0.01, 0.13]	**-0.19[-0.27, -0.11]**	**-0.19[-0.27, -0.12]**
Maths Self-concept			**0.34[0.18, 0.49]**	**0.55[0.43, 0.67]**
Maths Self-concept^2			**0.34[0.21, 0.48]**	0.11[0.01, 0.21]
Maths Self-concept^3			**-0.15[-0.25, -0.05]**	-0.06[-0.14, 0.01]
Language Self-concept			**0.42[0.27, 0.58]**	**0.53[0.38, 0.68]**
Language Self-concept^2			**0.23[0.10, 0.35]**	**0.15[0.02, 0.27]**
Educational Expectations			**0.16[0.11, 0.21]**	**0.22[0.17, 0.27]**

Note. Imputed sample size: $n = 2864$ girls, $n = 2749$ boys. Educational expectations, mathematics and language self-concept, and family income are time-varying. The remaining covariates are time-invariant. The median of the posterior distribution is reported with its 95% HDPI (highest density probability interval). Bold typeface highlights effects with a probability of direction > 95% that make a practically significant contribution to the outcome (Median ≥ 0.05). ^{2} quadratic and ^{3} cubic associations are reported where they make a non-negligible contribution in at least one of the two models for either boys or girls.

Table C.2

Bayesian Ordinal Logistic Regression Results for Girls and Boys: Behavioural Disengagement.

	Median [95% HDPI]			
	Model 1		Model 2	
Measure	Male	Female	Male	Female
Age	0.01[-0.03, 0.05]	**0.14[0.10, 0.17]**	-0.01[-0.05, 0.03]	**0.12[0.08, 0.16]**
Puberty	0.04[-0.05, 0.13]	**0.14[0.06, 0.22]**	0.05[-0.04, 0.14]	**0.13[0.05, 0.21]**
Covariates				
SEN Status	0.07[-0.04, 0.19]	-0.04[-0.17, 0.08]	0.04[-0.08, 0.15]	-0.07[-0.20, 0.05]
Income	**-0.22[-0.33, -0.11]**	**-0.27[-0.38, -0.17]**	**-0.24[-0.35, -0.14]**	**-0.30[-0.40, -0.19]**
Prior Ach	**-0.12[-0.20, -0.04]**	**-0.22[-0.30, -0.14]**	0.02[-0.06, 0.11]	-0.08[-0.16, 0.01]
Maths Self-concept			**-0.19[-0.35, -0.03]**	**-0.33[-0.46, -0.20]**
Maths Self-concept^2			**-0.15[-0.28, -0.01]**	0.02[-0.08, 0.13]
Maths Self-concept^3			**0.11[0.01, 0.22]**	**0.09[0.00, 0.17]**
Language Self-concept			**-0.37[-0.53, -0.21]**	**-0.24[-0.39, -0.07]**
Educational Expectations			**-0.13[-0.19, -0.08]**	**-0.16[-0.22, -0.11]**

Note. Imputed sample size: *n* = 2863 girls, *n* = 2748 boys. Educational expectations, mathematics and language self-concept, and family income are time-varying. The remaining covariates are time-invariant. The median of the posterior distribution is reported with its 95% HDPI (highest density probability interval). Bold typeface highlights effects with a probability of direction > 95% that make a practically significant contribution to the outcome (Median $\geq$ 0.05). ^{2} quadratic and ^{3} cubic associations are reported where they make a non-negligible contribution in at least one of the two models for either boys or girls.

Table C.3

Bayesian Ordinal Logistic Regression Results for Girls and Boys: Emotional Engagement.

	Median [95% HDPI]			
	Model 1		Model 2	
Measure	Male	Female	Male	Female
Age	**-0.16[-0.20, -0.13]**	**-0.26[-0.29, -0.23]**	**-0.12[-0.16, -0.09]**	**-0.23[-0.26, -0.19]**
Puberty	-0.02[-0.09, 0.06]	-0.07[-0.14, 0.00]	-0.03[-0.11, 0.04]	-0.07[-0.13, 0.00]
Covariates				
SEN Status	**-0.14[-0.24, -0.05]**	-0.03[-0.13, 0.08]	-0.07[-0.16, 0.02]	0.02[-0.08, 0.12]
Income	-0.05[-0.15, 0.04]	-0.05[-0.14, 0.04]	0.00[-0.09, 0.09]	0.00[-0.09, 0.08]
Income^2	**0.13[0.04, 0.21]**	**0.14[0.06, 0.23]**	0.06[-0.02, 0.14]	**0.08[0.01, 0.16]**
Prior Ach.	0.06[-0.01, 0.13]	**0.19[0.12, 0.26]**	**-0.24[-0.31, -0.17]**	**-0.11[-0.18, -0.04]**
Maths Self-concept			**0.35[0.21, 0.50]**	**0.67[0.56, 0.79]**
Maths Self-concept^2			**0.33[0.20, 0.45]**	0.06[-0.03, 0.15]
Maths Self-concept^3			**-0.20[-0.30, -0.11]**	**-0.10[-0.17, -0.03]**
Language Self-Concept			**0.54[0.40, 0.69]**	**0.63[0.49, 0.77]**
Language Self-Concept^2			**0.18[0.06, 0.29]**	0.07[-0.04, 0.19]
Language Self-Concept^3			**-0.12[-0.20, -0.04]**	-0.06[-0.14, 0.03]
Educational Expectations			**0.30[0.25, 0.34]**	**0.29[0.24, 0.34]**

Note. Imputed sample size: n = 2864 girls, n = 2749 boys. Educational expectations, mathematics and language self-concept, and family income are time-varying. The remaining covariates are time-invariant. The median of the posterior distribution is reported with its 95% HDPI (highest density probability interval). Bold typeface highlights effects with a probability of direction > 95% that make a practically significant contribution to the outcome (Median $\geq$ 0.05). ^{2} quadratic and ^{3} cubic associations are reported where they make a non-negligible contribution in at least one of the two models for either boys or girls.

Table C.4

Bayesian IRT Ordinal Logistic Regression Results for Girls and Boys: Emotional Disengagement.

	Median [95% HDPI]			
	Model 1		Model 2	
Measure	Male	Female	Male	Female
Age	0.01[-0.05, 0.07]	0.01[-0.09, 0.07]	0.00[-0.06, 0.06]	-0.04[-0.11, 0.04]
Puberty	0.04[-0.07, 0.16]	**-0.12[-0.20, -0.04]**	0.10[-0.02, 0.21]	**-0.13[-0.21, -0.05]**
Puberty*Age	0.01[-0.02, 0.03]	**0.08[0.05, 0.10]**	0.00[0.03, 0.03]	**0.08[0.05, 0.11]**
Covariates				
SEN Status	**0.19[0.11, 0.27]**	0.09[0.00, 0.19]	**0.15[0.07, 0.23]**	0.06[-0.03, 0.16]
Income	-0.03[-0.05, 0.00]	0.00[-0.03, 0.03]	-0.04[-0.07, -0.01]	-0.01[-0.04, 0.02]
Parent Ed	0.01[-0.02, 0.05]	-0.04[-0.08, 0.00]	0.05[0.02, 0.08]	-0.01[-0.05, 0.03]
Prior Ach	**-0.12[-0.18, -0.06]**	**-0.16[-0.22, -0.09]**	0.05[-0.02, 0.11]	0.01[-0.05, 0.08]
Maths Self-concept			**-0.23[-0.35, -0.11]**	**-0.44[-0.53, -0.35]**
Maths Self-concept^3			0.09[0.02, 0.16]	**0.12[0.07, 0.18]**
Language Self-concept			**-0.37[-0.50, -0.24]**	**-0.57[-0.69, -0.45]**
Language Self-concept^2			-0.05[-0.14, 0.05]	**0.16[0.06, 0.25]**
Language Self-concept^3			**0.10[0.03, 0.16]**	-0.03[-0.10, 0.04]
Educational Expectations			**-0.19[-0.22, -0.15]**	**-0.15[-0.19, -0.12]**

Note. Sample size: n = 2228 girls, n = 2257 boys. Educational expectations, mathematics and language self-concept, and family income are time-varying. The remaining covariates are time-invariant. The median of the posterior distribution is reported with its 95% HDPI (highest density probability interval). Bold typeface highlights effects with a probability of direction > 95% that make a practically significant contribution to the outcome (Median $\geq$ 0.05). ^{2} quadratic and ^{3} cubic associations are reported where they make a non-negligible contribution in at least one of the two models for either boys or girls.

C.3 Descriptive Statistics for SEN vs. TD Students

Table C.5

Descriptive Statistics for TD and SEN Students. School Engagement and Self-Concepts

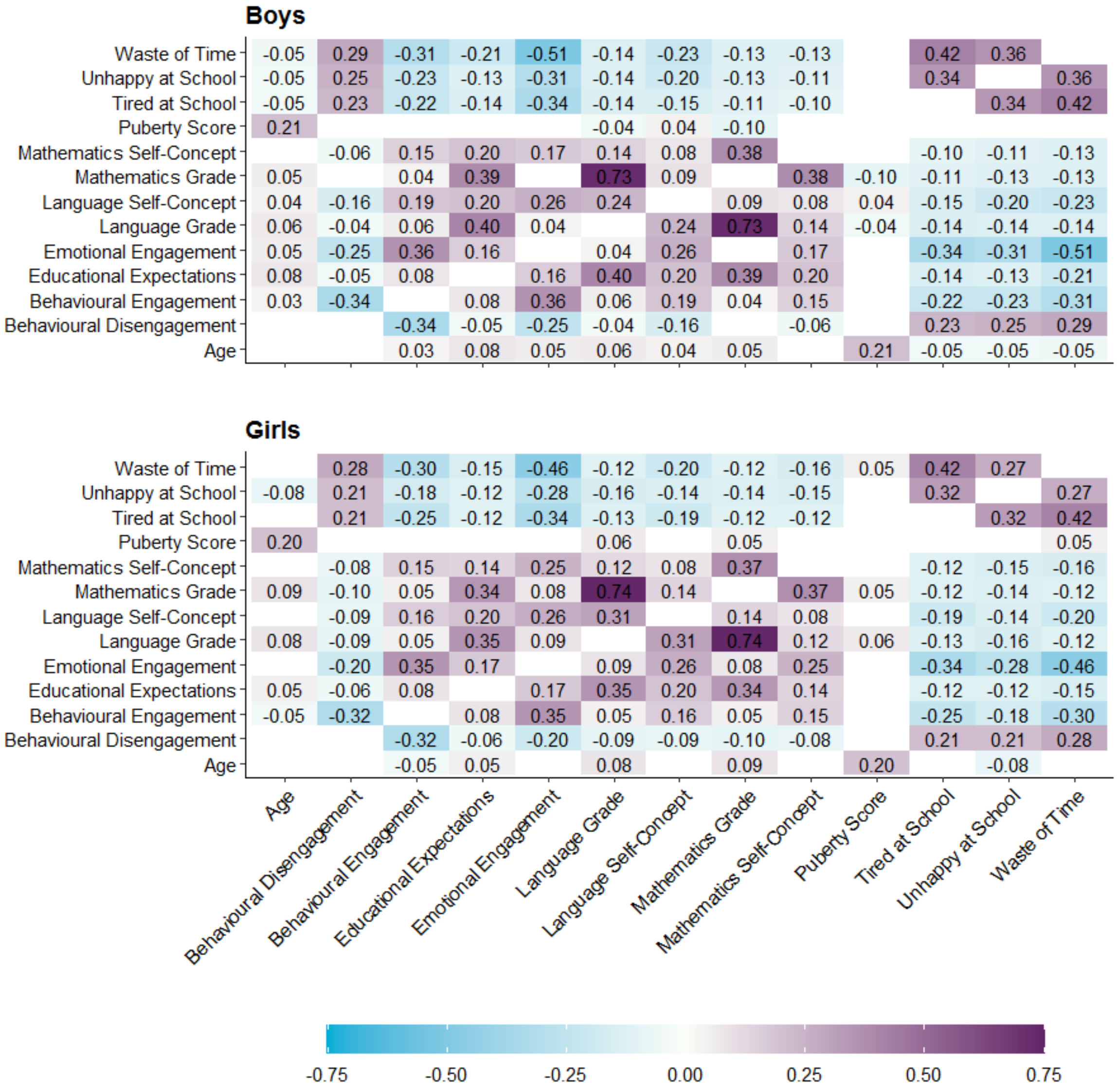
Measure	Wave	SEN / TD	*n*	*M*	*SD*
Behavioural Engagement	11	SEN	1356	3.49	0.63
		TD	4261	3.57	0.55
	14	SEN	1356	3.24	0.65
		TD	4216	3.23	0.60
Behavioural Disengagement	11	SEN	1356	1.59	0.67
		TD	4261	1.46	0.58
	14	SEN	1307	1.66	0.68
		TD	4204	1.60	0.61
Emotional Engagement	11	SEN	1356	2.74	0.75
		TD	4261	2.89	0.65
	14	SEN	1356	2.47	0.71
		TD	4261	2.49	0.66
Emotional Disengagement Tired at School	11	SEN	1309	2.18	0.85
		TD	4185	1.97	0.71
	14	SEN	1309	2.43	0.85
		TD	4203	2.45	0.80
Emotional Disengagement Unhappy at School	11	SEN	1313	1.96	0.72
		TD	4187	1.73	0.60
	14	SEN	1309	1.87	0.71
		TD	4203	1.80	0.69
Emotional Disengagement Waste of Time	11	SEN	1308	1.71	0.88
		TD	4186	1.50	0.70
	14	SEN	1308	1.83	0.84
		TD	4202	1.73	0.76
Language Self-Concept	11	SEN	1311	2.94	0.76
		TD	4184	3.18	0.65
	14	SEN	1309	2.85	0.73
		TD	4205	3.07	0.69
Mathematics Self-Concept	11	SEN	1307	3.22	0.80
		TD	4179	3.33	0.71
	14	SEN	1307	2.91	0.82
		TD	4205	3.06	0.78

Note. SEN - Special Educational Needs. TD - Typically Developing. Measures were reported on an ordinal 1-4 scale.

C.4 Age 11 Correlation Plots

Figure C.1

Bayesian Correlations for Boys and Girls of Age 11 Measures and Age 11 Achievement.



Note. Correlations that are not practically significant (*pd* < 0.95) are excluded.

C.5 Primary to Secondary Transitions

Table C.6

Descriptive Statistics for Year 6 and Year 7 Girls and Boys. School Engagement and Self-Concepts

Measure	Sex	Year Group	*n*	*M*	*SD*
Behavioural Engagement	M	6	2606	3.47	0.60
		7	59	3.51	0.54
	F	6	2733	3.63	0.59
		7	74	3.62	0.59
Behavioural Disengagement	M	6	2607	1.66	0.63
		7	60	1.48	0.60
	F	6	2726	1.33	0.53
		7	74	1.39	0.57
Emotional Engagement	M	6	2612	2.77	0.68
		7	60	2.83	0.69
	F	6	2728	2.93	0.66
		7	74	3.12	0.70
Emotional Disengagement Tired at School	M	6	2606	2.07	0.79
		7	59	2.00	0.79
	F	6	2723	1.97	0.72
		7	73	1.99	0.79
Emotional Disengagement Unhappy at School	M	6	2613	1.82	0.65
		7	58	1.67	0.63
	F	6	2723	1.77	0.63
		7	73	1.52	0.60
Emotional Disengagement Waste of Time	M	6	2602	1.69	0.82
		7	59	1.59	0.67
	F	6	2726	1.41	0.66
		7	74	1.46	0.67
Language Self-Concept	M	6	2609	3.05	0.68
		7	59	3.07	0.58
	F	6	2720	3.20	0.67
		7	74	3.19	0.73
Mathematics Self-Concept	M	6	2604	3.44	0.68
		7	59	3.32	0.71
	F	6	2719	3.17	0.76
		7	74	3.14	0.78

Note. Measures were reported on an ordinal 1-4 scale. School year 6 is the last year of primary education, year 7 is the first year of secondary education.

D Puberty and Age Associations with School Engagement. Supplementary Materials 2

D.1 Puberty Stage Calculation

Each cohort member's puberty stage was calculated at each wave to provide an alternate view of how girls' and boys' pubertal development differed over time. While originally proposed by Crockett (1988), the method was described and published by Carskadon and Acebo (1993). Puberty categorical scores use only three of the PDS measures. For boys, the responses for body hair, voice changes, and facial hair are used. For girls, the responses for body hair, breast growth, and menstruation are used. Scores are totalled and categorised as follows:

Boys

Pre-Pubertal: Score = 3.

Early Pubertal: Score = 4 - 5, no 3-point responses.

Mid-Pubertal: Score = 6 - 8, no 4-point responses.

Late Pubertal: Score = 9 - 11.

Post-Pubertal: Score = 12.

Girls

Pre-Pubertal: Score = 3.

Early Pubertal: Score = 3, no menstruation

Mid-Pubertal: Score = 4, no menstruation.

Late Pubertal: Score $\leq$ 7, and menstruation.

Post-Pubertal: Score = 8, and menstruation.

D.2 Bayesian Priors

Informative priors were set as indicated below for all Model 1 and 2 ordinal logistic regressions. Weak priors, assuming an equal likelihood of choosing each of the four ordinal responses, were set for the thresholds (τ_n) between ordinal responses for each school engagement outcome (Kurz, 2021, e.g.,). Priors for β_x were set reflecting the range of correlations in the literature for in which few correlations exceed 0.3 or 0.4, and some will be close to zero (Hattie, 2009). Also considered were the results of the two prior studies linking puberty with achievement motivation and school engagement (Martin et al., 2017, 2022). An exponential prior is used for the standard deviation (McElreath, 2020).

