Cambridge Assessment
Admissions Testing

Literature review on the gender effect in educational testing

Nathaniel Owen and Amy Devine

Literature review on the gender effect in educational testing

Nathaniel Owen and Amy Devine

Introduction

Gender differences exist in performance on admissions tests used for entry to higher education in Science, Technology, Engineering and Mathematics (STEM)[1] fields. For example, in the BioMedical Admissions Test (BMAT), which is developed by Cambridge Assessment Admissions Testing, males tend to score higher than females in Sections 1 (Thinking Skills[2]) and 2 (Scientific Knowledge and Applications), and females score higher than males in Section 3 (Writing Task) (Emery, 2013). In order to understand whether gender differences in admissions test performance may reflect true differences in the domains of interest and/or general societal biases in those pursuing scientific study, researchers at Cambridge Assessment Admissions Testing conducted a literature review.

Research Questions

The following questions were of particular interest:

1. To what extent do differences in test scores represent some truth in the domains of interest in terms of:
    a. Sampling?
    b. Ability?
    c. Test score variance?
    d. Developmental differences?

2. To what extent do differences in STEM admissions test scores reflect a general societal bias against females in terms of:
    a. Participation in STEM subjects?
    b. Method effect?

This review was conducted between 2015 and 2018 and covers work undertaken in relation to STEM subjects up to 2017. The studies reviewed cover a variety of age groups. In terms of sample selection, most studies reviewed include large-scale studies with some degree of range restriction that sample from the top end of their respective distributions, as university admissions tests focus on the selection of individuals who have already attained high grades in their subject specialisms.

The review will address the hypothesis that the score distributions of males and females are different, with males exhibiting greater score variance. The review will identify what differences occur in test results between males and females and consider to what extent these differences are replicated in different contexts and different age groups. It will evaluate the strength of the evidence

[1] STEM subjects here are defined according to the definition outlined by the House of Lords and JACS 3 subject categorisations provided by Higher Education Statistics Agency Limited (2014): https://publications.parliament.uk/pa/ld201213/ldselect/ldsctech/37/3705.htm#note9
This review largely focuses on physical and mathematical sciences, but also includes geography and computer sciences.

[2] Section 1 was formerly known as 'Aptitude and Skills'.

associated with claims they make. It will also investigate hypotheses regarding behavioural differences between males and females in terms of the academic pathways they select, what influences these choices, and how males and females react to competitive testing situations.

Table 1 Abbreviations used in this review

AMC American Mathematics Competitions
AAUW American Association of University Women
BMAT BioMedical Admissions Test
DPP Deprivation Pupil Premium
EAL English as an Additional Language
ECLS Early Childhood Longitudinal Study
ESSA Every Student Succeeds Act
FSM Free school meals
GGI Gender Gap Index
GPA Grade point average
HSB High School and Beyond
IEA Evaluation of International Achievement
MRI Magnetic resonance imaging
NAEP National Assessment of Educational Progress
NCLB No Child Left Behind
NLSY National Longitudinal Survey of Youth
OECD Organisation for Economic Co-operation and Development
PISA Programme for International Student Assessment
SAT Scholarly Aptitude Test (formerly)
SD Standard Deviation
SEN Special educational needs
SMS Scottish Mental Survey
ST Stereotype threat
STEM Science, Technology, Engineering and Mathematics
TFA true-false-abstain
TIMSS Trends in International Mathematics and Science Study
VR Variance ratio

Context

The study of differences in educational attainment between genders is controversial. It is a high-profile topic given that governments and media organisations focus on test scores as a barometer of the economic health of a nation. The *Wall Street Journal* proclaimed Programme for International Student Assessment (PISA) 2000 data, in which the United States ranked 20th out of 41 countries, an 'economic time bomb' (Kronholz, 2004). Similarly, when test scores reveal differences in performance between genders, many popular books are written on the subject with alarmist titles, e.g. *Failing at Fairness: How America's Schools Cheat Girls* (Sadker and Sadker, 1994), or *The War Against Boys* (Sommers, 2000). Public discussion becomes heated, with rival interpretations of whether these gender differences are based on genuine psychological or neurological differences in men and women, or the result of social conditioning from an early age which predisposes men to make stereotypically 'masculine' choices and women to make 'feminine' choices. Attention with regard to social inequality focuses almost exclusively on the most privileged positions in society, such as highlighting that Nobel Prize winners are almost exclusively male. Larry Summers (2005), at that time the President of Harvard University, drew attention to research that claimed this is due to over-representation of males at the top end of the population distribution in mental aptitude. However, various commentators have argued that such findings should not be used to make normative claims about gender differences, and some have implied that ideology is overriding impartial analysis of scientific data (e.g. 'The blank slate: The modern denial of human nature', Pinker, 2002).

Feminist thinking is inclined to consider visible differences in test scores, and the subsequent earning power of men and women, to be the result of historically-entrenched systemic social forces which both overtly and covertly favour men in the workplace and society more generally. They point to countries which demonstrate less inequality between males and females as the result of those countries' forward-thinking policies, which have been actively introducing legislation aimed at reducing gender inequality, thus proving that the inequality was the result of structural forces and institutional sexism rather than being reflective of innate ability. The Nordic countries are usually cited, for introducing policies such as mandatory paternal leave, generous parental leave benefits for both men and women, social insurance and tax incentives for employers have lowered the opportunity cost[3] of having children, and board quotas that have ensured 40 per cent representation for women on executive boards (Zahidi, 2013).