D.2.1 Bayesian Priors for Univariate Models

$$p(\text{school engagement} = k|\{\tau_k\}, \mu_i) = \Phi(\tau_k\text{-}\mu_i) - \Phi(\tau_{k-1}\text{-}\mu_i)$$

$$\mu_i = \beta_x i + u_i$$

$$u_i \sim \mathcal{N}(0, \sigma_u)$$

$$\tau_1 \sim \mathcal{N}(\text{-}0.674, 1)$$

$$\tau_2 \sim \mathcal{N}(0, 1)$$

$$\tau_3 \sim \mathcal{N}(0.674, 1)$$

$$\beta_{x1-5} \sim \mathcal{N}(0, 0.2)$$

$$\beta_{x6-7} \sim \mathcal{N}(0, 0.3)$$

$$\sigma_u \sim \text{Exponential}(1)$$

where

school engagement indicates the univariate outcomes: behavioural and emotional engagement and behavioural disengagement

τ_k are the thresholds between responses on the 1-4 ordinal scale

μ_i is the random mean for the i^{th} cohort member.

σ_u is the random standard deviation for the i^{th} cohort member.

β_{x1-5} are the estimates for the covariates and educational expectations.

β_{x6-7} are the estimates for self-concepts

D.2.2 Bayesian Priors for the Multivariate (IRT) Models

$$p(\text{response} = k|\{\tau_k\}, \mu_{ij}) = \Phi(\tau_k\text{-}\mu_{ij}) - \Phi(\tau_{k-1}\text{-}\mu_{ij})$$

$$\mu_{ij} = \beta_x i + u_i + v_j$$

$$u_i \sim \mathcal{N}(0, \sigma_u)$$

$$v_j \sim \mathcal{N}(0, \sigma_v)$$

$$\tau_1 \sim \mathcal{N}(\text{-}0.674, 1)$$

$$\tau_2 \sim \mathcal{N}(0, 1)$$

$$\tau_3 \sim \mathcal{N}(0.674, 1)$$

$$\beta_{x1-5} \sim \mathcal{N}(0, 0.2)$$

$$\beta_{x6-7} \sim \mathcal{N}(0, 0.3)$$

$$\sigma_u \sim \text{Exponential}(1)$$

$$\sigma_v \sim \text{Exponential}(1)$$

where

response is the item-specific response for each measure of emotional disengagement

τ_k are the thresholds between responses on the 1-4 ordinal scale

μ_{ij} is the random mean for the i^{th} cohort member and the j^{th} item.

u_i are person level parameters

v_i are item level parameters

σ_u is the random standard deviation for the i^{th} cohort member.

σ_v is the random standard deviation for the j^{th} item.

β_{x1-5} are the estimates for the covariates and educational expectations.

β_{x6-7} are the estimates for self-concepts

D.3 Unconditional Model Selection for Time Measures

Models A and B were also tested with unstructured residuals, allowing variances to differ between boys and girls and between survey waves. Allowing for variances to differ resulted in a significantly poorer model fit. The simpler models were, therefore, retained.

Table D.1

Bayesian Ordinal Logistic Regression Predicting Behavioural Engagement with Cohort Member Random Intercepts: Age, Puberty, and their Interaction

	Median [95% HDPI]			
Measure	Model A	Model B	Model C	Model D
Age	**-0.24[-0.33, -0.15]**		**-0.14[-0.25, -0.03]**	**-0.27[-0.32, -0.22]**
Puberty		**0.35[0.22, 0.48]**	**0.24[0.09, 0.40]**	0.08[-0.01, 0.17]
Puberty*Age	0.00[-0.03, 0.03]	**-0.11[-0.13, -0.09]**	-0.06[-0.10, -0.01]	
ELPD LOO				
Difference	-5.2[-11.6, 1.2]	-3.0[-8.0, 2.0]	0.0	-2.8[-8.0, 2.4]
Difference SE	3.2	2.5	0.0	2.6
Pareto-k > 0.7	13	13	15	10

Note. The median of the posterior distribution is reported with its 95% HDPI (highest density probability interval). Bold typeface highlights effects with a probability of direction $\geq$ 0.95 that make a practically significant contribution to the outcome (significance $\geq$ 0.90) (Makowski et al., 2019). $\widehat{R}$ = 1.00 for all estimates, indicating successful convergence, and all ESS > 1000, indicating estimates are reliable (Bürkner, 2017; Vehtari et al., 2017, 2021). ELPD LOO is used for model comparison, reporting ELPD differences and standard error in comparison to the preferred model (difference = 0.0) (Vehtari et al., 2017).

Table D.2

Bayesian Ordinal Logistic Regression for Emotional Engagement with Cohort Member Random Intercepts: Age, Puberty, and their Interaction

	Median [95% HDPI]			
Measure	Model A	Model B	Model C	Model D
Age	**-0.18[-0.26, -0.09]**		-0.13[-0.23, -0.03]	**-0.22[-0.26, -0.18]**
Puberty		**0.21[0.09, 0.33]**	0.11[-0.03, 0.25]	0.01[-0.08, 0.09]
Puberty*Age	-0.01[-0.04, 0.01]	**-0.09[-0.10, -0.07]**	-0.04[-0.08, 0.00]	
ELPD LOO				
Difference	0.0	-3.4[-10.8, 4.0]	-1.7[-5.3, 1.9]	-1.3[-4.1, 1.5]
Difference SE	0.0	3.7	1.8	1.4
Pareto-k > 0.7	7	5	10	6

Note. The median of the posterior distribution is reported with its 95% HDPI (highest density probability interval). Bold typeface highlights effects with a probability of direction $\geq$ 0.95 that make a practically significant contribution to the outcome (significance $\geq$ 0.90) (Makowski et al., 2019). $\widehat{R}$ = 1.00 for all estimates, indicating successful convergence, and all ESS > 1000, indicating estimates are reliable (Bürkner, 2017; Vehtari et al., 2017, 2021). ELPD LOO is used for model comparison, reporting ELPD differences and standard error in comparison to the preferred model (difference = 0.0) (Vehtari et al., 2017).

Table D.3

Bayesian Ordinal Logistic Regression Predicting Behavioural Disengagement with Cohort Member Random Intercepts: Age, Puberty, and their Interaction

	Median [95% HDPI]			
Measure	Model A	Model B	Model C	Model D
Age	**0.16[0.06, 0.26]**		**0.15[0.07, 0.23]**	**0.16[0.11, 0.21]**
Puberty		**-0.36[-0.50, -0.23]**	-0.24[-0.57, 0.08]	**-0.19[-0.29, -0.09]**
Puberty*Age	-0.02[-0.05, 0.01]	**0.07[0.05, 0.09]**	0.00[-0.02, 0.03]	
ELPD LOO				
Difference	-15.1[-23.9, -6.3]	0.0	-4.2[-7.6, -0.8]	-6.0[-12.0, 0.0]
Difference SE	4.4	0.0	1.7	3.0
Pareto-k > 0.7	45	33	52	29

Note. The median of the posterior distribution is reported with its 95% HDPI (highest density probability interval). Bold typeface highlights effects with a probability of direction ≥ 0.95 that make a practically significant contribution to the outcome (significance ≥ 0.90) (Makowski et al., 2019). $\widehat{R} = 1.00$ for all estimates, indicating successful convergence, and all ESS > 1000, indicating estimates are reliable (Bürkner, 2017; Vehtari et al., 2017, 2021). ELPD LOO is used for model comparison, reporting ELPD differences and standard error in comparison to the preferred model (difference = 0.0) (Vehtari et al., 2017).

Table D.4

Bayesian IRT Ordinal Logistic Regression Predicting Emotional Disengagement with Cohort Member Random Intercepts: Age, Puberty, and their Interaction

	Median [95% HDPI]			
Measure	Model A	Model B	Model C	Model D
Age	-0.05[-0.11, 0.01]		**-0.16[-0.23, -0.09]**	**0.10[0.07, 0.12]**
Puberty	**-0.14[-0.23, -0.06]**	**-0.27[-0.37, -0.18]**	0.04[-0.02, 0.10]	
Puberty * Age	0.05[0.03, 0.07]	0.05[0.04, 0.06]	**0.11[0.09, 0.14]**	
ELPD LOO				
Difference	-47.6[-59.2, -36.0]	-34.3[-44.1, -24.5]	0.0	-26.4[9.6, 43.2]
Difference SE	5.8	4.9	0.0	8.4
Pareto-k > 0.7	5	17	8	15

Note. The median of the posterior distribution is reported with its 95% HDPI (highest density probability interval). Bold typeface highlights effects with a probability of direction ≥ 0.95 that make a practically significant contribution to the outcome (significance ≥ 0.90) (Makowski et al., 2019). $\widehat{R} = 1.00$ for all estimates, indicating successful convergence, and all ESS > 1000, indicating estimates are reliable (Bürkner, 2017; Vehtari et al., 2017, 2021). ELPD LOO is used for model comparison, reporting ELPD differences and standard error in comparison to the preferred model (difference = 0.0) (Vehtari et al., 2017).

D.4 Imputed Boys Dataset for Emotional Disenagement

Table D.5

Bayesian IRT Ordinal Logistic Regression Results for Girls and Boys: Emotional Disengagement.

	Median [95% HDPI]	
	Model 1	Model 2
Age	0.02[-0.04, 0.07]	0.01[-0.04, 0.06]
Puberty	0.04[-0.06, 0.15]	0.07[-0.02, 0.17]
Puberty*Age	0.00[-0.02, 0.03]	0.00[-0.03, 0.02]
Covariates		
SEN Status	**0.21[0.15, 0.28]**	**0.16[0.09, 0.23]**
Income	-0.02[-0.08, 0.05]	-0.04[-0.10, 0.03]
Prior Ach.	**-0.08[-0.13, -0.03]**	0.05[-0.02, 0.11]
Maths Self-concept		**-0.20[-0.30, -0.11]**
Maths Self-concept^3		0.09[0.03, 0.15]
Language Self-concept		**-0.37[-0.47, -0.28]**
Language Self-concept^2		-0.02[-0.09, 0.06]
Language Self-concept^3		**0.08[0.03, 0.13]**
Educational Expectations		**-0.17[-0.20, -0.14]**

Note. Imputed sample size: n = 2752 boys. Educational expectations, mathematics and language self-concept, and family income are time-varying, and the remaining covariates are time-invariant. The median of the posterior distribution is reported with its 95% HDPI (highest density probability interval). Bold typeface highlights effects with a probability of direction > 95% that make a practically significant contribution to the outcome (Median $\geq$ 0.05). ^{2} quadratic and ^{3} cubic associations are reported where they make a non-negligible contribution in at least one of the two models for either boys or girls.

D.5 Main Results without Imputation

Table D.6

Bayesian Ordinal Logistic Regression Results for Girls and Boys: Behavioural Engagement.

	Median [95% HDPI]			
	Model 1		Model 2	
Measure	Male	Female	Male	Female
Age	0.00[-0.10, 0.10]	**-0.14[-0.26, -0.02]**	0.02[-0.08, 0.12]	-0.10[-0.23, 0.03]
Puberty	**0.17[0.00, 0.35]**	0.05[-0.07, 0.18]	0.11[-0.06, 0.29]	0.06[-0.07, 0.19]
Puberty*Age	**-0.09[-0.14, -0.05]**	-0.04[-0.09, 0.00]	**-0.08[-0.13, -0.04]**	-0.05[-0.10, -0.01]
Covariates				
SEN Status	-0.01[-0.13, 0.10]	0.02[-0.11, 0.15]	0.02[-0.10, 0.14]	0.05[-0.08, 0.18]
Income	0.09[-0.04, 0.22]	0.10[-0.03, 0.23]	0.13[-0.01, 0.26]	**0.14[0.01, 0.27]**
Parent Educ	-0.09[-0.26, 0.08]	0.10[-0.07, 0.28]	-0.16[-0.33, 0.01]	0.02[-0.16, 0.20]
Prior Ach	0.07[-0.02, 0.16]	0.08[-0.01, 0.17]	**-0.21[-0.30, -0.12]**	**-0.16[-0.25, -0.07]**
Maths Self-concept			**0.35[0.18, 0.53]**	**0.50[0.36, 0.64]**
Maths Self-concept^2			**0.31[0.16, 0.45]**	0.09[-0.03, 0.21]
Maths Self-concept^3			**-0.15[-0.26, -0.03]**	-0.07[-0.16, 0.03]
Language Self-concept			**0.42[0.24, 0.60]**	**0.47[0.29, 0.64]**
Language Self-concept^2			**0.26[0.11, 0.40]**	**0.16[0.01, 0.30]**
Educational Expectations			**0.18[0.12, 0.24]**	**0.21[0.15, 0.28]**

Note. Educational expectations, mathematics and language self-concept, and family income are time-varying; the remaining covariates are time-invariant. The median of the posterior distribution is reported with its 95% HDPI. Bold typeface highlights effects with a probability of direction > 95% that make a practically significant contribution to the outcome (Median $\geq$ 0.05, and significance $\geq$ 90%). ^{2} quadratic and ^{3} cubic associations are reported where they make a non-negligible contribution in at least one of the two models for either boys or girls. $\widehat{R} < 1.01$ for all estimates, indicating successful convergence and all ESS > 1000, indicating estimates are reliable (Bürkner, 2017; Vehtari et al., 2017, 2021).

Table D.7

Bayesian Ordinal Logistic Regression Results for Girls and Boys: Behavioural Disengagement.

Outcome: BEHAVIOURAL DISENGAGEMENT				
	Median [95% HDPI]			
	Model 1		Model 2	
Measure	Male	Female	Male	Female
Age	-0.01[-0.06, 0.03]	**0.16[0.12, 0.20]**	-0.06[-0.10, -0.01]	**0.14[0.10, 0.18]**
Puberty	0.07[-0.03, 0.17]	0.08[-0.01, 0.17]	**0.12[0.02, 0.22]**	0.08[-0.01, 0.17]
Covariates				
SEN Status	0.07[-0.06, 0.20]	-0.05[-0.19, 0.09]	0.03[-0.10, 0.16]	-0.09[-0.23, 0.05]
Income	**-0.27[-0.42, -0.13]**	**-0.36[-0.50, -0.22]**	**-0.30[-0.46, -0.16]**	**-0.37[-0.51, -0.23]**
Parent Educ	0.10[-0.09, 0.28]	-0.04[-0.22, 0.15]	0.15[-0.04, 0.35]	0.01[-0.18, 0.20]
Prior Ach	**-0.11[-0.21, -0.02]**	**-0.25[-0.34, -0.16]**	0.08[-0.03, 0.18]	-0.11[-0.21, -0.01]
Maths Self-concept			**-0.28[-0.46, -0.09]**	**-0.31[-0.46, -0.16]**
Language Self-concept			**-0.31[-0.50, -0.12]**	**-0.27[-0.44, -0.10]**
Language Self-concept^2			**-0.24[-0.39, -0.09]**	-0.11[-0.25, 0.04]
Language Self-concept^3			**0.11[0.01, 0.21]**	0.12[0.01, 0.23]
Educational Expectations			**-0.14[-0.20, -0.07]**	**-0.16[-0.23, -0.10]**

Note. Educational expectations, mathematics and language self-concept, and family income are time-varying; the remaining covariates are time invariant. The median of the posterior distribution is reported with its 95% HDPI. Bold typeface highlights effects with a probability of direction > 95% that make a practically significant contribution to the outcome (Median $\geq$ 0.05, and significance $\geq$ 90%). ^{2} quadratic and ^{3} cubic associations are reported where they make a non-negligible contribution in at least one of the two models for either boys or girls. $\widehat{R} < 1.01$ for all estimates, indicating successful convergence and all ESS > 1000, indicating estimates are reliable (Bürkner, 2017; Vehtari et al., 2017, 2021).

Table D.8

Bayesian Ordinal Logistic Regression Results for Girls and Boys: Emotional Engagement.

	Median [95% HDPI]			
	Model 1		Model 2	
Measure	Male	Female	Male	Female
Age	**-0.14[-0.18, -0.10]**	**-0.27[-0.30, -0.23]**	**-0.10[-0.14, -0.06]**	**-0.24[-0.28, -0.20]**
Puberty	-0.06[-0.14, 0.03]	-0.04[-0.12, 0.03]	-0.09[-0.18, 0.00]	-0.04[-0.11, 0.03]
Covariates				
SEN Status	-0.09[-0.20, 0.02]	-0.03[-0.15, 0.09]	-0.05[-0.15, 0.06]	0.01[-0.10, 0.14]
Income	-0.08[-0.21, 0.05]	-0.05[-0.17, 0.07]	-0.01[-0.14, 0.11]	0.01[-0.11, 0.13]
Income^2	0.11[0.01, 0.22]	**0.15[0.05, 0.25]**	0.09[-0.02, 0.19]	0.10[0.00, 0.02]
Parent Educ.	0.14[-0.02, 0.30]	**0.26[0.10, 0.43]**	0.04[-0.11, 0.19]	0.13[-0.03, 0.30]
Prior Ach.	0.08[0.00, 0.15]	**0.23[0.15, 0.31]**	**-0.25[-0.34, -0.17]**	-0.07[-0.15, 0.01]
Achievement Motivation				
Maths Self-concept			**0.28[0.11, 0.46]**	**0.65[0.52, 0.79]**
Maths Self-concept^2			**0.37[0.23, 0.51]**	0.05[-0.06, 0.16]
Maths Self-concept^3			**-0.19[-0.30, -0.08]**	**-0.12[-0.21, -0.03]**
Language Self-Concept			**0.52[0.35, 0.70]**	**0.66[0.49, 0.82]**
Language Self-Concept^2			**0.24[0.10, 0.38]**	0.01[-0.12, 0.14]
Language Self-Concept^3			**-0.15[-0.24, -0.05]**	0.01[-0.09, 0.11]
Educational Expectations			**0.28[0.23, 0.34]**	**0.29[0.23, 0.34]**

Note. Educational expectations, mathematics and language self-concept, and family income are time-varying; the remaining covariates are time invariant. The median of the posterior distribution is reported with its 95% HDPI (highest density probability interval). Bold typeface highlights effects with a probability of direction > 95% that make a practically significant contribution to the outcome (Median $\geq$ 0.05, and significance $\geq$ 90%). Effects that are close but do not meet these criteria are highlighted in the text. ^{2} quadratic and ^{3} cubic relationships associations are reported where they make a non-negligible contribution in at least one of the two models for either boys or girls. $\widehat{R} < 1.01$ for all estimates, indicating successful convergence and all ESS > 1000, indicating estimates are reliable (Bürkner, 2017; Vehtari et al., 2017, 2021).

D.6 Discarded Model B, Behavioural Disengagement

Model B regression results with missing-data datasets. Main analyses reverted to Model D, with simple main effects for puberty and age following these results. See the Analyses section in the main text for further details.

Table D.9

Bayesian Ordinal Logistic Regression Results for Girls and Boys: Behavioural Disengagement with puberty/age interaction

	Median [95% HDPI]			
	Model 1		Model 2	
Measure	Male	Female	Male	Female
Puberty	0.04[-0.12, 0.18]	-0.04[-0.16, 0.07]	0.14[-0.02, 0.30]	-0.03[-0.15, 0.09]
Puberty*Age	0.00[-0.02, 0.02]	0.06[0.04, 0.07]	-0.02[-0.04, 0.00]	0.05[0.04, 0.07]
Covariates				
SEN Status	0.07[-0.06, 0.19]	-0.05[-0.19, 0.09]	0.03[-0.01, 0.16]	-0.09[-0.24, 0.05]
Family Income	**-0.10[-0.15, -0.05]**	**-0.13[-0.17, -0.08]**	**-0.11[-0.16, -0.06]**	**-0.13[-0.18, -0.08]**
Parent Educ.	0.00[-0.01, 0.10]	0.00[-0.06, 0.05]	0.08[0.02, 0.14]	0.02[-0.04, 0.08]
Prior Ach.	0.11[-0.20, -0.02]	**-0.25[-0.35, -0.15]**	0.08[-0.02, 0.18]	-0.11[-0.21, -0.01]
Achievement Motivation				
Maths Self-concept			**-0.28[-0.47, -0.09]**	**-0.31[-0.46, -0.16]**
Maths Self-concept^2			-0.11[-0.27, 0.04]	0.06[-0.07, 0.18]
Maths Self-concept^3			0.12[0.00, 0.25]	0.07[-0.03, 0.16]
Language Self-concept			**-0.31[-0.50, -0.12]**	**-0.26[-0.44, -0.08]**
Language Self-concept^2			**-0.24[-0.39, -0.08]**	-0.10[-0.26, 0.06]
Language Self-concept^3			0.11[0.00, 0.21]	0.12[0.01, 0.23]
Educational Expectations			**-0.13[-0.20, -0.07]**	**-0.16[-0.22, -0.10]**

Note. Educational (Ed.) expectations, mathematics and language self-concept and family income are time-varying; the remaining covariates are time-invariant. The median of the posterior distribution is reported with its 95% HDPI (highest density probability interval). $\widehat{R} = 1.00$ for all estimates, indicating successful convergence and all ESS > 2900 indicating estimates are reliable (Bürkner, 2017; Vehtari et al., 2017, 2021). Bold typeface highlights significant effects.

References

Ahmad, F., Jhajj, A. K., Stewart, D. E., Burghardt, M., & Bierman, A. S. (2014). Single item measures of self-rated mental health: A scoping review. *BMC Health Services Research*, *14*, 1–11. https://doi.org/10.1186/1472-6963-14-398

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.), *Advances in Child Development and Behavior* (pp. 289–326, Vol. 64). JAI. https://doi.org/10.1016/bs.acdb.2022.10.006

Åhslund, I., & Boström, L. (2018). Teachers' Perceptions of Gender Differences: What about Boys and Girls in the Classroom? *International Journal of Learning, Teaching and Educational Research*, *17*(4), 28–44. https://doi.org/https://doi.org/10.26803/ijlter.17.4.2

Alan, S., Ertac, S., & Mumcu, I. (2018). Gender Stereotypes in the Classroom and Effects on Achievement. *The Review of Economics and Statistics*, *100*(5), 876–890. https://doi.org/10.1162/rest_a_00756

Allen, M., Iliescu, D., & Greiff, S. (2022). Single Item Measures in Psychological Science. *European Journal of Psychological Assessment*. https://doi.org/10.1027/1015-5759/a000699

Allen, R. (2015). *Missing Talent* (tech. rep.). Sutton Trust. https://www.suttontrust.com/our-research/missing-talent-disadvantaged-pupil-attainment/

Anders, J. (2012). The Link between Household Income, University Applications and University Attendance. *Fiscal Studies*, *33*(2), 185–210. https://doi.org/10.1111/j.1475-5890.2012.00158.x

Ang, L., & Eisend, M. (2018). Single versus multiple measurement of attitudes: A meta-analysis of advertising studies validates the single-item measure approach. *Journal of Advertising Research*, *58*(2), 218–227.

APPG, Men & Boys. (2023). *Closing the Gender Attainment Gap, Part1: Statistics* (Inquiry No. 4). All Party Parliamentary Group. UK. https://equi-law.uk/inquiry-4-boys-edu-underachievement/

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

Arden, R., & Plomin, R. (2006). Sex differences in variance of intelligence across childhood. *Personality and Individual Differences*, *41*(1), 39–48. https://doi.org/10.1016/j.paid.2005.11.027

Atit, K., Power, J. R., Veurink, N., Uttal, D. H., Sorby, S., Panther, G., Msall, C., Fiorella, L., & Carr, M. (2020). Examining the role of spatial skills and mathematics motivation on middle school mathematics achievement. *International Journal of STEM Education*, *7*(1), 38. https://doi.org/10.1186/s40594-020-00234-3

Autor, D. (2010). U.S. Labor Market Challenges over the Longer Term. Retrieved March 19, 2024, from https://economics.mit.edu/sites/default/files/publications/us%20labor%20market%20challenges%202010.pdf

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research*, *20*(1), 40–49. https://doi.org/10.1002/mpr.329

Bae, C. L., & DeBusk-Lane, M. (2019). Middle school engagement profiles: Implications for motivation and achievement in science. *Learning and Individual Differences*, *74*, 101753. https://doi.org/10.1016/j.lindif.2019.101753

Bagnall, C. L., Fox, C. L., & Skipper, Y. (2021). What emotional-centred challenges do children attending special schools face over primary–secondary school transition? *Journal of Research in Special Educational Needs*, *21*(2), 156–167. https://doi.org/10.1111/1471-3802.12507

Bakadorova, O., Lazarides, R., & Raufelder, D. (2020). Effects of social and individual school self-concepts on school engagement during adolescence. *European Journal of Psychology of Education*, *35*(1), 73–91. https://doi.org/10.1007/s10212-019-00423-x

Balzer, B. W. R., Garden, F. L., Amatoury, M., Luscombe, G. M., Paxton, K., Hawke, C. I., Handelsman, D. J., & Steinbeck, K. S. (2019). Self-rated Tanner stage and subjective measures of puberty are associated with longitudinal gonadal hormone changes. *Journal of Pediatric Endocrinology & Metabolism*, *32*(6), 569–576. https://doi.org/10.1515/jpem-2019-0017

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

Bear, G., Mantz, L., Glutting, J., Yang, C., & Boyer, D. (2015). Differences in Bullying Victimization Between Students With and Without Disabilities. *School Psychology Review*, *44*, 98–116. https://doi.org/10.17105/SPR44-1.98-116

Becherer, J., Köller, O., & Zimmermann, F. (2021). Externalizing behaviour, task-focused behaviour, and academic achievement: An indirect relation? *British Journal of Educational Psychology*, *91*(1), 27–45. https://doi.org/10.1111/bjep.12347

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

Bergkvist, L., & Rossiter, J. R. (2007). The predictive validity of multiple-item versus single-item measures of the same constructs. *Journal of Marketing Research*, *44*(2), 175–184. https://doi.org/https://doi.org/10.1509/jmkr.44.2.175

Berkman, E. T., Livingston, J. L., & Kahn, L. E. (2017). Finding The “Self” in Self-Regulation: The Identity-Value Model. *Psychological Inquiry*. https://doi.org/10.1080/1047840X.2017.1323463

Bernardi, F., & Valdés, M. T. (2021). Sticky educational expectations: A cross-country comparison. *Research in Social Stratification and Mobility*, *75*, 100624. https://doi.org/10.1016/j.rssm.2021.100624

Blakemore, S. (2010). The Developing Social Brain: Implications for Education. *Neuron*, *65*(6), 744–747. https://doi.org/10.1016/j.neuron.2010.03.004

Blossfeld, H.-P., Skopek, J., Triventi, M., & Buchholz, S. (2015). *Gender, education and employment: An international comparison of school-to-work transitions*. Edward Elgar Publishing.

Bodner, T. E. (2008). What Improves with increased missing data imputations? [Publisher: Routledge]. *Structural Equation Modeling: A Multidisciplinary Journal*, *15*(4), 651–675. https://doi.org/10.1080/10705510802339072

Bollen, K. A., Biemer, P. P., Karr, A. F., Tueller, S., & Berzofsky, M. E. (2016). Are Survey Weights Needed? A Review of Diagnostic Tests in Regression Analysis. *Annual Review of Statistics and Its Application*, *3*, 375–392. https://doi.org/10.1146/annurev-statistics-011516-012958

Borra, C., Iacovou, M., & Sevilla, A. (2023). Adolescent development and the math gender gap. *European Economic Review*, *158*, 104542. https://doi.org/10.1016/j.euroecorev.2023.104542

Breit, M., Scherrer, V., Tucker-Drob, E. M., & Preckel, F. (2024). The stability of cognitive abilities: A meta-analytic review of longitudinal studies. *Psychological Bulletin*, *150*(4), 399–439. https://doi.org/10.1037/bul0000425

Broeke, S., & Hamed, J. (2008). *Gender Gaps in Higher Education Participation* (DIUS Research Report No. 08-14). Department for Innovation, Universities and Skills. UK. https://dera.ioe.ac.uk/8717/1/DIUS-RR-08-14.pdf

Brooks-Gunn, J., Warren, M. P., Rosso, J., & Gargiulo, J. (1987). Validity of Self-Report Measures of Girls’ Pubertal Status. *Child Development*, *58*(3), 829–841. https://doi.org/10.2307/1130220

Brooks-Gunn, J., Petersen, A. C., & Eichorn, D. (1985). The study of maturational timing effects in adolescence. *Journal of Youth and Adolescence*, *14*(3), 149–161. https://doi.org/10.1007/BF02090316

Buchmann, M., Grütter, J., & Zuffianò, A. (2022). Parental educational aspirations and children’s academic self-concept: Disentangling state and trait components on their dynamic interplay. *Child Development*, *93*(1), 7–24. https://doi.org/10.1111/cdev.13645

Bugbee, B., Beck, K., Fryer, C., & Arria, A. (2019). Substance Use, Academic Performance, and Academic Engagement Among High School Seniors. *Journal of School Health*, *89*, 145–156. https://doi.org/10.1111/josh.12723

Burgess, S., & Greaves, E. (2013). Test Scores, Subjective Assessment, and Stereotyping of Ethnic Minorities. *Journal of Labor Economics*, *31*(3), 535–576. https://doi.org/10.1086/669340

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*, 102267. https://doi.org/10.1016/j.econedurev.2022.102267

Bürkner, P.-C. (2017). Brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*, 1–28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C. (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal*, *10*(1), 395–411. Retrieved June 5, 2024, from https://journal.r-project.org/archive/2018/RJ-2018-017/index.html

Bürkner, P.-C. (2021). Bayesian Item Response Modeling in R with brms and Stan. *Journal of Statistical Software*, *100*, 1–54. https://doi.org/10.18637/jss.v100.i05

Bürkner, P.-C. (2024, March). Handle Missing Values with brms. Retrieved October 1, 2024, from https://paulbuerkner.com/brms/articles/brms_missings.html

Bürkner, P.-C., & Vuorre, M. (2019). Ordinal Regression Models in Psychology: A Tutorial. *Advances in Methods and Practices in Psychological Science*, *2*(1), 77–101. https://doi.org/10.1177/2515245918823199

Campbell, T. (2015). Stereotyped at Seven? Biases in Teacher Judgement of Pupils' Ability and Attainment. *Journal of Social Policy*, *44*(3), 517–547. https://doi.org/10.1017/S0047279415000227

Caponera, E., Sestito, P., & Russo, P. M. (2016). The influence of reading literacy on mathematics and science achievement. *The Journal of Educational Research*, *109*(2), 197–204. https://www.jstor.org/stable/26586980

Cappon, P. (2011). Exploring the 'Boy Crisis' in Education. https://files.eric.ed.gov/fulltext/ED518173.pdf

Carlana, M. (2019). Implicit Stereotypes: Evidence from Teachers' Gender Bias. *The Quarterly Journal of Economics*, *134*(3), 1163–1224. https://doi.org/10.1093/qje/qjz008

Carroll, M. (2024). Sex gaps in education in England. https://www.cambridgeassessment.org.uk/Images/698454-sex-gaps-in-education-in-england.pdf

Carskadon, M. A., & Acebo, C. (1993). A Self-Administered Rating Scale for Pubertal Development. *Journal of Adolescent Health*, *14*, 190–198.

Carter, R., Leath, S., Butler-Barnes, S. T., Byrd, C. M., Chavous, T. M., Caldwell, C. H., & Jackson, J. S. (2017). Comparing Associations Between Perceived Puberty, Same-Race Friends and Same-Race Peers, and Psychosocial Outcomes Among African American and Caribbean Black Girls. *Journal of Black Psychology*, *43*(8), 836–862. https://doi.org/10.1177/0095798417711024

Carter, R., Mustafaa, F. N., & Leath, S. (2018). Teachers' Expectations of Girls' Classroom Performance and Behavior: Effects of Girls' Race and Pubertal Timing. *The Journal of Early Adolescence*, *38*(7), 885–907. https://doi.org/10.1177/0272431617699947

Casey, B. M., & Ganley, C. M. (2021). An examination of gender differences in spatial skills and math attitudes in relation to mathematics success: A bio-psycho-social model. *Developmental Review*, *60*, 100963. https://doi.org/10.1016/j.dr.2021.100963

Caspi, A., & Moffitt, T. E. (1991). Individual differences are accentuated during periods of social change: The sample case of girls at puberty. *Journal of Personality and Social Psychology*, *61*(1), 157–168. https://doi.org/10.1037/0022-3514.61.1.157

Castellanos-Ryan, N., Parent, S., Vitaro, F., Tremblay, R. E., & Séguin, J. R. (2013). Pubertal development, personality, and substance use: A 10-year longitudinal study from childhood to adolescence. *Journal of Abnormal Psychology*, *122*(3), 782–796. https://doi.org/10.1037/a0033133

Castro, M. S., Bahli, B., Ferreira, J. J., & Figueiredo, R. (2023). Comparing Single-Item and Multi-Item Trust Scales: Insights for Assessing Trust in Project Leaders. *Behavioral Sciences*, *13*(9), 786. https://doi.org/10.3390/bs13090786

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. https://doi.org/10.1177%2F019027250707000207

Ceci, S. J. (2017). Women in Academic Science: Experimental Findings From Hiring Studies. *Educational Psychologist*, *53*(1), 22–41. https://doi.org/10.1080/00461520.2017.1396462

Ceci, S. J., & Williams, W. M. (2010). Sex Differences in Math-Intensive Fields. *Current Directions in Psychological Science*, *19*(5), 275–279. https://doi.org/10.1177/0963721410383241

Chaku, N., & Hoyt, L. T. (2019). Developmental Trajectories of Executive Functioning and Puberty in Boys and Girls. *Journal of Youth and Adolescence*, *48*(7), 1365–1378. https://doi.org/10.1007/s10964-019-01021-2

Champely, S. (2020). Pwr: Basic Functions for Power Analysis. https://CRAN.R-project.org/package=pwr.

Chang, D.-F., Chien, W.-C., & Chou, W.-C. (2016). Meta-analysis approach to detect the effect of student engagement on academic achievement. *ICIC Express Letters*, *10*, 2441–2446.