Stereotyped claims about innate differences between boys and girls such as 'brain researchers have proven that boys learn differently than girls', an opinion made by a teacher reported in Halpern et al. (2011), are typical of claims made in popular science books which address the topic of male/female differences. These books often make claims or form narratives on the basis of isolated studies rather than meta-analysis relating to large samples which are more representative of populations (Halpern et al., 2011). Nonetheless, the view persists that gender differences in abilities are innate, i.e., due to genetic inheritance and other biological factors, rather than considering the interaction of genetics/biology with environmental factors. This review will not postulate on innate vs. environmental factors in impacting on male and female test scores. Instead, it is concerned with the range of variables impacting on *developmental* differences between males and females and to what extent these are observable in test scores.

It should also be noted that the literature referred to within this review often (incorrectly) conflates the terms 'sex' and 'gender' and as such, has not examined each construct separately or

[3] Opportunity cost refers to 'the potential benefits that an individual […] misses out on when choosing one alternative over another' (Fernando, 2022): https://www.investopedia.com/terms/o/opportunitycost.asp#:~:text=Opportunity%20cost%20is%20the%20forgone,and%20weighed%20against%20the%20others

acknowledged the interaction between the two (see Rioux et al., 2022, for a discussion of this issue). Moreover, the reviewed research literature conceptualises gender/sex as representing two mutually exclusive categories: boys/men/males and girls/women/females. We recognise that such a view is problematic (ibid.) as the research findings may not reflect the diversity of experiences of individuals that fall outside these binaries. However, for accuracy we summarise the research literature referring to the terminology and the grouping of samples used in the reviewed studies. We expect that future research may take a more inclusive view of gender identity when investigating gender differences in STEM.

This review is organised under headings related to the research questions outlined above, starting with studies looking at identifying differences between males and females in test scores, and explanations of differences encountered in terms of sampling and developmental differences. It then examines studies which explore wider issues around participation in STEM subjects and how male and female students react differently to test situations which may impact upon test scores.

Sampling and Reporting

Hyde (1981, p.899) identified sampling issues in comparing males to females in different cognitive abilities. Comparing genders in a population in which males and females have elected the same pathway (e.g. an advanced mathematics course) means that claims cannot be made about a *population* after cut-off points have been established to determine abilities. Thus, studies attempting to make claims about males and females in general require random sampling of a population rather than a truncated, self-selecting sample. Emergent differences should then be analysed to identify causes, be they developmental or cultural.

Studies which take convenience samples from undergraduate courses have already sampled for exceptional verbal and mathematical reasoning (depending on the academic pathways). Studies comparing males and females in high school physics may select students after they have elected to study the subject, biasing towards the top end of the distribution in potential for success in that subject (Willingham & Cole, 2010). Even tests which gather information about large numbers of test takers tend to target specific populations which are often self-selecting. The SAT (formerly the Scholastic Aptitude Test), for example, targets individuals who are applying for higher education in the United States. The test population is therefore pre-selected for talent.

Questions of the representativeness of individual samples are relatively easy to address given that national populations are usually well-defined by census information. Studies which sample talented populations, that is, pre-selected for talent in some respect, will have already skewed their sample and will therefore not shed much light on the origins of the differences between males and females with respect to test scores. Biases with respect to sampling may be relatively small, especially in large-scale studies involving thousands of participants, but if the difference between males and females is also small, then these differences may mask or exaggerate any potential emergent differences (Hedges & Nowell, 1995, p.42).

However, studies analysing extremely large samples also create a problem, in that such studies are more likely to produce statistically significant outcomes which are a function of the massive sample size; they may be very small differences yet are determined to be significant, whilst accounting for very little overall variance. This is known as the ‘crud factor’ (Meehl, 1990) and is important in considering the meaningfulness of statistically significant findings. This emphasises the importance of establishing *effect size* of individual studies. Small-scale, locally produced studies of difference between genders do not often make claims about differences between males and females as a

whole, but make claims about the males and females within their local population, or within the programme on which they are enrolled. Such studies use this information to tailor course materials or pedagogical practices to the requirements of their students, rather than make psychological claims about why these differences occurred. These studies will therefore not form the basis of discussion, but may be included if they shed light on how differences between males and females occur in specific contexts.

Taking all this into consideration, the majority of studies described in this review will be those which have attempted to sample randomly across school-aged populations in one region or country, and therefore have extremely large samples (n > 10,000), and which use instruments that target general verbal, spatial and quantitative ability which may be replicated, rather than locally-produced tests which may have validity or reliability issues and which cannot easily be replicated.

Analysis of differences between males' and females' test scores is usually achieved through national surveys or standardised tests, as these are conducted on large sample sizes in order to generalise the findings. Specifically, it is investigated through analysis of standardised differences in mean score between males and females divided by the standard deviation in the population (the average of male and female standard deviation). National surveys tend to include a range of subject areas in which either males or females are traditionally regarded as strong. These include vocabulary, reading comprehension, verbal ability, arithmetic, physical and biological sciences, and a composite of spatial and verbal reasoning.

In order to establish a general claim of the differences between males and females in terms of test scores in scientific subjects, the review focuses on meta-analytic studies. *Meta-analysis* is a means of combining results of multiple quantitative research studies, commonly citing effect sizes as a standardised method of comparing outcomes. This is preferred to statistical significance, as outcomes are highly dependent on sample size. Meta-analysis compares multiple effect sizes from different large-scale studies to determine consistency and magnitude of a finding across contexts.

Ability (United States and UK)

The review includes several high-profile large-scale meta-analytic studies published in *Science*. These tend to be conducted in the United States, but have received widespread attention in establishing what claims can be made regarding gender-based differences in STEM subject scores (Dee, 2006; Hedges & Nowell, 1995; Hyde & Linn, 2006; Hyde, Lindberg, Linn, Ellis, & Williams, 2008; Maccoby & Jacklin, 1974; Nowell & Hedges, 1998). Maccoby and Jacklin (1974) concluded that gender differences have reliably emerged across studies: males tend to outperform females in visual and spatial ability and females outperform males in verbal ability. Their conclusion is indicative of widespread claims in education, based on an extensive body of literature, which has been increasingly critiqued.

Hyde (1981), for example, questioned these claims and subjected the studies included in Maccoby and Jacklin's analysis to two meta-analytic techniques to assess the magnitude of gender effects (omega-squared and *d*). Hyde specifically argued that statistical significance to identify a difference is insufficient to make a claim of an established gender difference given the large overlap between populations, and argued that the magnitude (recorded in effect size) was more important in determining the significance of identifiable gender differences (Hyde, 1981). In her meta-analysis, she determined that effect sizes of the studies cited by Maccoby and Jacklin varied between .01 and .45 – small to medium effects (using Cohen's (1977) effect size interpretation: above 0.8 is large, above 0.5 is medium and 0.2 is small; see Table 2). Hyde declared that the 'well-established' gender

differences in these areas were in fact minimal, with gender accounting for 1–5 per cent of the variance in test scores.

Table 2 Cohen's (1977) effect size interpretation

Magnitude of effect	**Cohen's *d***
Large	> 0.8
Medium	> 0.5
Small	> 0.2

Hedges and Nowell (1995) collected data from six large datasets (United States national surveys) containing data from 1960 to 1992. The sample sizes of these studies ranged from 11,914 to 73,425, with ages ranging from 14–22. Sampling weights from each survey were used to create estimates of national means and variances for males and females. The effect size of the difference between males and females was estimated by subtracting the mean of female scores from male scores and dividing the difference by the estimate of the whole population's national standard deviation. The authors found that, on average, females performed better on tests of reading comprehension, perceptual speed, and associative memory, whereas males performed better on tests of mathematics and social studies. However, commensurate with the findings of Hyde (1981), the majority of effect sizes are small to medium. Large effect sizes were recorded in stereotypically male domains, such as mechanical reasoning, electronics and auto repair procedures. No evidence emerged of substantive changes to these trends occurring over the three decades for which data was collected. However, each of these datasets were based on different cohorts at different ages at different times, using different instruments. Thus, the study cannot claim to represent a longitudinal study across the 30-year period for which data was collected.

Nowell and Hedges (1998) went some way to mitigating the weaknesses of the 1995 study. Their study used eight samples of test scores, using data gathered for the 1995 study updated with additional data from the National Assessment of Educational Progress (NAEP) dated 1971 to 1994. The study focused on the United States twelfth-grade population to estimate trends in the data from 1960 to 1994 and, between-instrument variance was computed as part of the study to estimate how much score variability could be determined by choice of instrument. NAEP data was analysed separately. Nowell and Hedges replicated the findings from the original study, with males scoring higher on tests of mathematics and science, and females scoring higher on tests of reading, perceptual speed, and writing, although the effect sizes of differences were once again small (in the magnitude of 0.1 to 0.3). This study found that mean differences in test scores for gender have not significantly changed in the time period for which data was gathered.

White (1992) identified a noticeable and growing gap between male and female test scores from as young as nine years old in the NAEP dataset. In 1990, the mathematics performance of nine-year-old girls equalled their male counterparts, but in science, scored three points lower (on a 300-point scale) than boys, a statistically significant difference. The gap in science performance narrowed between 1973 and 1990. Matyas (1985) argued that by the age of nine, girls already suffer from a deficit in access to scientific investigations in the classroom, which may not cause this difference, but certainly exacerbates it. By the age of 13, the mathematics achievement scores remained consistent in the NAEP dataset, but the science gap had increased to seven points. This gap widened to a 10-point difference between males and females among 17-year-olds. The author notes that this gap has grown between 1973 and 1992 rather than shrank. A more recent analysis of NAEP data from 1990 to 2011 confirmed the existence of small gender differences in mathematics and science favouring males (Reilly, Neumann, & Andrews, 2015).

Regarding early years, the Department of Education's Early Childhood Longitudinal Study (ECLS) data suggests that when children enter kindergarten, girls and boys perform similarly on tests of both reading and mathematics (Dee, 2006). The difference in reading appears to be established by the age of 13, whereas male advantages in science and mathematics progress smoothly from the ages of 9–17, although the difference in science is greater than in mathematics. But by the ages of 9–10, boys outperform girls in mathematics and science, while girls outperform boys in reading. By the age of 17, girls outperform boys in reading by 0.31 SD, and boys outperform girls in science by 0.22 SD and in mathematics by 0.1 SD. However, there is evidence that when observed on a larger scale, these differences become more marginal.

Hyde and Linn (2006) argued that there is no meaningful difference between male and female mathematics ability in the wider population (prior to adolescence) and that this was true for all ethnic groups. The authors presented evidence from a meta-study of 46 reports conducting meta-analysis of gender influence in test scores, which together summarise more than 5,000 individual studies with seven million participants. Of the findings across the 46 papers, 30 per cent were trivial ($d < 0.10$), and 48 per cent were small. Additionally, when Hyde and Linn (2006, p.600) examined NAEP data from 2005 (to update the findings of Nowell & Hedges, 1998), they identified a 4-point difference between boys and girls (on a 300-point scale) in fourth grade science scores. This difference was statistically significant due to the large sample size (approximately 100,000), but this returned an effect size of only 0.12 for fourth grade scores and 0.11 for twelfth grade, consistent with the earlier findings, and indicative of consistency throughout middle and high school education.

Wider social changes regarding the participation of women in STEM subjects and careers appear to reflect the closing gap. In 2001, women earned 48 per cent of the bachelor's degrees in mathematics in the United States (Hyde & Linn, 2006, p.599). Hyde et al. (2008) conducted a further meta-analysis by obtaining detailed information from 10 states in the United States for one year across grades 2–11 for mathematics performance. Consistent with the earlier study, this dataset was combined to form a sample size of more than seven million test takers. Effect sizes for gender differences across all ages were less than 0.10.