Chase, P. A., Hilliard, L. J., John Geldhof, G., Warren, D. J. A., & Lerner, R. M. (2014). Academic Achievement in the High School Years: The Changing Role of School Engagement. *Journal of Youth and Adolescence*, *43*(6), 884–896. https://doi.org/10.1007/s10964-013-0085-4

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). https://doi.org/10.1177/2332858416673617

Coelho, V. A., & Romão, A. M. (2017). The Impact of Secondary School Transition on Self-Concept and Self-Esteem. *Revista de Psicodidáctica (English ed.)*, *22*(2), 85–92. https://doi.org/10.1016/j.psicoe.2016.10.001

Coleman, L., & Coleman, J. (2002). The measurement of puberty: A review. *Journal of Adolescence*, *25*(5), 535–550. https://doi.org/10.1006/jado.2002.0494

Collado, R. (2019). Reducing the gender dgap in spatial skills in high school physics. *GIREP-ICPE-EPEC 2017 Conference*, *1286*. https://doi.org/10.1088/1742-6596/1286/1/012013

Costa, P. T., Terracciano, A., & McCrae, R. R. (2001). Gender differences in personality traits across cultures: Robust and surprising findings. *Journal of Personality and Social Psychology*, *81*(2), 322–331. https://doi.org/10.1037/0022-3514.81.2.322

Crockett, L. J. (1988). Pubertal Development Scale:Pubertal Categories.

Crockett, L. J., Schulenberg, J. E., & Petersen, A. C. (1987). Congruence between Objective and Self-Report Data in a Sample of Young Adolescents. *Journal of Adolescent Research*, *2*(4), 383–392. https://doi.org/10.1177/074355488724006

Crone, E. A., & Dahl, R. E. (2012). Understanding adolescence as a period of social-affective engagement and goal flexibility. *Nature Reviews Neuroscience*, *13*(9), 636–650. https://doi.org/10.1038/nrn3313

Crouter, A. C., Whiteman, S. D., McHale, S. M., & Osgood, D. W. (2007). Development of Gender Attitude Traditionality Across Middle Childhood and Adolescence. *Child Development*, *78*(3), 911–926. https://doi.org/10.1111/j.1467-8624.2007.01040.x

Cuff, B. M. (2017). *Perceptions of subject difficulty and subject choices: Are the two linked, and if so, how?* (Tech. rep. No. Ofqual/17/6288). Ofqual. UK. https://assets.publishing.service.gov.uk/media/5a81f60d40f0b62305b91bbb/Perceptions_of_subject_difficulty_and_subject_choices.pdf

Curran, P. J., Obeidat, K., & Losardo, D. (2010). Twelve Frequently Asked Questions About Growth Curve Modeling. *Journal of Cognition and Development*, *11*(2), 121–136. https://doi.org/10.1080/15248371003699969

Cvencek, D., Meltzoff, A. N., & Greenwald, A. G. (2011). Math-Gender Stereotypes in Elementary School Children: Gender Stereotypes. *Child Development*, *82*(3), 766–779. https://doi.org/10.1111/j.1467-8624.2010.01529.x

Dai, D. Y. (2001). A comparison of gender differences in academic self-concept and motivation between high-ability and average Chinese adolescents. *Journal of Secondary Gifted Education*, *13*(1), 22–32. https://doi.org/jsge-2001-361

Daniel, J., & Wang, H. (2023). Gender differences in special educational needs identification. *Review of Education*, *11*(3), e3437. https://doi.org/10.1002/rev3.3437

Davis-Kean, P. E., Huesmann, L. R., Jager, J., Bates, J. E., Collins, W. A., & Lansford, J. E. (2008). Changes in the Relation of Self-Efficacy Beliefs and Behaviors across Development. *Child Development*, *79*(5), 1257–1269. https://www.jstor.org/stable/27563551

De Bolle, M., De Fruyt, F., McCrae, R. R., Löckenhoff, C., Costa Jr, P. T. C., Aguilar-Vafaie, M. E., Ahn, C.-k., Ahn, H.-n., Alcalay, L., Allik, J., Avdeyeva, T. V., Bratko, D., Brunner-Sciarra, M., Cain, T. R., Chan, W., Chittcharat, N., Crawford, J. T., Fehr, R., Ficková, E., … Terracciano, A. (2015). The emergence of sex differences in personality traits in early adolesence: A cross-sectional, cross-cultural study. *Journal of Personality and Social Psychology*, *108*(1), 171–185. https://doi.org/10.1037/a0038497

De Fraine, B., Damme, J. V., & Onghena, P. (2007). A longitudinal analysis of gender differences in academic self-concept and language achievement: A multivariate multilevel latent growth approach. *Contemporary Educational Psychology*, *32*(1), 132–150. https://doi.org/10.1016/j.cedpsych.2006.10.005

De Fruyt, F., Van Leeuwen, K., De Bolle, M., & De Clercq, B. (2008). Sex differences in school performance as a function of conscientiousness, imagination and the mediating role of problem behaviour. *European Journal of Personality*, *22*(3), 167–184. https://doi.org/10.1002/per.675

Deardorff, J., Hoyt, L. T., Carter, R., & Shirtcliff, E. A. (2019). Next Steps in Puberty Research: Broadening the Lens Toward Understudied Populations. *Journal of Research on Adolescence*, *29*(1), 133–154. https://doi.org/10.1111/jora.12402

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

Deci, E. L., & Ryan, R. (1985). *Intrinsic Motivation and Self-Determination in Human Behavior*. Springer US. https://doi.org/10.1007/978-1-4899-2271-7

Dégeilh, F., Guillery-Girard, B., Dayan, J., Gaubert, M., Chételat, G., Egler, P.-J., Baleyte, J.-M., Eustache, F., & Viard, A. (2015). Neural Correlates of Self and Its Interaction With Memory in Healthy Adolescents. *Child Development*, *86*(6), 1966–1983. https://doi.org/10.1111/cdev.12440

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), 92–101. https://doi.org/10.5334/irsp.82

Denissen, J. J. A., Zarrett, N. R., & Eccles, J. S. (2007). I Like to Do It, I'm Able, and I Know I Am: Longitudinal Couplings Between Domain-Specific Achievement, Self-Concept, and Interest. *Child Development*, *78*(2), 430–447. https://doi.org/10.1111/j.1467-8624.2007.01007.x

Department for Business, Innovation and Skills. (2011). The 2011 Skills for Life Survey: A Survey of Literacy, Numeracy and ICT Levels in England [BIS Research Paper Number 81]. https://www.gov.uk/government/publications/2011-skills-for-life-survey

Department for Business, Innovation and Skills. (2013). *The Benefits of Higher Education Participation for Individuals and Society: Key findings and reports "The Quadrants"* (BIS Research Paper No. 146). GOV.UK. https://assets.publishing.service.gov.uk/media/5a7ba5fee5274a7318b9004a/bis-13-1268-benefits-of-higher-education-participation-the-quadrants.pdf

Department for Education. (2010). *SFR 33/2010 Main Text* (Statistical First Release No. 33/2010). London. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/218856/sfr33-2010.pdf

Department for Education. (2012). *SFR 19/2012 Main Text (provisional)* (Statistical First Release No. 19/2012). London. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/219204/sfr19-2012.pdf

Department for Education. (2015). National curriculm in England: Science programmes of study.

https://www.gov.uk/government/publications/national-curriculum-in-england-science-programmes-of-study/national-curriculum-in-england-science-programmes-of-study

Department for Education. (2016). Progress_8_school_performance_measure_13_oct_update.pdf. Retrieved September 29, 2024, from https://dera.ioe.ac.uk/id/eprint/27619/1/Progress_8_school_performance_measure_13_Oct_update.pdf

Department for Education. (2017). *SFR 69/2017 Main Text (revised)* (Statistical First Release No. 69/2017). London.

Department for Education. (2018). *SFR01/2018 Subject tables* (Statistical First Release No. 01/2018). Department for Education. London. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/784809/SFR01_2018_Subject_tables.xlsx

Department for Education. (2019, August). Guidance: English Baccalaureate (Ebacc). https://www.gov.uk/government/publications/english-baccalaureate-ebacc/english-baccalaureate-ebacc

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

Department for Education. (2022, June). *Special Educational Needs and Disability: An analysis and summary of data sources* (tech. rep.).

Department for Education. (2023, June). *Special educational needs and disability: An analysis and summary of data sources* (tech. rep. No. SEN-2023). https://assets.publishing.service.gov.uk/media/64930eef103ca6001303a3a6/Special_educational_needs_and_disability_an_analysis_and_summary_of_data_sources.pdf

Devine, A., Fawcett, K., Szűcs, D., & Dowker, A. (2012). Gender differences in mathematics anxiety and the relation to mathematics performance while controlling for test anxiety. *Behavioral and Brain Functions*, *8*(1), 33. https://doi.org/10.1186/1744-9081-8-33

Diamond, A. (2013). Executive functions. *Annual Review of Psychology*, *64*, 135–168.

Dijkstra, P., Kuyper, H., van der Werf, G., Buunk, A. P., & van der Zee, Y. G. (2008). Social Comparison in the Classroom: A Review. *Review of Educational Research*, *78*(4), 828–879. https://doi.org/10.3102/0034654308321210

Dimler, L. M., & Natsuaki, M. N. (2015). The effects of pubertal timing on externalizing behaviors in adolescence and early adulthood: A meta-analytic review. *Journal of Adolescence*, *45*, 160–170. http://www.sciencedirect.com/science/article/pii/S0140197115002249

Ding, C. S., Song, K., & Richardson, L. I. (2006). Do Mathematical Gender Differences Continue? A Longitudinal Study of Gender Difference and Excellence in Mathematics Performance in the U.S. *Educational Studies*, *40*(3), 279–295. https://doi.org/10.1080/00131940701301952

Dockery, A. M., Koshy, P., & Li, I. W. (2022). Parental expectations of children's higher education participation in Australia. *British Educational Research Journal*, *48*, 617–639. https://doi.org/10.1002/berj.3786

Dorn, L. D., Susman, E. J., & Ponirakis, A. (2003). Pubertal Timing and Adolescent Adjustment and Behavior: Conclusions Vary by Rater. *Journal of Youth and Adolescence*, *32*(3), 157–167. https://doi.org/10.1023/A:1022590818839

Dreber, A., Essen, E. v., & Ranehill, E. (2012). Age at Pubertal Onset and Educational Outcomes. *SSRN Electronic Journal*, 22. https://doi.org/10.2139/ssrn.1934188

Dubas, J. S., Graber, J. A., & Petersen, A. C. (1991). The Effects of Pubertal Development on Achievement during Adolescence. *American Journal of Education*, *99*(4), 444–460. https://doi.org/10.1086/443993

Duckworth, A. L., & Seligman, M. E. P. (2006). Self-discipline gives girls the edge: Gender in self-discipline, grades, and achievement test scores. *Journal of Educational Psychology*, *98*(1), 198–208. https://doi.org/10.1037/0022-0663.98.1.198

Duncan, T. E., Duncan, S. C., & Strycker, L. A. (2013). *An Introduction to Latent Variable Growth Curve Modeling: Concepts, Issues, and Application, Second Edition* (2nd ed.). Routledge. https://doi.org/10.4324/9780203879962

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, *105*(3), 399–412. https://doi.org/10.1111/bjop.12046

Eccles, J. S., 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.

Eccles, J. S., Midgley, C., Wigfield, A., Buchanan, C. M., Reuman, D., Flanagan, C., & Iver, D. M. (1993). Development during adolescence. The impact of stage-environment fit on young adolescents' experiences in schools and in families. *The American Psychologist*, *48*(2), 90–101. https://doi.org/10.1037//0003-066x.48.2.90

Eccles, J. S., & Roeser, R. W. (2013). Schools, academic motivation, and stage-environment fit. In R. M. Lerner & L. Steinberg (Eds.), *Handbook of Adolescent Psychology* (2nd ed., pp. 125–153). John Wiley & Sons, Inc.

Eccles, J. S., & Wang, M.-T. (2012). Part I Commentary: So What Is Student Engagement Anyway? In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of Research on Student Engagement* (pp. 133–145). Springer US. https://doi.org/10.1007/978-1-4614-2018-7_6

Eccles, J. S., & Wigfield, A. (1995). In the Mind of the Actor: The Structure of Adolescents' Achievement Task Values and Expectancy-Related Beliefs. *Personality and Social Psychology Bulletin*, *21*(3), 215–225. https://doi.org/10.1177/0146167295213003

Eccles, J. S., & Wigfield, A. (2020). From expectancy-value theory to situated expectancy-value theory: A developmental, social cognitive, and sociocultural perspective on motivation. *Contemporary Educational Psychology*, *61*. https://doi.org/10.1016/j.cedpsych.2020.101859

Education Select Committee. (2024). Call for Evidence. Solving the SEND Crisis. https://committees.parliament.uk/call-for-evidence/3517

Eichorn, D. H. (1975). Asynchronizations in adolescent development. In S. E. Dragastin & G. H. Elder (Eds.), *Adolescence in the life cycle: Psychological change and social context. Hemisphere.* (pp. xi, 324–xi, 324). Hemisphere.

Ellis, H. (1934). *Man and Woman. A Study of Human Secondary and Tertiary Sexual Characters.* (8th ed.). W. Heinemann Ltd.

Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-National Patterns of Gender Differences in Mathematics: A Meta-Analysis. *Psychological Bulletin*, *136*(1), 103–127. https://doi.org/10.1037/a0018053

Ermisch, J., & Bono, E. D. (2010). Education Mobility in England. http://www.suttontrust.com/reports/Education_mobility_in_england.pdf

Erola, J., Jalonen, S., & Lehti, H. (2016). Parental education, class and income over early life course and children's achievement. *Research in Social Stratification and Mobility*, *44*, 33–43. https://doi.org/10.1016/j.rssm.2016.01.003

Ethnicity Facts and Figures Service. (2023a). Entry rates into higher education. Retrieved March 9, 2025, from https://www.ethnicity-facts-figures.service.gov.uk/education-skills-and-training/higher-education/entry-rates-into-higher-education/latest/

Ethnicity Facts and Figures Service. (2023b). Income distribution. Retrieved February 27, 2025, from https://www.ethnicity-facts-figures.service.gov.uk/work-pay-and-benefits/pay-and-income/income-distribution/latest/

Fagence, S., & Hansom, J. (2018). Influence of finance on higher education decision-making. https://assets.publishing.service.gov.uk/media/5ab3b40840f0b65bb5842993/Influence_of_finance_on_higher_education_decision-making.pdf

Feingold, A. (1994). Gender differences in variability in intellectual abilities: A cross-cultural perspective. *Sex Roles*, *30*(1), 81–92. https://doi.org/10.1007/BF01420741

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., Tomasello, M., Mervis, C. B., & Stiles, J. (1994). Variability in Early Communicative Development. *Monographs of the Society for Research in Child Development*, *59*(5), i–185. https://doi.org/10.2307/1166093

Fischbach, A., Keller, U., Preckel, F., & Brunner, M. (2013). PISA proficiency scores predict educational outcomes. *Learning and Individual Differences*, *24*, 63–72. https://doi.org/https://doi.org/10.1016/j.lindif.2012.10.012

Flouri, E., & Hawkes, D. (2008). Ambitious mothers - successful daughters: Mothers' early expectations for children's education and children's earnings and sense of control in adult life. *British Journal of Educational Psychology*, *78*(3), 411–433. https://doi.org/10.1348/000709907X251280

Francis-Devine, B. (2024). *Income Inequality in the UK* (tech. rep. No. CBP-7484). House of Commons. London. Retrieved February 27, 2025, from https://researchbriefings.files.parliament.uk/documents/CBP-7484/CBP-7484.pdf

Frawley, D., Banks, J., & McCoy, S. (2014). Resource allocation for students with special educational needs and disabilities. In J. Cullinan, S. Lyons, & B. Nolan (Eds.), *The Economics of Disability*. Manchester University Press. https://doi.org/10.7765/9781847799814.00012

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School Engagement: Potential of the Concept, State of the Evidence. *Review of Educational Research*, *74*(1), 59–109. https://doi.org/10.3102/00346543074001059

Fredricks, J. A., & McColskey, W. (2012). The Measurement of Student Engagement: A Comparative Analysis of Various Methods and Student Self-report Instruments. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of Research on Student Engagement* (pp. 763–782). Springer US. https://doi.org/10.1007/978-1-4614-2018-7_37

Fredricks, J. A., Parr, A. K., Amemiya, J. L., Wang, M.-T., & Brauer, S. (2019). What Matters for Urban Adolescents' Engagement and Disengagement in School: A Mixed-Methods Study. *Journal of Adolescent Research*, *34*(5), 491–527. https://doi.org/10.1177/0743558419830638

Frenzel, A. C., Goetz, T., Pekrun, R., & Watt, H. M. G. (2010). Development of mathematics interest in adolescence: Influences of gender, family and school context. *Journal of Research on Adolescence*, *20*(2), 507–537. https://doi.org/10.1111/j.1532-7795.2010.00645.x

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 Jr., R. G., & 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

Fumagalli, L. (2019). Short and long term barriers to university education in England. What do expectation data tell us? https://www.iser.essex.ac.uk/wp-content/uploads/files/projects/information-expectations-transitions-he/barriers-university-education-england.pdf

Gallagher, A. L., Galvin, R., Robinson, K., Murphy, C.-A., Conway, P. F., & Perry, A. (2020). The characteristics, life circumstances and self-concept of 13 year olds with and without disabilities in Ireland: A secondary analysis of the Growing Up in Ireland (GUI) study. *PloS One*, *15*(3). https://doi.org/10.1371/journal.pone.0229599

Galván, A. (2013). The Teenage Brain: Sensitivity to Rewards. *Current Directions in Psychological Science*, *22*(2), 88–93. https://doi.org/10.1177/0963721413480859

Ge, X., Conger, R. D., & Elder, G. H. (1996). Coming of Age Too Early: Pubertal Influences on Girls' Vulnerability to Psychological Distress. *Child Development*, *67*(6), 3386–3400. http://www.jstor.org/stable/1131784

Ge, X., Conger, R. D., & Elder, G. H. (2001). Pubertal transition, stressful life events, and the emergence of gender differences in adolescent depressive symptoms. *Developmental Psychology*, *37*(3), 404–417. https://doi.org/10.1037//0012-1649.37.3.404

Ge, X., Kim, I. J., Brody, G. H., Conger, R. D., Simons, R. L., Gibbons, F. X., & Cutrona, C. E. (2003). It's about timing and change: Pubertal transition effects on symptoms of major depression among African American youths. *Developmental Psychology*, *39*(3), 430–439. https://doi.org/10.1037/0012-1649.39.3.430

Geary, D. C., Hoard, M. K., & Nugent, L. (2023). Boys' advantage in solving algebra word problems is mediated by spatial abilities and mathematics anxiety. *Developmental Psychology*, *59*(3), 413–430. https://doi.org/10.1037/dev0001450

Gelman, A. (2007). Struggles with Survey Weighting and Regression Modeling. *Statistical Science*, *22*(2), 153–164. https://doi.org/10.1214/088342306000000691

Giedd, J., Blumenthal, J., Jeffries, N. O., Castellanos, F., Liu, H., Zijdenbos, A., Paus, T., Evans, A., & Rapoport, J. (1999). Brain Development during Childhood and Adolescence: A Longitudinal MRI Study. *Nature*, *2*, 861–863.