A number of robust studies have been conducted to identify differences between male and female test scores in the last few decades, which have been effectively summarised in meta-analytic studies. These have received considerable attention in the literature and suggest that mean differences between genders exist in science and mathematics scores, in which males outperform females, although the gap has narrowed (suggesting a movement towards equality). However, the gender difference in reading and writing (in which females outperform males) remains consistent. These trends were observed across instruments and samples. However, effect sizes of these differences tend to be either small or medium. The narrowing of the performance gap between males and females in mathematics may reflect the result of successive policies to encourage greater participation by women in STEM subjects in the United States, beginning with the 1972 Education Amendments which prohibited sex discrimination at all levels of education.

With regard to gender differences in test scores in the UK specifically, Bramley, Vidal Rodeiro and Vitello (2015) conducted an analysis of GCSE entry and performance in 2014. Although boys outnumbered girls in some STEM GCSE specifications such as engineering and computer studies, entries were similar for mathematics and single science specifications. In terms of performance, there was a tendency for girls to outperform boys across all subjects. There were small gender gaps in the number of candidates gaining A* or C grades in mathematics and science specifications. However, the authors note that the interpretation of these differences is complicated by grade boundary decisions. Thus, gender differences in performance were also examined in terms of a scale-free measure, the 'probability of superiority' statistic. This statistic is 'the probability that a

randomly sampled boy would have a higher score than a randomly sampled girl, with the (hypothetical) sampling coming from the actual distribution of scores on the exam' (Bramley et al., 2015, p.8). Using this measure to analyse gender differences in OCR GCSE specifications, the authors reported that girls tended to do better than boys in all STEM subjects, with the exception of two applied mathematics specifications. Generally, the probability of superiority statistic was smaller for STEM subjects than for humanities, languages and expressive subjects. Collectively these findings suggest that males and females have similar performance in STEM subjects at GCSE level, but it should be noted that this analysis inspected one year only.

Ability (Internationally)

In order to provide a more international picture beyond the United States and UK, the review also considered studies that used data from the Programme for International Student Assessment (PISA) and Trends in International Mathematics and Science Study (TIMSS).

PISA surveys are produced by the Organisation for Economic Co-operation and Development (OECD) and take place in three-year cycles[4]. The first survey took place in 2000. They collect information about 15-year-old students in the areas of mathematics, science and reading from across 90 countries and economies. Tests are reproduced and translated across all OECD members and select partner countries. Since its inception, the total number of student participants has grown from 265,000 to more than 500,000. It therefore forms the basis of comparison of different countries' educational performance. One of the aims of the surveys is to discern differences between countries in the relationships between student-level factors (such as gender and socio-economic background) and achievement (OECD, 2012). PISA tests are designed by the OECD to be free of cultural biases, although this intention does not guarantee that they are, as Soh (2014) notes that overall country performance is skewed if a large proportion of the test takers speak the language of instruction as a second language.

TIMSS is an international assessment of mathematics and science learning in Year 5 and Year 9 students which takes place in four-year cycles[5]. TIMSS surveys are conducted by the International Association for the Evaluation of International Achievement (IEA), in collaboration with Statistics Canada and the Educational Testing Service. TIMSS assesses the attained curriculum and focuses more on teacher- and school-level variables than PISA.

González de San Román and de la Rica (2012) argue that female students underperform relative to males in PISA maths scores (2009 cohort), although females tend to outperform males in the PISA reading tests. The findings suggested that social norms and the parental transmission of role attitudes to children were crucial determinants of these gender differences and the authors noted that enhanced gender equality was associated with substantially improved maths performance by girls. Gender equality was measured through the World Economic Forum's Gender Gap Index (GGI). The GGI measures four dimensions of equality: health (e.g. life expectancy), political participation, economic participation, and educational attainment. The OECD (2015a) reported problem solving PISA scores for 2012 by gender: boys generally outperformed girls. Of the 44 countries and regions included in the PISA data, in 24 boys outperformed girls by a statistically significant margin, in four girls outperformed boys by a significant margin and in 16 there was no significant difference between boys and girls. In the top six performing countries/regions, boys outperformed girls by a significant margin.

[4] https://www.oecd.org/pisa/

[5] https://timssandpirls.bc.edu/timss-landing.html

Guiso, Monte, Sapienza and Zingales (2008) demonstrated that there were no systemic differences in mean attainment between males and females. The authors used 2003 PISA data that reported on 276,165 15-year-old students from 40 countries. There was wide variation between male and female test scores between countries, with boys outscoring girls by an overall average of 10.5 points. However, this discrepancy is inversely correlated with measures of gender equality across participating countries, with gender differences in test scores disappearing as countries become more politically and socially equal. Those countries with the highest gender equality scores such as Norway and Sweden did not report a difference between males and females in mathematics scores.

Else-Quest, Hyde and Linn (2010) reported similar findings from their analysis of the 2003 PISA and TIMSS data sets. These authors found that the size and direction of gender differences in maths and science performance varied by nation. In line with previous studies, most effect sizes were low to moderate and gender differences were smallest or non-existent in more gender equal countries.

Caution should be taken with regard to interpreting the GGI. Rankings strictly relate to differences between men and women; therefore, countries which exhibit low inequality but low overall health and wealth indices are ranked highly. Moreover, others have questioned the association between the gender equality of a nation and the size of the gender difference. Stoet and Geary (2013; 2015) analysed data from four PISA assessments (involving 1.5 million children), including the 2003 data set mentioned above in the analyses by Guiso et al. (2008) and Else-Quest et al. (2010). Stoet and Geary (2013) found that several countries (Iceland, Sweden, Norway and Finland) were driving the correlation between GGI measures and the magnitude of the gender difference in the 2003 assessment; once these countries were removed, the correlation between GGI and the gender difference in mathematics diminished. It is important to point out that despite having the lowest mathematics gender gap, in terms of the overall achievement gap between males and females (the average across mathematics, reading and science), Finland, Norway and Iceland were the top three OECD countries. Collectively these findings suggested that no country had successfully reduced gender differences in achievement in all subjects.

Despite their influence, PISA findings have been criticised by a number of academics for the negative effect they have on individual countries' education policies by encouraging short-term thinking to move up PISA rankings (Meyer et al., 2014). It has also been criticised for its lack of longitudinal datasets, making year-on-year comparisons difficult (if not impossible) and for its uneven sampling methodology across regions/countries (Goldstein, 2004). Kreiner & Christensen (2014) argue that the uneven sampling results in model misfit, arguing that the Rasch model is inappropriate for interpreting PISA data. Individual countries also have specific circumstances which hinder interpretation of their data in relation to other countries. For example, socio-economic factors are thought to contribute to gender differences in educational outcomes and/or the gender ratio of the samples that participate in the PISA assessments; these factors may contribute to the large gender differences in PISA scores reported for some countries such as Bulgaria and United Arab Emirates (European Commission, 2013; Gortazar, Herrera-Sosa, Kutner, Moreno, & Gautam, 2012; IEGE, 2013; McFarlane, 2014; OECD, 2015a; 2015b).

It should be noted that the findings of more recent PISA and TIMMS assessments have been published since this review was conducted.

Test Score Variance

The majority of studies that compare genders usually extrapolate claims to whole populations based on large-scale sampling, rather than focusing on self-selected or truncated samples of the right tail (i.e. higher end) of the distribution, which is of interest to this review. However, some studies also include measures of 'talent' in their research agendas, which are presented as *variance ratios* (Hedges & Nowell, 1995; Nowell & Hedges, 1998).

Variance ratios (VR) are calculated by comparing the proportions of males to females in the right-hand (top five and/or 10 per cent) of the score distribution for each dataset. Greater male variance is indicated by VR > 1.0 and greater female variance is indicated by <1.0, assuming normality in the distribution of scores. 'Talent' in the context of studies of cognitive ability refers to the highest percentiles of the right tail of the distribution.

Studies focusing on talent query the imbalance between high-performing men and women in STEM fields which are linked to economic success. The wage gap between men and women is usually attributed to lack of participation in these disciplines. This difference in subject participation is also considered ultimately responsible for highly visible indicators of gender difference, such as the awarding of Nobel prizes or the Field Medal in mathematics, which are heavily skewed towards male recipients.

Johnson, Carothers and Deary (2008) provide an overview of the historical thinking associated with variability of male and female cognitive abilities, proclaiming that the idea emerged in the late nineteenth century and became entrenched in mainstream thinking in the early twentieth century with the advent of Thorndike's (1906) testing of *general intelligence*. The authors used data from two population-representative samples of general intelligence test scores; the Scottish Mental Survey of 1932 (SMS32) and the Scottish Mental Survey of 1947 (SMS47). SMS data indicated greater variance in males than in females and found some evidence that this greater variance existed at the higher end of general intelligence, but they also found that the lower half of the distribution of general intelligence made a greater contribution to this greater variance than did the upper half. A caveat is that this data comes from children aged 11–12, prior to the development of cognitive maturity that may reveal different distribution shapes and mean differences.

Maccoby and Jacklin (1974), Feingold (1988, 1992) and Hyde et al. (2008) argued that males' test scores in mathematical ability and spatial reasoning demonstrate greater variability than those of females. Hedges and Nowell (1995) performed a meta-analysis of six large datasets between 1960 and 1992 and found that male test scores exhibit greater variability than female test scores in 10 of the 12 subject areas measured in the datasets. Female variance scores demonstrated greater variability in word memory and a coding speed test (the ability to find numbers quickly in a table). However, the latter finding only occurred in one dataset. Three other instruments measuring the same construct all demonstrated greater male variance (1995, p.43). Male variance was typically 5–20 per cent greater than female test score variance overall.

Nowell and Hedges (1998) also found that across all subjects, instruments and years, the ratio of male variance divided by female variance was greater than one in nearly every instance (i.e. VR > 1.0). Thus, greater variance in male test scores than female test scores appears to be a general trend.

However, it is notable that while male variance is generally greater than female variance, this trend is erratic across instruments. Instruments with smaller sample sizes tend to record greater discrepancies in variance. Between-study variance, which measures variance between instruments,

indicates that the choice of instrument has an impact on the outcomes and therefore undermines claims regarding the population (Nowell & Hedges, 1998, p.24).

For example, the National Longitudinal Survey of Youth dataset from 1980 (NLSY80) displayed male-female variance ratios of 1.23 and 1.28 for mathematical knowledge and arithmetic reasoning, respectively, from a sample size of 592, whereas larger datasets from similar years (High School and beyond (HSB), 1980, 1982) recorded slightly lower variance ratios of 1.16 and 1.18, respectively, from samples sizes of 12,534 and 11,623 (Nowell & Hedges, 1998, pp.30–31). Smaller variance ratios appear to be linked to larger-scale studies, giving a truer picture of the difference in distribution shapes in the population.

Furthermore, data from Nowell and Hedges (1998, pp.32–33) seems to confirm stereotypes, in that males are over-represented in the upper tails of science and mathematics score distributions by 2:1 and 1.5:1, whereas females are over-represented in the upper tails of reading and writing at a rate of 1.4:1 and 2.5:1, respectively. These trends appear to hold across instruments and across time periods in the data presented. The gender differences in the tails of the distributions are mostly accounted for by mean differences in scores between males and females rather than differences in variance.