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

Gill, D., Del Greco M, F., Rawson, T. M., Sivakumaran, P., Brown, A., Sheehan, N. A., & Minelli, C. (2017). Age at Menarche and Time Spent in Education: A Mendelian Randomization Study. *Behavior Genetics*, *47*(5), 480–485. https://doi.org/10.1007/s10519-017-9862-2

GL Assessment. (2020). Read All Ahout It: Why reading success is key to GCSE success. https://reports.gl-assessment.co.uk/whyreading/

Goering, M., Albright, M. G., & Mrug, S. (2023). The Effects of Pubertal Timing on Academic Performance in Adolescence and Career Success in Adulthood: Evidence from a 16-year Longitudinal Study. *Journal of Youth and Adolescence*, *52*(9), 1769–1787. https://doi.org/10.1007/s10964-023-01814-6

Goering, M., & Mrug, S. (2022). The Distinct Roles of Biological and Perceived Pubertal Timing in Delinquency and Depressive Symptoms from Adolescence to Adulthood. *Journal of Youth and Adolescence*, *51*(11), 2092–2113. https://doi.org/10.1007/s10964-022-01657-7

Goetz, T., Bieg, M., Lüdtke, O., Pekrun, R., & Hall, N. C. (2013). Do Girls Really Experience More Anxiety in Mathematics? *Psychological Science*, *24*(10), 2079–2087. https://doi.org/10.1177/0956797613486989

Gorard, S. (2003). *Quantitative Methods in Social Science Research*. Bloomsbury Publishing Plc.

Gortazar, L., Martinez de Lafuente, D., & Vega-Bayo, A. (2022). Comparing teacher and external assessments: Are boys, immigrants, and poorer students undergraded? *Teaching and Teacher Education*, *115*, 103725. https://doi.org/10.1016/j.tate.2022.103725

Gottfried, M. A. (2009). Excused Versus Unexcused: How Student Absences in Elementary School Affect Academic Achievement. *Educational Evaluation and Policy Analysis*, *31*(4), 392–415. https://doi.org/10.3102/0162373709342467

GOV.UK. (n.d.). The national curriculum. Retrieved April 9, 2025, from https://www.gov.uk/national-curriculum/key-stage-3-and-4

Graber, J. A., Seeley, J. R., Brooks-Gunn, J., & Lewinsohn, P. M. (2004). Is Pubertal Timing Associated With Psychopathology in Young Adulthood? *Journal of the American Academy of Child & Adolescent Psychiatry*, *43*(6), 718–726. http://www.sciencedirect.com/science/article/pii/S0890856709613216

Graber, J. A., Lewinsohn, P. M., Seeley, J. R., & Brooks-Gunn, J. (1997). Is Psychopathology Associated With the Timing of Pubertal Development? *Journal of the American*

Academy of Child & Adolescent Psychiatry, *36*(12), 1768–1776. https://doi.org/10.1097/00004583-199712000-00026

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

Green, J., Liem, G. A. D., Martin, A. J., Colmar, S., Marsh, H. W., & McInerney, D. (2012). Academic motivation, self-concept, engagement, and performance in high school: Key processes from a longitudinal perspective. *Journal of Adolescence*, *35*(5), 1111–1122. https://doi.org/10.1016/j.adolescence.2012.02.016

Greenland, S., & Finkle, W. D. (1995). A Critical Look at Methods for Handling Missing Covariates in Epidemiologic Regression Analyses. *American Journal of Epidemiology*, *142*(12), 1255–1264. https://doi.org/10.1093/oxfordjournals.aje.a117592

Grimm, K. J., Ram, N., & Estabrook, R. (2016). Chapter 8 - Multivariate Growth Models and Dynamic Predictors [ISBN: 0165025409]. In *Growth Modeling: Structural Equation and Mulitlevel Modeling Approaches* (Vol. 33). Guildord Press.

Grissom, N. M., & Reyes, T. M. (2019). Let's call the whole thing off: Evaluating gender and sex differences in executive function. *Neuropsychopharmacology*, *44*(1), 86–96. https://doi.org/10.1038/s41386-018-0179-5

Haddad, C., Sacre, H., Zeenny, R. M., Hajj, A., Akel, M., Iskandar, K., & Salameh, P. (2022). Should samples be weighted to decrease selection bias in online surveys during the COVID-19 pandemic? Data from seven datasets. *BMC Medical Research Methodology*, *22*, 63. https://doi.org/10.1186/s12874-022-01547-3

Hallfors, D., Vevea, J. L., Iritani, B., Cho, H., Khatapoush, S., & Saxe, L. (2002). Truancy, Grade Point Average, and Sexual Activity: A Meta-Analysis of Risk Indicators for Youth Substance Use. *Journal of School Health*, *72*(5), 205–211. https://doi.org/10.1111/j.1746-1561.2002.tb06548.x

Halpern, D. F. (2012). *Sex Differences in Cognitive Abilities* (4th). 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. (2019). Sex Differences in Intelligence. In R. Sternberg (Ed.), *The Cambridge Handbook of Intelligence* (pp. 317–345). Cambridge University Press. https://doi.org/10.1017/9781108770422.015

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

Hannah, E. F., & Topping, K. (2013). The transition from primary to secondary school: Perspectives of students with autism spectrum disorder and their parents. *International Journal of Special Education*, *28*, 145–157.

Hannah, E. F., & Topping, K. J. (2012). Anxiety Levels in Students with Autism Spectrum Disorder Making the Transition from Primary to Secondary School. *Education and Training in Autism and Developmental Disabilities*, *47*(2), 198–209. https://www.jstor.org/stable/23880100

Hansen, D., & Jessop, N. (2017). A Context for Self-Determination and Agency: Adolescent Developmental Theories. In M. L. Wehmeyer, K. A. Shogren, T. D. Little, & S. J. Lopez (Eds.), *Development of Self-Determination Through the Life-Course* (pp. 27–46). Springer Netherlands. https://doi.org/10.1007/978-94-024-1042-6_3

Hansen, K., & Jones, E. M. (2011). Ethnicity and gender gaps in early childhood. *British Educational Research Journal*, *37*(6), 973–991. https://www.jstor.org/stable/23077019

Harlen, W. (2005). Trusting teachers' judgement: Research evidence of the reliability and validity of teachers' assessment used for summative purposes. *Research Papers in Education*, *20*(3), 245–270. https://doi.org/10.1080/02671520500193744

Hartley, B. L., & Sutton, R. M. (2013). A Stereotype Threat Account of Boys' Academic Underachievement. *Child Development*, *84*(5), 1716–1733. https://doi.org/10.1111/cdev.12079

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

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

Herlitz, A., & Rehnman, J. (2008). Sex Differences in Episodic Memory: *Current Directions in Psychological Science*, *17*(1). https://doi.org/10.1111/j.1467-8721.2008.00547.x

Herlitz, A., Reuterskiöld, L., Lovén, J., Thilers, P. P., & Rehnman, J. (2013). Cognitive sex differences are not magnified as a function of age, sex hormones, or puberty development during early adolescence. *Developmental Neuropsychology*, *38*(3), 167–179. https://doi.org/10.1080/87565641.2012.759580

Heyder, A., Kessels, U., & Steinmayr, R. (2017). Explaining academic-track boys' underachievement in language grades: Not a lack of aptitude but students' motivational beliefs and parents' perceptions? *British Journal of Educational Psychology*, *87*(2), 205–223. https://doi.org/10.1111/bjep.12145

Hill, J. P., & Lynch, M. E. (1983). The Intensification of Gender-Related Role Expectations during Early Adolescence. In J. Brooks-Gunn & A. C. Petersen (Eds.), *Girls at Puberty: Biological and Psychosocial Perspectives* (pp. 201–228). Springer US. https://doi.org/10.1007/978-1-4899-0354-9_10

Hillman, N., & Robinson, N. (2016). Boys to Men: The underachievement of young men in higher education – and how to start tackling it. https://www.hepi.ac.uk/wp-content/uploads/2016/05/Boys-to-Men.pdf

Hoffmann, J. P. (2020). Academic Underachievement and Delinquent Behavior. *Youth & Society*, *52*(5), 728–755. https://doi.org/10.1177/0044118X18767035

Hoyt, L. T., Niu, L., Pachucki, M. C., & Chaku, N. (2020). Timing of puberty in boys and girls: Implications for population health. *SSM - Population Health*, *10*, 100549. https://doi.org/10.1016/j.ssmph.2020.100549

Hsieh, T.-y., 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*, 101974. https://doi.org/10.1016/j.cedpsych.2021.101974

Huang, C. (2011). Self-concept and academic achievement: A meta-analysis of longitudinal relations. *Journal of School Psychology*, *49*(5), 505–528. https://doi.org/10.1016/j.jsp.2011.07.001

Huang, C. (2013). Gender differences in academic self-efficacy: A meta-analysis. *European Journal of Psychology of Education*, *28*(1), 1–35. https://doi.org/10.1007/s10212-011-0097-y

Huang, J., Xie, X., Pan, Y., Li, G., Zhang, F., & Cui, N. (2022). Trajectories of Self-Esteem Development for Adolescence Based on China Family Panel Studies: A Piecewise Growth Mixture Model Analysis. *Sage Open*, *12*(2), 21582440221086611. https://doi.org/10.1177/21582440221086611

Hughes, L. A., Banks, P., & Terras, M. M. (2013). Secondary school transition for children with special educational needs: A literature review. *Support for Learning*, *28*(1), 24–34. https://doi.org/10.1111/1467-9604.12012

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

Hulin, C., Netemeyer, R., & Cudeck, R. (2001). Can a Reliability Coefficient Be Too High? *Journal of Consumer Psychology*, *10*, 55–58. https://doi.org/10.2307/1480474

Hyde, J. S. (2005). The gender similarities hypothesis. *American Psychologist*, *60*(6), 581–592. https://doi.org/10.1037/0003-066X.60.6.581

Institute of Education. (2019). Millennium Cohort Study. A guide to the linked education administrative datasets. https://cls.ucl.ac.uk/wp-content/uploads/2019/07/User-guide-to-the-Linked-Education-Datasets-2019-Version2_May2019.pdf

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. https://www.jstor.org/stable/3696372

Jansen, M., Boda, Z., & Lorenz, G. (2022). Social Comparison Effects on Academic Self-Concepts—Which Peers Matter Most? *Developmental Psychology*, *58*(8), 1541–1556. https://doi.org/10.1037/dev0001368

Jerrim, J. (2019, October). How do GCSE grades relate to PISA scores? Retrieved August 26, 2024, from https://blogs.ucl.ac.uk/ioe/2019/10/29/how-do-gcse-grades-relate-to-pisa-scores/

Jerrim, J. (2021). PISA 2018 in England, Northern Ireland, Scotland and Wales: Is the data really representative of all four corners of the UK? *Review of Education*, *9*(3), e3270. https://doi.org/10.1002/rev3.3270

Jerrim, J., & Palma Carvajal, M. (2025). What happens to bright 5-year-olds from poor backgrounds? Longitudinal evidence from the Millennium Cohort Study. *Research in Social Stratification and Mobility*, *97*. https://doi.org/10.1016/j.rssm.2025.101038

Jindal-Snape, D., Hannah, E. F., Cantali, D., Barlow, W., & MacGillivray, S. (2020). Systematic literature review of primary–secondary transitions: International research. *Review of Education*, *8*(2), 526–566. https://doi.org/10.1002/rev3.3197

Jing, Y., Lin, Y., Yu, L., & Zhang, Z. (2021). A Review Study of the "Boy Crisis". *Proceedings of the 2021 International Conference on Public Relations and Social Sciences (ICPRSS 2021)*, 797–801. https://doi.org/10.2991/assehr.k.211020.259

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. https://www.jstor.org/stable/40212273

Jones, E. M., & Ketende, S. (2010). *User Guide to Analysing MCS Data Using SPSS* (tech. rep.). https://cls.ucl.ac.uk/wp-content/uploads/2017/06/Complex-Samples-in-SPSS.pdf

Jormanainen, E., Fröjd, S., Marttunen, M., & Kaltiala-Heino, R. (2014). Is pubertal timing associated with involvement in bullying in middle adolescence? *Health Psychology and Behavioral Medicine*, *2*(1), 144–159. https://doi.org/10.1080/21642850.2014.881259

Kashikar, L., Soemers, L., Lüke, T., & Grosche, M. (2023). Does the 'Learning Disability' label lower teachers' performance expectations? *Social Psychology of Education*, *26*(4), 971–1000. https://doi.org/10.1007/s11218-023-09775-1

Kashikar, L., Soemers, L., Lüke, T., & Grosche, M. (2024). Influence of the 'Learning Disability' label on teachers' performance expectations—a matter of attitudes towards inclusion? *Journal of Research in Special Educational Needs*, *24*(3), 696–712. https://doi.org/10.1111/1471-3802.12664

Kaufman, S. B. (2007). Sex differences in mental rotation and spatial visualization ability: Can they be accounted for by differences in working memory capacity? *Intelligence*, *35*(3), 211–223. https://doi.org/10.1016/j.intell.2006.07.009

Kenney-Benson, G. A., Pomerantz, E. M., Ryan, A., & 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

Kessels, U., Heyder, A., Latsch, M., & Hannover, B. (2014). How gender differences in academic engagement relate to students' gender identity. *Educational Research*, *56*(2), 220–229. https://doi.org/10.1080/00131881.2014.898916

Ketende, S., & Jones, E. M. (2011). *User Guide to Analysis MCS Data Using STATA* (tech. rep.). https://cls.ucl.ac.uk/wp-content/uploads/2017/07/User-Guide-to-Analysing-MCS-Data-using-Stata.pdf

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

Klapwijk, E. T., Goddings, A.-L., Burnett Heyes, S., Bird, G., Viner, R. M., & Blakemore, S.-J. (2013). Increased functional connectivity with puberty in the mentalising network involved in social emotion processing. *Hormones and Behavior*, *64*(2), 314–322. https://doi.org/10.1016/j.yhbeh.2013.03.012

Klimstra, T. A., Hale III, W. W., Raaijmakers, Q. A. W., Branje, S. J. T., & Meeus, W. H. J. (2010). Identity Formation in Adolescence: Change or Stability? *Journal of Youth and Adolescence*, *39*(2), 150–162. https://doi.org/10.1007/s10964-009-9401-4

Kline, R. B. (2016). *Principles and practice of structural equation modeling, 4th ed* [Pages: xvii, 534]. Guilford Press.

Kling, K. C., Noftle, E. E., & Robins, R. W. (2013). Why Do Standardized Tests Underpredict Women's Academic Performance? The Role of Conscientiousness. *Social Psychological and Personality Science*, *4*(5), 600–606. https://doi.org/10.1177/1948550612469038

Koerselman, K., & Pekkarinen, T. (2017). The Timing of Puberty and Gender Differences in Educational Achievement. *IZA Discussion Paper*, (10889).

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

Kollmayer, M., Schober, B., & Spiel, C. (2018). Gender stereotypes in education: Development, consequences, and interventions. *European Journal of Developmental Psychology*. https://doi.org/10.1080/17405629.2016.1193483

Koshy, P., Dockery, A. M., & Seymour, R. (2019). Parental expectations for young people's participation in higher education in Australia. *Studies in Higher Education*, *44*(1), 302–317. https://doi.org/10.1080/03075079.2017.1363730

Kremer, K. P., Vaughn, M. G., & Loux, T. M. (2018). Parent and peer social norms and youth's post-secondary attitudes: A latent class analysis. *Children and Youth Services Review*, *93*, 411–417. https://doi.org/10.1016/j.childyouth.2018.08.026

Kriegbaum, K., Becker, N., & Spinath, B. (2018). The relative importance of intelligence and motivation as predictors of school achievement: A meta-analysis. *Educational Research Review*, *25*, 120–148. https://doi.org/10.1016/j.edurev.2018.10.001

Kuncel, N. R., Credé, M., & Thomas, L. L. (2005). The Validity of Self-Reported Grade Point Averages, Class Ranks, and Test Scores: A Meta-Analysis and Review of the Literature. *Review of Educational Research*, *75*(1), 63–82. https://doi.org/10.3102/00346543075001063

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

Kurz, S. (2021, December). Notes on the Bayesian cumulative probit. https://solomonkurz.netlify.app/blog/2021-12-29-notes-on-the-bayesian-cumulative-probit/

Lam, S.-f., Jimerson, S., Kikas, E., Cefai, C., Veiga, F. H., Nelson, B., Hatzichristou, C., Polychroni, F., Basnett, J., Duck, R., Farrell, P., Liu, Y., Negovan, V., Shin, H., Stanculescu, E., Wong, B. P. H., Yang, H., & Zollneritsch, J. (2012). Do girls and boys perceive themselves as equally engaged in school? The results of an international study from 12 countries. *Journal of School Psychology*, *50*(1), 77–94. https://doi.org/10.1016/j.jsp.2011.07.004

Latsch, M., & Hannover, B. (2014). Smart Girls, Dumb Boys!? *Social Psychology*, *45*(2), 112–126. https://doi.org/10.1027/1864-9335/a000167

Laube, C., & Fuhrmann, D. (2020). Is early good or bad? Early puberty onset and its consequences for learning. *Current Opinion in Behavioral Sciences*, *36*, 150–156. https://doi.org/10.1016/j.cobeha.2020.10.005

Lavy, V., & Megalokonomou, R. (2019). Persistency in Teachers' Grading Bias and Effects on Longer-Term Outcomes: University Admissions Exams and Choice of Field of Study. https://doi.org/10.3386/w26021

Lavy, V., & Sand, E. (2018). On the origins of gender gaps in human capital: Short- and long-term consequences of teachers' biases. *Journal of Public Economics*, *167*, 263–279. https://doi.org/10.1016/j.jpubeco.2018.09.007

Lee, M. W. (2021). Summer 2021 student-level equalities analysis - GCSE and A level. Retrieved May 4, 2025, from https://www.gov.uk/government/publications/analysis-of-results-a-levels-and-gcses-summer-2021/summer-2021-student-level-equalities-analysis-gcse-and-a-level

Lee, T., & Shi, D. (2021). A comparison of full information maximum likelihood and multiple imputation in structural equation modeling with missing data. *Psychological Methods*, *26*(4), 466–485. https://doi.org/10.1037/met0000381

Lei, H., Cui, Y., & Zhou, W. (2018). Relationships between student engagement and academic achievement: A meta-analysis. *Social Behavior and Personality: an international journal*, *46*, 517–528. https://doi.org/10.2224/sbp.7054

Lenroot, R. K., Gogtay, N., Greenstein, D. K., Wells, E. M., Wallace, G. L., Clasen, L. S., Blumenthal, J. D., Lerch, J., Zijdenbos, A. P., Evans, A. C., Thompson, P. M., & Giedd, J. N. (2007). Sexual dimorphism of brain developmental trajectories during childhood and adolescence. *NeuroImage*, *36*(4), 1065–1073. https://doi.org/10.1016/j.neuroimage.2007.03.053

Li, Y., & Lerner, R. M. (2011). Trajectories of school engagement during adolescence: Implications for grades, depression, delinquency, and substance use. *Developmental Psychology*, *47*(1), 233–247. https://doi.org/10.1037/a0021307

Liem, G. A. D., Marsh, H. W., Martin, A. J., McInerney, D. M., & Yeung, A. S. (2013). The Big-Fish-Little-Pond Effect and a National Policy of Within-School Ability Streaming: Alternative Frames of Reference. *American Educational Research Journal*, *50*(2), 326–370. https://doi.org/10.3102/0002831212464511

Lim, A. J.-M., & Cheung, M. W.-L. (2022). Evaluating FIML and multiple imputation in joint ordinal-continuous measurements models with missing data. *Behavior Research Methods*, *54*(3), 1063–1077. https://doi.org/10.3758/s13428-021-01582-w

Lindberg, J. L., Petersen, S. M., & Hyde, J. S. (2010). New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin*, *136*(6), 1123–1135.

Liu, W. C., Wang, C. K. J., & Parkins, E. J. (2005). A longitudinal study of students' academic self-concept in a streamed setting: The Singapore context. *The British Journal of Educational Psychology*, *75*(Pt 4), 567–586. https://doi.org/10.1348/000709905X42239

Lo Bello, S., & Morchio, I. (2022). Like father, like son: Occupational choice, intergenerational persistence and misallocation. *Quantitative Economics*, *13*(2), 629–679. https://doi.org/10.3982/QE1375

Lochner, L., & Moretti, E. (2004). The Effect of Education on Crime: Evidence from Prison Inmates, Arrests, and Self-Reports. *The American Economic Review*, *94*(1), 155–189. Retrieved March 19, 2024, from https://www.jstor.org/stable/3592774

Lumley, T. (2010). *Complex Surveys: A Guide to Analysis Using R*. John Wiley & Sons, Inc.

Lumley, T. (2024). Survey: Analysis of complex survey samples. R package version 4.4.

Lundberg, S. (2020). Educational Gender Gaps. *Southern economic journal*, *87*(2), 416–439. https://doi.org/10.1002/soej.12460

Machin, S., Marie, O., & Vujić, S. (2011). The Crime Reducing Effect of Education*. *The Economic Journal*, *121*(552), 463–484. https://doi.org/10.1111/j.1468-0297.2011.02430.x

Machin, S., & Pekkarinen, T. (2008). Global Sex Differences in Test Score Variability. *Science*, *322*(5906), 1331–1332. https://doi.org/10.1126/science.1162573

Magnusson, D., Stattin, H., & Allen, V. L. (1985). Biological maturation and social development: A longitudinal study of some adjustment processes from mid-adolescence to adulthood. *Journal of Youth and Adolescence*, *14*(4), 267–283. https://doi.org/10.1007/BF02089234

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

Makowski, D., Ben-Shachar, M. S., Chen, S. H. A., & Lüdecke, D. (2019). Indices of Effect Existence and Significance in the Bayesian Framework. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.02767

Makowski, D., Wiernik, B. M., Patil, I., Lüdeke, D., & Ben-Shachar, M. S. (2022). Correlation: Methods for correlation analysis [R package].

Malouff, J., & Thorsteinsson, E. (2016). Bias in grading: A meta-analysis of experimental research findings. *Australian Journal of Education*, *60*. https://doi.org/10.1177/0004944116664618

Marceau, K., Ram, N., Houts, R. M., Grimm, K. J., & Susman, E. J. (2011). Individual differences in boys' and girls' timing and tempo of puberty: Modeling development with nonlinear growth models. *Developmental Psychology*, *47*(5), 1389–1409. https://doi.org/10.1037/a0023838

Marsh, H. W. (1987). The big-fish-little-pond effect on academic self-concept. *Journal of Educational Psychology*, *79*(3), 280. https://doi.org/10.1037/0022-0663.79.3.280

Marsh, H. W., Barnes, J., Cairns, L., & Tidman, M. (1984). Self-Description Questionnaire: Age and sex effects in the structure and level of self-concept for preadolescent children. *Journal of Educational Psychology*, *76*(5), 940–956. https://doi.org/10.1037/0022-0663.76.5.940

Marsh, H. W., Craven, R. G., & Debus, R. (1991). Self-concepts of young children 5 to 8 years of age: Measurement and multidimensional structure. *Journal of Educational Psychology*, *83*(3), 377–392. https://doi.org/10.1037/0022-0663.83.3.377

Marsh, H. W., & Hau, K.-T. (2003). Big Fish Little Pond Effect on Academic Self-concept: A cross-cultural (26-country) test of the negative effects of academically selective schools. *The American Psychologist*, *58*(5), 364*761–21. https://doi.org/10.1037/0003-066x.58.5.364

Marsh, H. W., Parker, P. D., Guo, J., Basarkod, G., Niepel, C., & Van Zanden, B. (2021). Illusory gender-equality paradox, math self-concept, and frame-of-reference effects: New integrative explanations for multiple paradoxes. *Journal of Personality and Social Psychology*, *121*(1), 168–183. https://doi.org/10.1037/pspp0000306

Marsh, H. W., Pekrun, R., Murayama, K., Arens, A. K., Parker, P. D., Guo, J., & Dicke, T. (2018). An integrated model of academic self-concept development: Academic self-concept, grades, test scores, and tracking over 6 years. *Developmental Psychology*, *54*(2), 263–280. https://doi.org/10.1037/dev0000393

Marsh, H. W., Seaton, M., Trautwein, U., Lüdtke, O., Hau, K.-T., O'Mara, A. J., & Craven, R. G. (2008). The big-fish–little-pond-effect stands up to critical scrutiny: Implications for theory, methodology, and future research [ISBN: 1040-726X Publisher: Springer]. *Educational Psychology Review*, *20*, 319–350. https://doi.org/10.1007/s10648-008-9075-6

Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2005). Academic self-concept, interest, grades and standardized test scores: Reciprocal effects models of causal ordering. *Child Development*, *76*(2), 397–416. https://doi.org/10.1111/j.1467-8624.2005.00853.x

Marshall, W. A., & Tanner, J. M. (1969). Variations in Pattern of Pubertal Changes in Girls. *Archives of Disease in Childhood*, *44*, 291–303. https://doi.org/10.1136/adc.44.235.291

Marshall, W. A., & Tanner, J. M. (1970). Variations in the Pattern of Pubertal Changes in Boys. *Archives of Disease in Childhood*, *45*(239), 13–23. Retrieved September 3, 2024, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2020414/

Martin, A. J. (2012). Motivation and Engagement. https://eric.ed.gov/?id=ED532692

Martin, A. J., Balzer, B., Garden, F., Handelsman, D. J., Hawke, C., Luscombe, G., Paxton, K., Skinner, S. R., & 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*, 102213. https://doi.org/10.1016/j.lindif.2022.102213

Martin, A. J., Strnadová, I., O'Neill, S. C., & Cumming, T. M. (2017). The role of perceived competence in the lives of children with ADHD, emotional and behavioral disorder, learning disability, and developmental disability: A positive psychology perspective. In

Self: Driving Positive Psychology and Well-Being (pp. 1–26). IAP Information Age Publishing.

Martin, D. J., & Hoover, H. D. (1987). Sex Differences in Educational Achievement: A Longitudinal Study. *The Journal of Early Adolescence*, *7*(1), 65–83. https://doi.org/10.1177/0272431687071007

Matějů, P., & Smith, M. L. (2015). Are boys that bad? Gender gaps in measured skills, grades and aspirations in Czech elementary schools. *British Journal of Sociology of Education*, *36*(6), 871*895. https://doi.org/10.1080/01425692.2013.874278

Mattern, K. D., & Patterson, B. F. (2014). Synthesis of Recent SAT Validity Findings: Trend Data over Time and Cohorts.

Mayes, S. D., Castagna, P. J., & Waschbusch, D. A. (2020). Sex Differences in Externalizing and Internalizing Symptoms in ADHD, Autism, and General Population Samples. *Journal of Psychopathology and Behavioral Assessment*, *42*(3), 519–526. https://doi.org/10.1007/s10862-020-09798-4

McArdle, J. J., & Epstein, D. (1987). Latent growth curves within developmental structural equation models. *Child Development*, *58*(1), 110–133.

McDonald, R. P. (2013). *Test theory: A unified treatment* (1st Edition). Psychology Press. https://doi.org/10.4324/9781410601087

McDougal, E., Riby, D. M., & Hanley, M. (2020). Profiles of academic achievement and attention in children with and without Autism Spectrum Disorder. *Research in Developmental Disabilities*, *106*, 103749. https://doi.org/10.1016/j.ridd.2020.103749

McElreath, R. (2020). *Statistical Rethinking* (2nd). doi.org/10.1201/9780429029608

McLatchie, Y., & Vehtari, A. (2023, September). Efficient estimation and correction of selection-induced bias with order statistics. http://arxiv.org/abs/2309.03742

Mead, S. (2006). *The evidence suggests otherwise: The truth about boys and girls* (tech. rep. No. 378705) (Issue: June). Education Sector.

Meimoun, E., Bonnot, V., Berenguer, J., & Aelenei, C. (2024). How are gender counter-stereotypical adolescents evaluated by their peers? Investigating the backlash effect in French schools. *Social Psychology of Education*, *27*(3), 833–857. https://doi.org/10.1007/s11218-023-09807-w

Mendle, J. (2014). Beyond Pubertal Timing: New Directions for Studying Individual Differences in Development. *Current Directions in Psychological Science*, *23*(3), 215–219. https://doi.org/10.1177/0963721414530144

Mendle, J., & Ferrero, J. (2012). Detrimental psychological outcomes associated with pubertal timing in adolescent boys. *Developmental Review*, *32*(1), 49–66. https://doi.org/10.1016/j.dr.2011.11.001

Mendle, J., Harden, K. P., Brooks-Gunn, J., & Graber, J. A. (2010). Development's tortoise and hare: Pubertal timing, pubertal tempo, and depressive symptoms in boys and girls. *Developmental Psychology*, *46*(5), 1341–1353. https://doi.org/10.1037/a0020205

Mendle, J., & Koch, M. K. (2019). The Psychology of Puberty: What Aren't We Studying That We Should? *Child Development Perspectives*, *13*(3), 166–172. https://doi.org/10.1111/cdep.12333

Mendle, J., Turkheimer, E., & Emery, R. E. (2007). Detrimental Psychological Outcomes Associated with Early Pubertal Timing in Adolescent Girls. *Developmental Review*, *27*(2), 151–171. https://doi.org/10.1016/j.dr.2006.11.001

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

Miratrix, L. W., Sekhon, J. S., Theodoridis, A. G., & Campos, L. F. (2018). Worth Weighting? How to Think About and Use Weights in Survey Experiments. *Political Analysis*, *26*(3), 275–291. https://doi.org/10.1017/pan.2018.1

Möller, J., & Marsh, H. W. (2013). Dimensional comparison theory. *Psychological Review*, *120*(3), 544–560. https://doi.org/10.1037/a0032459

Möller, J., Pohlmann, B., Köller, O., & Marsh, H. W. (2009). A meta-analytic path analysis of the internal/ external frame of reference model of academic achievement and academic self-concept. *Review of Educational Research*, *79*(3), 1129–1167. https://doi.org/10.3102/0034654309337522

Montacute, R. (2018). Potential for Success: Fulfilling the promise of highly able students in secondary schools. *The Sutton Trust*. https://www.suttontrust.com/our-research/potential-for-success-schools-high-attainers/

Moore, S., Harden, K., & Mendle, J. (2014). Pubertal Timing and Adolescent Sexual Behavior in Girls. *Developmental psychology*, *50*. https://doi.org/10.1037/a0036027

Moreira, P. A. S., Bilimória, H., Pedrosa, C., de Fátima Pires, M., de Jesus Cepa, M., de Deus Mestre, M., Ferreira, M., & Serra, N. (2015). Engagement with school in students with special educational needs. *International Journal of Psychology & Psychological Therapy*, *15*(3), 361–375.

Mulvey, K. L., & Killen, M. (2015). Challenging Gender Stereotypes: Resistance and Exclusion [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cdev.12317]. *Child Development*, *86*(3), 681–694. https://doi.org/10.1111/cdev.12317

Murray, C., & Greenberg, M. T. (2001). Relationships with teachers and bonds with school: Social emotional adjustment correlates for children with and without disabilities. *Psychology in the Schools*, *38*(1), 25–41. https://doi.org/10.1002/1520-6807(200101)38:1<25::AID-PITS4>3.0.CO;2-C

Murray, C., & Greenberg, M. T. (2006). Examining the Importance of Social Relationships and Social Contexts in the Lives of Children With High-Incidence Disabilities. *The Journal of Special Education*, *39*(4), 220–233. https://doi.org/10.1177/00224669060390040301

Muthén, L., & Muthén, B. O. (1998). *Mplus User's Guide.* (6th ed.).

Muyeed, A. Z. (2008). Adolescent Health and Development. In W. Kirch (Ed.), *Encyclopedia of Public Health* (pp. 14–17). Springer Netherlands. https://doi.org/10.1007/978-1-4020-5614-7_57

National Literacy Trust. (2011). *What do adult literacy levels mean?* (Tech. rep.). https://literacytrust.org.uk/parents-and-families/adult-literacy/what-do-adult-literacy-levels-mean/

Nelson, E. E., Leibenluft, E., Tone, E., & Pine, D. (2005). The social re-orientation of adolescence: A neuroscience perspective on the process and its relation to psychopathology. *Psychological Medicine*, *35*, 163–74. https://doi.org/10.1017/S0033291704003915

Nelson, E. E., Jarcho, J. M., & Guyer, A. E. (2016). Social re-orientation and brain development: An expanded and updated view. *Developmental Cognitive Neuroscience*, *17*, 118–127. https://doi.org/10.1016/j.dcn.2015.12.008

Newcombe, N. (1989). The Development of Spatial Perspective Taking. *Advances in Child Development and Behaviour*, *22*, 203–247. https://doi.org/10.1016/s0065-2407(08)60415-2

Oakley, C. M., Pekrun, R., & Stoet, G. (2024). Sex Differences of School Grades in Childhood and Adolescence: A Longitudinal Analysis. *Intelligence*, *107*. https://doi.org/10.1016/j.intell.2024.101857

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

O'Donnell, K. C., & Reschly, A. L. (2020). Student Engagement and Learning: Attention, Behavioral, and Emotional Difficulties in School. In *Handbook of Educational Psychology and Students with Special Needs*. Routledge.

OECD. (2023, September). *Education at a Glance 2023 Sources, Methodologies and Technical Notes*. https://doi.org/10.1787/d7f76adc-en

Office for National Statistics. (2018, October). *Population estimate by output areas, electoral, health and other geographies, England and Wales: Mid 2017* (tech. rep.).