However, even with small average differences between male and female populations, these differences will be amplified at the tails of the distribution if there is unequal variance in female and male populations respectively. For the top five per cent, difference in sample variance contributes more in accounting for the remaining difference in representation (although the influence of mean difference remains consistent). This trend was repeated across instruments. Yu and Shauman (2003) note that the fraction of males to females who score in the top five per cent in high school mathematics remained constant at 2:1 between 1983 and 2003. Hyde et al. (2008) also observed a trend of greater male variance in test scores. All VRs, by state and grade, are >1.0 (range 1.11 to 1.21) (Hyde et al., 2008, p.495). Hyde et al. note that in the dataset, above the 99th percentile, there were a ratio of 1.45:1 males for every female for white ethnic groups, although the trend is reversed for Asian Americans, with a ratio of 0.91:1 males for every female. Nowell and Hedges conclude that the lack of change in proportion ratios at the top end of the distribution is the most alarming threat to achieving equality, as the unequal number of men and women within this talent pool inevitably results in more males than females being represented in STEM occupations.

When factoring in gender equality, Guiso et al. (2008) found that countries which score highly on the GGI measure (described in the previous section) exhibit a smaller gender gap in mean math performance as well as in the tail of the distribution. Using PISA 2003 data, they found that women performed well in both mathematics and reading in societies with greater gender equality. Consistent with Guiso et al., are the findings of Pope and Sydnor (2010), who demonstrated that, in regions of the United States where men and women are viewed as more equal, there were smaller gender disparities in both mathematics and reading. Hyde and Mertz (2009) found a positive correlation between the number of female representatives in International Mathematical Olympiad teams and those countries' GGI (using 2007 data).

In contrast with the findings of Guiso et al. (2008), Machin and Pekkarinen (2008) found no relationship between gender equality and the gender gap at the top end of the distribution. Using the same data from the 2003 PISA cohort, they investigated whether higher variance in male test scores is replicated across borders and cultures and also explored the relationship between gender equality and the magnitude of the gender gap at the top end of the distribution. In 35 of the 41 countries for which data was obtained, the variance ratio indicated that boys' scores had greater variance than girls' scores for both mathematics and reading. Boys generally outscored girls in

mathematics ability with the girls outscoring boys in reading ability. These trends impacted on the number of boys and girls in the tails of the distributions – for maths, in 35 countries, there are more boys than girls in the top five per cent. For reading, 36 countries have more girls than boys in the top five per cent, and in 39 countries, more boys than girls in the bottom five per cent of scores. The authors noted that the variance difference was in fact greater in countries with better overall performance.

Benbow and Stanley (1980; 1983) showed that males outperform females by approximately half a standard deviation (i.e. 0.5 SD) in the mathematics SAT test and that this representation was extended to a ratio of 13:1 for those who score more than 700. The authors argued that the causes of this disparity must be biological on the basis that the participants in the studies had all experienced the same education. This reasoning has been questioned by many in the field, and is undermined, for example, by the findings of Halpern (1992) who reported that the 'mathematics gap' was decreasing over time.

Ellison and Swanson (2010) point out that as SAT scores increase beyond 600, the proportion of females as a proportion of total test takers declines rapidly. There is a 2.1:1 male to female ratio of American students who score more than 800 on the SAT in mathematics when applying for university. They use this statistic to justify examining performance more closely at the top end of the distribution. The authors adopted an innovative approach to examining gender gaps in the upper tails of the distribution in mathematics performance by looking at entrants to the American Mathematics Competitions (AMC). One drawback of this approach is that the sample is self-selected as participation is voluntary, although the questions are designed to discriminate between students at the top end. A second drawback is that women are more underrepresented among high scorers in the AMC than among students with comparable performance on the SATs, suggesting that girls are less attracted to participating in the AMC (Ellison & Swanson, 2010, p.110). This is borne out by the imbalance of females to males at the top end of the AMC. Of test takers who score 100 or more (out of 150), there is a male to female ratio of 4.2:1 (ibid., p.116) (five girls versus 90 boys scoring 136.5 or higher on the AMC 12 test in the 2007 cohort). The authors find that the highest-achieving girls in the United States are concentrated in a small set of elite schools, in contrast to the wider participation among boys, suggesting that not all girls with strong mathematics abilities are being well-served by the American education system. They also note that schools with girls who score highly on the AMC tend to have smaller gender gaps overall.

Ellison and Swanson (2010) demonstrated that fewer girls volunteer for consideration in mathematics competitions. The authors suggest that the lack of participation in mathematics competition is symptomatic of a United States educational system which fails to develop girls' mathematical talent (ibid., p.124). If social factors make girls less likely to join math teams or take advanced online courses, then they will be more underrepresented when we examine achievement levels at the top end of the distribution. These authors noted that high-achieving girls come almost exclusively from high-achieving schools and suggested that girls can thrive in environments which encourage advanced learning. This claim and the data on which it is based are also compatible with the explanation that there is less variance in the female population with regard to mathematical and numerical reasoning.

Wai, Cacchio, Putallaz and Makel (2010) examined the ratio of males to females among the top 0.01 per cent of seventh graders (12–13 years old) taking the SAT test of mathematical reasoning between 1981 and 2010 (a sample size of more than 1.6 million). In the early 1980s, the ratio of boys to girls was 13.5:1. By the early 1990s, this ratio declined rapidly to approximately 4:1. For the two most recent decades for which there are complete data (1991 to 2010), the male/female ratio has remained relatively stable, with a ratio of 3.83:1 for the period 2006 to 2010 (ibid., p.415). This trend

therefore appears to have persisted despite significant attention from academics and politicians to address the gender gap. However, girls have consistently outnumbered boys at the highest levels of verbal reasoning and writing ability across the same time period. There are equal numbers of males and females at the bottom of the distribution, but at the top, the male/female ratio declines to 0.7–0.8 males for every female. Females from the United States therefore seem to have a small but clear and consistent advantage over males in conventions of standard written English and verbal reasoning ability.