Ofqual. (2018, February). GCSE 9 to 1 grades: A brief guide for parents. https://ofqual.blog.gov.uk/2018/03/02/gcse-9-to-1-grades-a-brief-guide-for-parents/

O'Mara-Eves, A., Green, J., & Marsh, H. (2006). Administering self-concept interventions in schools: No training necessary? A meta-analysis. *International Education Journal*, *7*(4), 524–533.

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

Oostdam, R., van Diepen, M., Zijlstra, B., & Fukkink, R. (2024). Effects of the COVID-19 school lockdowns on language and math performance of students in elementary schools: Implications for educational practice and reducing inequality. *European Journal of Psychology of Education*, *39*(1), 129–149. https://doi.org/10.1007/s10212-023-00679-4

Ortiz-Gervasi, L. (2020). What shape great expectations? Gender and social-origin effects on expectation of university graduation. *Research in Social Stratification and Mobility*, *69*, 100527. https://doi.org/10.1016/j.rssm.2020.100527

Paikoff, R. L., & Brooks-Gunn, J. (1991). Do parent-child relationships change during puberty? *Psychological Bulletin*, *110*(1), 47–66. https://doi.org/10.1037/0033-2909.110.1.47

Paus, T., Wong, A. P.-Y., Syme, C., & Pausova, Z. (2017). Sex differences in the adolescent brain and body: Findings from the saguenay youth study. *Journal of Neuroscience Research*, *95*(1-2), 362–370. https://doi.org/10.1002/jnr.23825

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*.

Pekrun, R. (2024). Overcoming Fragmentation in Motivation Science: Why, When, and How Should We Integrate Theories? *Educational Psychology Review*, *36*(1), 27. https://doi.org/10.1007/s10648-024-09846-5

Pekrun, R., & Linnenbrink-Garcia, L. (2012). Academic Emotions and Student Engagement. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of Research on Student Engagement* (pp. 259–282). Springer US. https://doi.org/10.1007/978-1-4614-2018-7_12

Pennington, C. R., Kaye, L. K., Qureshi, A. W., & Heim, D. (2018). Controlling for Prior Attainment Reduces the Positive Influence that Single-Gender Classroom Initiatives Exert on High School Students' Scholastic Achievements. *Sex Roles*, *78*(5-6), 385–393. https://doi.org/10.1007/s11199-017-0799-y

Peper, J. S., & Dahl, R. E. (2013). The Teenage Brain: Surging Hormones-Brain-Behavior Interactions During Puberty. *Current Directions in Psychological Science*, *22*(2), 134–139. https://doi.org/10.1177/0963721412473755

Perinetti Casoni, V., & Barg, K. (2023). Ethnic biases in English state primary schools. https://bpb-eu-w2.wpmucdn.com/blogs.bristol.ac.uk/dist/f/860/files/2023/10/Ethnic-biases-in-English-state-primary-schools.pdf

Petersen, A. C., Crockett, L. J., Richards, M., & Boxer, A. (1988). A self-report measure of pubertal status: Reliability, validity, and initial norms. *Journal of Youth and Adolescence*, *17*(2), 117–133. https://doi.org/10.1007/BF01537962

Petersen, A. C., & Taylor, B. (1980). The biological approach to adolescence: Biological change and psychological adaptation. In J. Adelson (Ed.), *Handbook of Adolescent Psychology* (pp. 117–155). John Wiley & Sons, Inc.

Pfeifer, J. H., & Berkman, E. T. (2018). The Development of Self and Identity in Adolescence: Neural Evidence and Implications for a Value-Based Choice Perspective on Motivated Behavior. *Child Development Perspectives*, *12*(3), 158–164. https://doi.org/10.1111/cdep.12279

Pfeifer, J. H., Kahn, L. E., Merchant, J. S., Peake, S. J., Veroude, K., Masten, C. L., Lieberman, M. D., Mazziotta, J. C., & Dapretto, M. (2013). Longitudinal Change in the Neural Bases of Adolescent Social Self-Evaluations: Effects of Age and Pubertal Development. *The Journal of Neuroscience*, *33*(17), 7415–7419. https://doi.org/10.1523/JNEUROSCI.4074-12.2013

Pingault, J. B., Côté, S. M., Petitclerc, A., Vitaro, F., & Tremblay, R. E. (2015). Assessing the Independent Contribution of Maternal Educational Expectations to Children's Educational Attainment in Early Adulthood: A Propensity Score Matching Analysis. *PLoS ONE*, *10*(3), e0119638. https://doi.org/10.1371/journal.pone.0119638

Pinquart, M., & Ebeling, M. (2020). Parental Educational Expectations and Academic Achievement in Children and Adolescents—a Meta-analysis. *Educational Psychology Review*, *32*(2), 463–480. https://doi.org/10.1007/s10648-019-09506-z

Platt, L., & Parsons, S. (2017). *Is the future female? Educational and occupational aspirations of teenage boys and girls in the UK* (Working paper No. 17/2017). Centre for Longitudinal Studies.

Plewis, I. (2007). *The Millennium Cohort Study: Technical Report on Sampling* (tech. rep.) (4th Edition). Centre for Longitudinal Studies. http://doc.ukdataservice.ac.uk/doc/4683/mrdoc/pdf/mcs_technical_report_on_sampling_4th_edition.pdf

Poropat, A. E. (2009). A Meta-Analysis of the Five-Factor Model of Personality and Academic Performance. *Psychological Bulletin*, *135*(2), 322–338. https://doi.org/10.1037/a0014996

Pulkkinen, J., & Rautopuro, J. (2022). The correspondence between PISA performance and school achievement in Finland. *International Journal of Educational Research*, *114*, 102000. https://doi.org/10.1016/j.ijer.2022.102000

Putwain, D. W., & Wood, P. (2023). Anxiety in the mathematics classroom: Reciprocal relations with control and value, and relations with subsequent achievement. *ZDM – Mathematics Education*, *55*(2), 285–298. https://doi.org/10.1007/s11858-022-01390-2

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 No. 2013-15). Institute for Social and Economic Research (ISER), University of Essex.

Ramsden, S., Richardson, F. M., Josse, G., Thomas, M. S. C., Ellis, C., Shakeshaft, C., Seghier, M. L., & Price, C. J. (2011). Verbal and non-verbal intelligence changes in the teenage brain. *Nature*, *479*(7371), 113–116. https://doi.org/10.1038/nature10514

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). https://doi.org/10.3102/0013189X1876210

Rees, D. I., Sabia, J. J., & Argys, L. M. (2009). A head above the rest: Height and adolescent psychological well-being. *Economics & Human Biology*, *7*(2), 217–228. https://doi.org/10.1016/j.ehb.2009.04.002

Reeve, J. (2012). A Self-determination Theory Perspective on Student Engagement. In S. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of Research on Student Engagement* (pp. 149–172). Springer, Boston, MA. https://doi.org/10.1007/978-1-4614-2018-7_7

Reilly, D., Neumann, D., & 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

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

Reschly, A. L., & Christenson, S. L. (2012). Jingle, Jangle, and Conceptual Haziness: Evolution and Future Directions of the Engagement Construct. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of Research on Student Engagement* (pp. 3–19). Springer US. https://doi.org/10.1007/978-1-4614-2018-7_1

Retelsdorf, J., Asbrock, F., & Schwartz, K. (2015). “Michael Can’t Read!” Teachers’ Gender Stereotypes and Boys’ Reading Self-Concept. *Journal of Educational Psychology*, *107*, 186–194. https://doi.org/10.1037/a0037107

Revelle, W. (2024). Using R and the psych package to find ω. https://personality-project.org/r/psych/HowTo/omega.pdf

Reynolds, M. R., Scheiber, C., Hajovsky, D. B., Schwartz, B., & Kaufman, A. S. (2015). Gender Differences in Academic Achievement: Is Writing an Exception to the Gender Similarities Hypothesis? *The Journal of Genetic Psychology*, *176*(4), 211–234. https://doi.org/10.1080/00221325.2015.1036833

Roberts, N., & Long, R. (2025). Research Briefing. Special Educational Needs: Support in England. https://commonslibrary.parliament.uk/research-briefings/sn07020/

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

Robitzsch, A., & Grund, S. (2022). Miceadds: Some Additional Multiple Imputation Functions, Especially for ’mice’. https://CRAN.R-project.org/package=miceadds

Rogers, M., Bélanger-Lejars, V., Toste, J. R., & Heath, N. L. (2015). Mismatched: ADHD symptomatology and the teacher–student relationship. *Emotional and Behavioural Difficulties*. https://doi.org/10.1080/13632752.2014.972039

Rosseel, Y. (2012). Lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02.

RStudio Team. (2020). RStudio: Integrated Development for R. 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

Rudman, L. A., & Fairchild, K. (2004). Reactions to Counterstereotypic Behavior: The Role of Backlash in Cultural Stereotype Maintenance. *Journal of Personality and Social Psychology*, *87*(2), 157–176. https://doi.org/10.1037/0022-3514.87.2.157

Rudman, L. A., Moss-Racusin, C. A., Glick, P., & Phelan, J. E. (2012, January). Chapter four - Reactions to Vanguards: Advances in Backlash Theory. In P. Devine & A. Plant (Eds.), *Advances in Experimental Social Psychology* (pp. 167–227, Vol. 45). Academic Press. https://doi.org/10.1016/B978-0-12-394286-9.00004-4

Ryan, A. (2001). The peer group as a context for the development of young adolescent motivation and achievement. *Child Development*, *72*(4), 1135. https://doi.org/10.1111/1467-8624.00338

Ryan, R., & Deci, E. L. (2000a). Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being. *American Psychologist*, *55*(1), 68*78. https://doi.org/10.1037110003-066X.55.1.68

Ryan, R., & Deci, E. L. (2000b). Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions. *Contemporary Educational Psychology*, *25*(1), 54–67. https://doi.org/10.1006/ceps.1999.1020

Santos, C., 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

Santos, S., Ferreira, H., Martins, J., Gonçalves, J., & Castelo-Branco, M. (2022). Male sex bias in early and late onset neurodevelopmental disorders: Shared aspects and differences in Autism Spectrum Disorder, Attention Deficit/hyperactivity Disorder, and Schizophrenia. *Neuroscience & Biobehavioral Reviews*, *135*, 104577. https://doi.org/10.1016/j.neubiorev.2022.104577

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

Schaffhuser, K., Allemand, M., & Schwarz, B. (2017). The Development of Self-Representations During the Transition to Early Adolescence: The Role of Gender, Puberty, and School Transition. *The Journal of Early Adolescence*, *37*(6), 774–804. https://doi.org/10.1177/0272431615624841

Schmitz, K. E., Hovell, M. F., Nichols, J. F., Irvin, V. L., Keating, K., Simon, G. M., Gehrman, C., & Jones, K. L. (2004). A Validation Study of Early Adolescents' Pubertal Self-Assessments. *The Journal of Early Adolescence*, *24*(4), 357–384. https://doi.org/10.1177/0272431604268531

Schnitzler, K., Holzberger, D., & Seidel, T. (2021). All better than being disengaged: Student engagement patterns and their relations to academic self-concept and achievement. *European Journal of Psychology of Education*, *36*(3), 627–652. https://doi.org/10.1007/s10212-020-00500-6

Senia, J. M., Donnellan, M. B., & Neppl, T. K. (2018). Early pubertal timing and adult adjustment outcomes: Persistence, attenuation, or accentuation? *Journal of Adolescence*, *65*, 85–94. https://doi.org/10.1016/j.adolescence.2018.03.003

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

Shenouda, C. K. (2014). *Effects of gender stereotypes on children's beliefs, interests, and performance in STEM fields* [Ph.D.]. Michigan State University. https://www.proquest.com/docview/1553436242/abstract/7A9B9DFC5D45462DPQ/1

Shepherd, P., & Gilbert, E. (2019). *Millennium Cohort Study. Ethical Review and Consent* (tech. rep.). Centre for Longitudinal Studies, UCL Institute of Education. London. https://cls.ucl.ac.uk/wp-content/uploads/2017/07/MCS-Ethical-Approval-and-Consent-2019.pdf

Shirtcliff, E. A., Dahl, R. E., & Pollak, S. D. (2009). Pubertal Development: Correspondence Between Hormonal and Physical Development. *Child Development*, *80*(2), 327–337. https://doi.org/10.1111/j.1467-8624.2009.01263.x

Shulman, E. P., Harden, K. P., Chein, J., & Steinberg, L. (2014). The Development of Impulse Control and Sensation-Seeking in Adolescence: Independent or Interdependent Processes? *Journal of Research on Adolescence*, *26*(1), 1–8. https://doi.org/10.1111/jora.12181

Shulman, E. P., Smith, A. R., Silva, K., Icenogle, G., Duell, N., Chein, J., & Steinberg, L. (2016). The dual systems model: Review, reappraisal, and reaffirmation. *Developmental Cognitive Neuroscience*, *17*, 103–117. https://doi.org/10.1016/j.dcn.2015.12.010

Simmons, R. G., & Blyth, D. A. (1987). *Moving into adolescence: The impact of pubertal change and school context.* Aldine.

Simpkins, S. D., Fredricks, J. A., & Eccles, J. S. (2012). Charting the Eccles' expectancy-value model from mothers' beliefs in childhood to youths' activities in adolescence. *Developmental Psychology*, *48*(4), 1019–1032. https://doi.org/10.1037/a0027468

Sisk, C. L., & Foster, D. L. (2004). The neural basis of puberty and adolescence. *Nature Neuroscience*, *7*, 1040. https://doi.org/10.1038/nn1326

Sivula, T., Magnusson, M., & Vehtari, A. (2020). Uncertainty in Bayesian Leave-One-Out Cross-Validation Based Model Comparison. https://doi.org/10.48550/arXiv.2008.10296

Skinner, E. A. (2023). Four Guideposts toward an Integrated Model of Academic Motivation: Motivational Resilience, Academic Identity, Complex Social Ecologies, and Development. *Educational Psychology Review*, *35*(3), 80. https://doi.org/10.1007/s10648-023-09790-w

Skinner, E. A., Furrer, C., Marchand, G., & Kindermann, T. (2008). Engagement and disaffection in the classroom: Part of a larger motivational dynamic? *Journal of Educational Psychology*, *100*(4), 765–781. https://doi.org/10.1037/a0012840

Skinner, E. A., Kindermann, T. A., & Furrer, C. J. (2009). A Motivational Perspective on Engagement and Disaffection: Conceptualization and Assessment of Children's Behavioral and Emotional Participation in Academic Activities in the Classroom. *Educational and Psychological Measurement*, *69*(3), 493–525. https://doi.org/10.1177/0013164408323233

Skinner, E. A., & Raine, K. E. (2022). Unlocking the Positive Synergy Between Engagement and Motivation. In A. L. Reschly & S. L. Christenson (Eds.), *Handbook of Research on Student Engagement* (pp. 25–56). Springer International Publishing. https://doi.org/10.1007/978-3-031-07853-8_2

Skipper, Y., & Fox, C. (2022). Boys will be boys: Young people's perceptions and experiences of gender within education. *Pastoral Care in Education*, *40*(4), 391–409. https://doi.org/10.1080/02643944.2021.1977986

Smith-Woolley, E., Rimfeld, K., & Plomin, R. (2017). Weak associations between pubertal development and psychiatric and behavioral problems. *Translational Psychiatry*, *7*(4). https://doi.org/10.1038/tp.2017.63

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

Soto, C. J., John, O. P., Gosling, S. D., & Potter, J. (2011). Age Differences in Personality Traits From 10 to 65: Big Five Domains and Facets in a Large Cross-Sectional Sample. *Journal of Personality and Social Psychology*, *100*(2), 330–348. https://doi.org/10.1037/a0021717

Spear, L. P. (2000). The adolescent brain and age-related behavioral manifestations. *Neuroscience & Biobehavioral Reviews*, *24*(4), 417–463. http://www.sciencedirect.com/science/article/pii/S0149763400000142

Spear, L. P. (2013). Adolescent neurodevelopment. *The Journal of adolescent health : official publication of the Society for Adolescent Medicine*, *52*(2, Supplement 2), S7–S13. https://doi.org/https://doi.org/10.1016/j.jadohealth.2012.05.006

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

Starr, A., Haider, Z. F., & von Stumm, S. (2024). Do school grades matter for growing up? Testing the predictive validity of school performance for outcomes in emerging adulthood [Place: US Publisher: American Psychological Association]. *Developmental Psychology*, *60*(4), 665–679. https://doi.org/10.1037/dev0001548

Stattin, H., & Magnusson, D. (1990). *Pubertal maturation in female development* [Pages: viii, 402]. Lawrence Erlbaum Associates, Inc.