Lastly, exploring trends in test score variance in the UK, Bramley et al. (2015) also inspected test score variability in their analysis of gender differences in GCSE performance. Again, using the 'probability of superiority' statistic, they found that girls were overrepresented at the top end of the distribution compared to boys in all GCSE subjects, with the exception of four STEM specifications. However, the authors note that girls were overrepresented at the top end of the distribution in alternative specifications offered for mathematics and physics. In contrast, boys were overrepresented in the bottom five per cent of the distribution for the majority of specifications. With regard to A-levels, data from 2014 candidates shows that a slightly higher percentage of boys gained A* grades in maths, further maths and chemistry than girls, however slightly more girls gained A* grades in geography, biology and physics than boys (Smithers, 2014); thus, boys are not consistently overrepresented at the top end of the distribution in STEM subjects at A-level either.

Developmental Differences

Neuroscientists have found few sex differences in children's brains beyond the larger volume of boys' brains with more neurons (even correcting for body size) (Halpern et al., 2011), though women have a higher percentage of grey matter, and their brain growth is completed earlier.

Lenroot et al. (2007) examine the trajectories of brain development in males and females by performing multiple magnetic resonance imaging (MRI) scans on a cohort of 387 between the ages of 3 and 27. Individuals with learning, developmental, or medical conditions linked to brain development were not accepted into the study. The authors noted that average male brain size was 10 per cent larger than females' and that total cerebral volume peaked at age 10.5 in females and 14.5 in males. Both cortical and subcortical grey matter trajectories follow an inverted U-shaped path with peak sizes one to two years earlier in females than males. Research attempting to link neural gender differences to cognitive gender differences has not yielded many conclusive findings (Hines, 2011); thus, it remains unclear how these findings relate to learning, either generally or within specific subject areas (Pinker, 2002, p.346).

Impact of School on Development

Regardless of the ferocity of the nature/nurture debate, considerable uncertainty exists regarding how males and females are impacted by the experience of the school environment and how this shapes their learning. At the cognitive level, Spelke (2005) argues that some gender differences stem from differing strategy choices, citing evidence that males tend to use spatial strategies more than females, which confers advantages to males in tasks which lend themselves to spatial problem solving (e.g., Johnson, 2004). Importantly, research has shown that gender gaps are reduced by altering the presentation of mathematics problems or by encouraging the use of spatial strategies in all students (Gallagher, Levin, & Cahalan, 2002).

Moreover, students with similar cognitive abilities may receive different test scores if incentives associated with high performance differ or are perceived to differ (Gneezy & Rustichini, 2000). Niederle and Vesterlund (2010) argue that the observed difference in males and females in mathematics at the top end of the distribution is the result of the differential manner in which men and women respond to a competitive test-taking environment. Eccles (1998) argues that girls and boys with approximately similar mathematics test scores hold differing opinions of their relative ability. This gender confidence gap is wider among gifted children than the less gifted (Preckel, Goetz, Pekrun, & Kleine, 2008).

Children are also sensitive to the gender-stereotyped beliefs and expectancies of their socialisers (Eccles, Adler, & Kaczala, 1982a; Eccles, Kaczala, & Meece, 1982b). For example, research in the United States has shown that parents hold differential attributions for girls' and boys' success and failure in mathematics, such as attributing boys' success to talent, and girls' success to hard work. Furthermore, children's self-concept of their mathematics ability and their mathematical confidence are influenced more by their parents' expectations than by children's own past performance (Eccles et al., 1982a; Frome & Eccles, 1998).

Niederle and Vesterlund (2010) studied the effect of an incentive (reward) on the performance of males and females in given tasks. They conclude that females react differently when placed in a competitive environment, and that these different reactions result in observable performance differences between the genders. The biggest difference occurred in the *mixed-sex* experimental group receiving a payment. The authors concluded that women are not averse to competition, but are more averse to competing in situations with men rather than with other women, and that this affected their performance. This trend was visible in the proportion of top performers – the proportion of women in the top fifth performers dropped from 40 to 24 per cent. The authors claimed that the results demonstrated women shy away from competition while men embrace it (ibid., p.135).

Thus, girls appear to be sensitive to competition and the gender of their competitors affects their performance. In mathematics or other STEM subjects which are favoured by large numbers of male applicants, competitive anxiety is heightened for girls. If the competitive element of the task is part of the construct, such as performance in a competitive maths environment (e.g., taking part in the International Mathematics Olympiad), then high-pressure testing environments serve as good predictors of subsequent performance in this domain. However, if the competitive pressure is not part of the construct, then test producers need to consider how this biasing element may impact girls' scores and lead to skewed selection procedures.

Huguet and Régner (2007) demonstrated that girls underperformed in mixed-sex groups when completing a complex figure recall memory task which was described as measuring geometry ability, compared to girls who completed the same task when it was described as measuring drawing ability. However, girls did not underperform in the mathematics condition when completing the task in all-female groups. No significant effect of task condition emerged in boys, and boys tended to perform better in same-sex environments than in mixed-sex environments. The abovementioned studies therefore suggest that boys are generally more confident than girls in terms of task completion, although the relationship between (over/under) confidence and subsequent performance in tests remains unclear.

Dee (2006; 2007) and Carrell, Page and West (2009) also noted that girls' mathematics and science test performance was affected by their teacher's gender in those subjects. The authors found that having a female mathematics or science teacher improved performances in these subjects by girls, and the effect was magnified for particularly gifted female students. This difference was exacerbated

by the relative numbers of male and female mathematics teachers. Dee (2007) reported that the proportion of female mathematics and science teachers decreased as students moved through school. A teacher's gender did have a statistically significant effect on student test performance, teacher perceptions of students, and students' engagement with academic material ($p < .05$), although Dee does not report effect sizes for these findings which are important given the large sample sizes being used. Given a sample size of more than 20,000, the reported difference of an increase in performance of 0.042 standard deviations for girls and 0.046 standard deviations for boys suggests this would result in a small effect size.