Steinberg, L., Albert, D., Cauffman, E., Banich, M., Graham, S., & Woolard, J. (2008). Age differences in sensation seeking and impulsivity as indexed by behavior and self-report: Evidence for a dual systems model. *Developmental Psychology*, *44*(6), 1764–1778. https://doi.org/10.1037/a0012955

Steinberg, L., Icenogle, G., Shulman, E. P., Breiner, K., Chein, J., Bacchini, D., Chang, L., Chaudhary, N., Giunta, L. D., Dodge, K. A., Fanti, K. A., Lansford, J. E., Malone, P. S., Oburu, P., Pastorelli, C., Skinner, A. T., Sorbring, E., Tapanya, S., Tirado, L. M. U., … Takash, H. M. S. (2018). Around the world, adolescence is a time of heightened sensation seeking and immature self-regulation. *Developmental Science*, *21*(2), 1–13. https://doi.org/10.1111/desc.12532

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). https://doi.org/10.1371/journal.pone.0057988

Stoet, G., & Geary, D. C. (2015). Sex differences in academic achievement are not related to political, economic, or social equality. *Intelligence*, *48*, 137–151. https://doi.org/10.1016/j.intell.2014.11.006

Stoet, G., & Geary, D. C. (2018). The Gender-Equality Paradox in Science, Technology, Engineering, and Mathematics Education. *Psychological Science*, *29*(4), 581–593. https://doi.org/10.1177/0956797617741719

Stoet, G., & Geary, D. C. (2020a). Gender differences in the pathways to higher education. *Proceedings of the National Academy of Sciences*, *117*(25), 14073–14076. https://doi.org/10.1073/pnas.2002861117

Stoet, G., & Geary, D. C. (2020b). Sex-specific academic ability and attitude patterns in students across developed countries. *Intelligence*, *81*(May), 101453. https://doi.org/10.1016/j.intell.2020.101453

Stoet, G., & Yang, J. (2016). The boy problem in education and a 10-point proposal to do something about it. *New Male Studies*, *5*(2), 1839.

Strand, S. (2010). Do some schools narrow the gap? Differential school effectiveness by ethnicity, gender, poverty, and prior achievement. *School Effectiveness and School Improvement*, *21*(3), 289–314. https://doi.org/10.1080/09243451003732651

Strand, S. (2011). The limits of social class in explaining ethnic gaps in educational attainment. *British Educational Research Journal*, *37*, 197–229. https://doi.org/10.1080/01411920903540664

Strand, S. (2014). Ethnicity, gender, social class and achievement gaps at age 16: Intersectionality and 'getting it' for the white working class. *Research Papers in Education*, *29*(2), 131–171. https://doi.org/10.1080/02671522.2013.767370

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

Susperreguy, M. I., Davis-Kean, P. E., Duckworth, K., & Chen, M. (2018). Self-Concept Predicts Academic Achievement Across Levels of the Achievement Distribution: Domain Specificity for Math and Reading. *Child Development*, *89*(6), 2196–2214. https://doi.org/10.1111/cdev.12924

Sutton, A., Langenkamp, A. G., Muller, C., & Schiller, K. S. (2018). Who Gets Ahead and Who Falls Behind During the Transition to High School? Academic Performance at the Intersection of Race/Ethnicity and Gender. *Social Problems*, *65*(2), 154–173. https://doi.org/10.1093/socpro/spx044

Suutela, M., Miettinen, P. J., Kosola, S., Rahkonen, O., Varimo, T., Tarkkanen, A., Hero, M., & Raivio, T. (2022). Timing of puberty and school performance: A population-based study. *Frontiers in Endocrinology*, *13*, 936005. https://doi.org/10.3389/fendo.2022.936005

Svetaz, M., Ireland, M., & Blum, R. (2000). Adolescents With Learning Disabilities: Risk and Protective Factors Associated With Emotional Well-Being: Findings From the National Longitudinal Study of Adolescent Health. *The Journal of Adolescent Health*, *27*, 340–8. https://doi.org/10.1016/S1054-139X(00)00170-1

Symonds, J. E., & Galton, M. (2014). Moving to the next school at age 10–14 years: An international review of psychological development at school transition. *Review of Education*, *2*(1), 1–27. https://doi.org/10.1002/rev3.3021

Symonds, J. E., & Hargreaves, L. (2016). Emotional and Motivational Engagement at School Transition: A Qualitative Stage-Environment Fit Study. *The Journal of Early Adolescence*, *36*(1), 54–85. https://doi.org/10.1177/0272431614556348

Szumski, G., Smogorzewska, J., & Karwowski, M. (2017). Academic achievement of students without special educational needs in inclusive classrooms: A meta-analysis. *Educational Research Review*, *21*, 33–54. https://doi.org/10.1016/j.edurev.2017.02.004

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, *2*, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

Terrier, C. (2020). Boys lag behind: How teachers' gender biases affect student achievement. *Economics of Education Review*, *77*, 101981. https://doi.org/10.1016/j.econedurev.2020.101981

Testing and Standards Agency. (2018). Teacher assessment frameworks at the end of key stage 1. For use from the 2018/19 academic year onwards. https://dera.ioe.ac.uk/id/eprint/31093/1/Teacher_assessment_frameworks_at_the_end_of_key_stage_1_for_use_from_the_2018_to_2019_academic_year_onwards.pdf

Testing and Standards Agency. (2025). 2025 key stage 2 national curriculum tests: Information for parents. https://www.gov.uk/government/publications/key-stage-2-national-curriculum-tests-and-results-information-for-parents/2025-key-stage-2-national-curriculum-tests-information-for-parents

Tinklin, T., Croxford, L., Ducklin, A., & Frame, B. (2001). *Gender and pupil performance in Scotland's schools* (tech. rep.). The Scottish Executive Education Department / Edinburgh University. http://www.ces.ed.ac.uk/PDF%20Files/Gender_Report.pdf

Tinklin, T. (2003). Gender Differences and High Attainment. *British Educational Research Journal*, *29*(3), 307–325. https://www.jstor.org/stable/1502255

Tomasetto, C., Alparone, F. R., & Cadinu, M. (2011). Girls' math performance under stereotype threat: The moderating role of mothers' gender stereotypes. *Developmental Psychology*, *47*, 943–949. https://doi.org/10.1037/a0024047

Tomaszewski, W., Xiang, N., & Western, M. (2020). Student engagement as a mediator of the effects of socio-economic status on academic performance among secondary school students in Australia. *British Educational Research Journal*, *46*(3). https://doi.org/10.1002/berj.3599

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

UCL Centre for Longitudinal Studies. (2020). *Millennium Cohort Study, User Guide (Surveys 1-5). 9th Edition.* (tech. rep.). Institute for Education, University of London.

UCL Centre for Longitudinal Studies. (2021a). Millennium Cohort Study: Age 11, Sweep 5, 2012. [data collection] [University of London, Institute of Education], (5th Edition). https://doi.org/10.5255/UKDA-SN-7464-5

UCL Centre for Longitudinal Studies. (2021b). Millennium Cohort Study: Age 14, Sweep 6, 2015. [data collection] [University of London, Institute of Education], (7th Edition). https://doi.org/10.5255/UKDA-SN-8156-7

UCL Centre for Longitudinal Studies. (2021c). Millennium Cohort Study: Age 17, Sweep 7, 2018. [data collection]. [University of London, Institute of Education], (2nd Edition). https://doi.org/10.5255/UKDA-SN-8682-2

UCL Centre for Longitudinal Studies. (2021d). Millennium Cohort Study: Age 7, Sweep 4, 2008. [data collection] [University of London, Institute of Education], (8th Edition). https://doi.org/10.5255/UKDA-SN-6411-8

Ullsperger, J. M., & Nikolas, M. A. (2017). A meta-analytic review of the association between pubertal timing and psychopathology in adolescence: Are there sex differences in risk? *Psychological Bulletin*, *143*(9), 903–938. https://doi.org/10.1037/bul0000106

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*, *9*(2), 219–224. https://doi.org/10.1177/1745691614522067

Van Houtte, M. (2023). Understanding the gender gap in school (dis)engagement from three gender dimensions: The individual, the interactional and the institutional [Publisher: Routledge _eprint: https://doi.org/10.1080/03055698.2020.1842722]. *Educational Studies*, *49*(2), 260–278. https://doi.org/10.1080/03055698.2020.1842722

van Buuren, S., & Groothhuis-Oudshoorn, K. (2011). Mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*(3), 1–67.

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

van Hoogdalem, A., & Bosman, A. M. (2024). Intelligence tests and the individual: Unsolvable problems with validity and reliability. *Methodological Innovations*, *17*(1), 6–18. https://doi.org/10.1177/20597991231213871

van Rijn, R., Lee, N. C., Hollarek, M., Sijtsma, H., Walsh, R. J., van Buuren, M., Braams, B. R., & Krabbendam, L. (2023). The Effect of Relative Pubertal Maturation and Perceived Popularity on Symptoms of Depression and Social Anxiety in Adolescent Boys and Girls. *Journal of Youth and Adolescence*, *52*(11), 2384–2403. https://doi.org/10.1007/s10964-023-01836-0

van Tetering, M. A. J., Jolles, J., van der Elst, W., & Jolles, D. D. (2022). School Achievement in Early Adolescence Is Associated With Students' Self-Perceived Executive Functions. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.734576

van Tetering, M. A. J., Laan, A. M., de Kogel, K., Groot, R., & Jolles, J. (2020). Sex differences in self-regulation in early, middle and late adolescence: A large-scale cross-sectional study. *PLOS ONE*, *15*, e0227607. https://doi.org/10.1371/journal.pone.0227607

Veater, M. (2022). *An exploration of the affective engagement of children with special educational needs attending mainstream, rural primary schools* [Doctoral]. UCL (University College London). https://discovery.ucl.ac.uk/id/eprint/10154659/

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved R-hat for assessing convergence of MCMC. *Bayesian Analysis*, *16*(2). https://doi.org/10.1214/20-BA1221

Veiga, F. H., García, F., Reeve, J., Wentzel, K., & Garcia, O. (2015). When Adolescents with High Self-Concept Lose their Engagement in School // Cuando se pierde la motivación escolar de los adolescentes con mejor autoconcepto. *Journal of Psychodidactics*, *20*(2), 305–320. https://doi.org/10.1387/RevPsicodidact.12671

von Hippel, P. T. (2020). How Many Imputations Do You Need? A Two-stage Calculation Using a Quadratic Rule. *Sociological Methods and Research*, *49*(3). https://journals.sagepub.com/doi/abs/10.1177/0049124117747303

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

Voyer, D., Voyer, S. D., & Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. *Psychological Bulletin*, *117*(2), 250–270. https://doi.org/10.1037/0033-2909.117.2.250

Voyer, D., Voyer, S. D., & Saint-Aubin, J. (2017). Sex differences in visual-spatial working memory: A meta-analysis. *Psychonomic Bulletin & Review*, *24*(2), 307–334. https://doi.org/10.3758/s13423-016-1085-7

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

Wallace, T., Anderson, A. R., Bartholomay, T., & Hupp, S. C. (2002). An ecobehavioral examination of high school classrooms that include students with disabilities. *Exceptional children*, *68*(3), 345–359. https://doi.org/10.1177/001440290206800304

Wang, M.-T., Chow, A., Hofkens, T., & Salmela-Aro, K. (2015). The trajectories of student emotional engagement and school burnout with academic and psychological development: Findings from Finnish adolescents. *Learning and Instruction*, *36*, 57–65. https://doi.org/10.1016/j.learninstruc.2014.11.004

Wang, M.-T., & Degol, J. (2014). Staying Engaged: Knowledge and Research Needs in Student Engagement. *Child Development Perspectives*, *8*(3), 137–143. https://doi.org/10.1111/cdep.12073

Wang, M.-T., & Degol, J. L. (2017). Gender Gap in Science, Technology, Engineering, and Mathematics (STEM): Current Knowledge, Implications for Practice, Policy, and Future Directions. *Educational Psychology Review*, *29*(1), 119–140. https://doi.org/10.1007/s10648-015-9355-x

Wang, M.-T., Degol, J. L., & Henry, D. A. (2019). An integrative development-in-sociocultural-context model for children's engagement in learning. *American Psychologist*, *74*(9), 1086–1102. https://doi.org/10.1037/amp0000522

Wang, M.-T., & Eccles, J. S. (2012). Adolescent Behavioral, Emotional, and Cognitive Engagement Trajectories in School and Their Differential Relations to Educational Success. *Journal of Research on Adolescence*, *22*(1), 31–39. https://doi.org/10.1111/j.1532-7795.2011.00753.x

Wang, M.-T., & Eccles, J. S. (2013). School context, achievement motivation, and academic engagement: A longitudinal study of school engagement using a multidimensional perspective. *Learning and Instruction*, *28*, 12–23. https://doi.org/10.1016/j.learninstruc.2013.04.002

Wang, M.-T., Eccles, J. S., & Kenny, S. (2013). Not Lack of Ability but More Choice. *Psychological science*, *24*. https://doi.org/10.1177/0956797612458937

Wang, M.-T., Fredricks, J. A., Ye, F., Hofkens, T., & Linn, J. S. (2017). Conceptualization and assessment of adolescents' engagement and disengagement in school: A Multidimensional School Engagement Scale. *European Journal of Psychological Assessment*, *35*(4), 592–606. https://doi.org/10.1027/1015-5759/a000431

Wang, M.-T., & Peck, S. C. (2013). Adolescent educational success and mental health vary across school engagement profiles. *Developmental Psychology*, *49*(7), 1266–1276. https://doi.org/10.1037/a0030028

Wang, M.-T., Willett, J. B., & Eccles, J. S. (2011). The assessment of school engagement: Examining dimensionality and measurement invariance by gender and race/ethnicity. *Journal of School Psychology*, *49*(4), 465–480. https://doi.org/10.1016/j.jsp.2011.04.001

Wang, S., Rubie-Davies, C. M., & Meissel, K. (2018). A systematic review of the teacher expectation literature over the past 30 years. *Educational Research and Evaluation*, *24*(3-5), 124–179. https://doi.org/10.1080/13803611.2018.1548798

Wanous, J. P., Reichers, A. E., & Hudy, M. J. (1997). Overall job satisfaction: How good are single-item measures? *Journal of Applied Psychology*, *82*(2), 247.

Watson, P. W. S. J., Rubie-Davies, C. M., & Meissel, K. (2019). Mathematics Self-Concept in New Zealand Elementary School Students: Evaluating Age-Related Decline. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.02307

Welmond, M. J., & Gregory, L. (2021). Educational Underachievement Among Boys and Men. https://documents1.worldbank.org/curated/en/111041644611110155/pdf/Educational-Underachievement-Among-Boys-and-Men.pdf

West, P., Sweeting, H., & Young, R. (2010). Transition matters: Pupils' experiences of the primary–secondary school transition in the West of Scotland and consequences for well-being and attainment. *Research Papers in Education*, *25*(1), 21–50. https://doi.org/10.1080/02671520802308677

Widlund, A., Tuominen, H., & Korhonen, J. (2021). Development of school engagement and burnout across lower and upper secondary education: Trajectory profiles and educational outcomes. *Contemporary Educational Psychology*, *66*, 101997. https://doi.org/10.1016/j.cedpsych.2021.101997

Wigfield, A. (1994). Expectancy-value theory of achievement motivation: A developmental perspective. *Educational Psychology Review*, *6*(1), 49–78. https://doi.org/10.1007/BF02209024

Wigfield, A., & Eccles, J. S. (2000). Expectancy–Value Theory of Achievement Motivation. *Contemporary Educational Psychology*, *25*(1), 68–81. https://doi.org/10.1006/ceps.1999.1015

Wigfield, A., Tonks, S., & Klauda, S. L. (2009). Expectancy-value theory. In *Handbook of motivation at school* (pp. 55–75). Routledge/Taylor & Francis Group.

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. (2013). *Gender and fair assessment*. Routledge.

Wilson, M., Donnelly, S., & Stainer, J. (2018). Barriers to participation and progression in education. https://www.endingtheharm.com/wp-content/uploads/2022/04/Barriers-participation-progression-report.pdf

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

Wong, W. I., Shi, S. Y., & Yeung, S. P. (2023). Girls Are Better Students but Boys Will Be More Successful at Work: Discordance Between Academic and Career Gender Stereotypes in Middle Childhood. *Archives of Sexual Behavior*, *52*(3), 1105–1121. https://doi.org/10.1007/s10508-022-02523-0

Wong, Z. Y., Liem, G. A. D., Chan, M., & Datu, J. A. D. (2024). Student engagement and its association with academic achievement and subjective well-being: A systematic review and meta-analysis. *Journal of Educational Psychology*, *116*(1), 48–75. https://doi.org/10.1037/edu0000833

Woods, S. A., & Patterson, F. (2024). A critical review of the use of cognitive ability testing for selection into graduate and higher professional occupations. *Journal of Occupational and Organizational Psychology*, *97*(1), 253–272. https://doi.org/10.1111/joop.12470

Woolcock, N. (2024). Boys set to close gap with girls for top A-level grades. *The Times*. Retrieved August 19, 2024, from https://www.thetimes.com/uk/education/article/a-level-results-2024-boys-expected-higher-girls-kc7sqkpqt

Wu, M. (2010). *Comparing the Similarities and Differences of PISA 2003 and TIMSS* (tech. rep. No. 32). Assessment Research Centre, University of Melbourne. Paris, France. https://doi.org/10.1787/5km4psnm13nx-en

Wyse, D., Selwyn, N., Smith, E., Suter, L., Sullivan, A., & Calderwood, L. (2017). Surveys: Longitudinal, Cross-sectional and Trend Studies. In *The BERA/SAGE Handbook of Educational Research: Two Volume Set* (pp. 395–415). SAGE Publications Ltd. https://doi.org/10.4135/9781473983953

Yu, J. (2019). *Motivational processes underlying gender gaps in school engagement and achievement* [Doctoral dissertation, Homerton College, University of Cambridge].

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

Zhang, Y., Haddad, E., Torres, B., & Chen, C. (2011). The Reciprocal Relationships Among Parents' Expectations, Adolescents' Expectations, and Adolescents' Achievement: A

Two-Wave Longitudinal Analysis of the NELS Data. *Journal of Youth and Adolescence*, *40*(4), 479–489. https://doi.org/10.1007/s10964-010-9568-8

Zuckerman, M. (1994). *Behavioral expressions and biosocial bases of sensation seeking*. Cambridge University Press.