Collectively, these findings may have implications for single-sex schooling, which could be advocated as a solution to intervening variables such as female confidence and interpersonal communication. A provision for single-sex schooling in the United States was made via the *No Child Left Behind* (NCLB) act (Dee, 2006) which remains effective as of 2016 under title IX provisions of the successor legislation *Every Student Succeeds Act* (ESSA). Title IX of the United States Education Amendments of 1972 outlawed discrimination on the basis of sex in educational programs receiving federal funds (N.B. not impacting on the provision of fee-paying educational institutions), prohibiting single-sex classes in coeducational schools unless a specific educational objective or government target mandated such a change. Despite this, Dee (2006, p.69) reports that as of April 2006, 223 public schools in the United States offer gender-separate learning environments, from only four in 1998.

Outcomes associated with single-sex education may initially appear positive with more successful educational outcomes than coeducational schools. However, 'apparent advantages dissolve when outcomes are corrected for pre-existing differences' (Halpern et al, 2011, p.1,706). Single-sex education reduces boys' and girls' opportunities to work together and thus perpetuates negative stereotypes which are inculcated from society rather than their own personal experience and which may falsify these general impressions. Sullivan, Joshi and Leonard (2010) examined the impact of single-sex education on males' and females' educational and professional outcomes at various life stages. Single-sex schooling was found to have a positive effect on educational achievement for girls (but not boys). The most noticeable outcomes of single-sex education in this study were increased participation, and better performance, in gender-atypical disciplines in education and subsequent career. This appears to contradict the claim of Halpern et al that more consistent engagement with the opposite sex can play a positive role in reducing stereotype threat.

However, there are problems with claiming that educational outcomes are improved for females in single-sex schools. Single-sex schools tend to be private rather than public, with competitive admissions procedures. Therefore, the data from these schools is from pre-selected samples. Any claims about the efficacy of single-sex education on the basis of this data should therefore be interpreted cautiously. Additionally, United States data suggests that children in single-sex schools who are underperforming relative to their peers may transfer out of the single-sex school prematurely, which may inflate final school performance outcomes (Sweet, 2010, cited by Halpern et al., 2011). Ultimately, without blind assessment, randomized assignment to treatment or control groups with consideration of additional variables, judging the effectiveness of single-sex education is extremely difficult.

Hannay (2016) demonstrated that socio-economic variables likely impact on student attainment more than the gender make-up of the educational institution. Regarding overall GCSE attainment, the author showed that the gap between single-sex schools and mixed schools closes from 20 percentage points to less than three percentage points, when controlling for admissions testing, prior academic attainment at KS2, proportion of pupils with English as an additional language (EAL), the proportion of disadvantaged pupils measured by eligibility for free school meals (FSM) and the

Deprivation Pupil Premium (DPP)[6], and special educational needs (SEN). He concludes that single-sex secondary schooling for girls offers some benefits as measured by GCSE performance, but that any gain by boys is marginal.

Stereotype Threat

An additional explanation of observed differences may be 'stereotype threat' (ST). ST refers to lower test performance by individuals who feel at risk of conforming to negative stereotypes of their group abilities (Huguet & Régner, 2009, p.1,024; Steele & Aronson, 1995). Devine (1989) highlights that stereotypes are entrenched in children before they develop the necessary cognitive abilities for mathematical reasoning or the critical skills necessary to question the impressions conveyed to them by adults.

There are a few studies which support the theory that stereotyping may be a factor in observed differences in performance between girls and boys. For example, Stevenson and Newman (1986) found that mothers' perceptions of daughters' cognitive abilities made in the fifth grade were good predictors of the daughters' attitude towards reading in the tenth grade. Jacobs (1991) found that parents' perceptions of their children's mathematics abilities mediated their children's perception of their own abilities. Frome and Eccles (1998) expanded upon these earlier studies by using structural equation modelling with longitudinal test data and attitude questionnaires to investigate a potential causal connection between parental beliefs and girls' and boys' self-evaluation. Using data from 914 sixth-grade students (age 11–12), the authors found that students' impressions were mediated by the gendered nature of parents' perceptions, independent of what grades they were actually awarded, and that this trend became noticeable between the fifth and seventh grades.

Stereotype activation is therefore in place from an early age and is likely to be an integral component of ST. Using the same experimental manipulation as described above, in which school students completed a complex figure recall memory task when it was described as either a geometry or drawing task, Huguet and Régner (2009) reported that girls demonstrated the effects of ST even if they explicitly denied the negative gender stereotype when completing a self-evaluation questionnaire. Therefore, participants need to be primed in the stereotype to produce the deficit in performance, but do not need to endorse the stereotype. Aronson et al.'s (1999) study suggests that high achievers are more susceptible to this effect, or that the effect is more noticeable in high achievers – those who identify strongly with the subject as well as the stereotype will be affected more by the threat. It is noteworthy that boys are also subject to the effects of ST. For example, Huguet and Régner (2009) also demonstrated better test performance in boys that were told they were taking a geometry test compared to boys that were told they were completing an art exercise.

To mitigate ST effects, Steele (1997) proposed an approach called 'wise schooling'. This is an approach to education that seeks to establish strong teacher-student relationships and positive role models, to emphasise potential student ability rather than focusing on their limitations, emphasis on intelligence as something malleable rather than fixed or deterministic, and workshops to share personal experiences. Steele (1997) introduced a programme with these features as part of a controlled experiment. Outcomes demonstrated that African American students enrolled in a programme of 'wise-schooling' achieved significantly higher grade point averages (GPAs) than others not enrolled on the programme. Importantly, African Americans' test scores were more consistent with European-American participants on the programme, demonstrating that a more egalitarian approach to education results in more egalitarian outcomes.

[6] The DPP is additional funding provided to publicly funded schools in England to raise the attainment of disadvantaged pupils (https://www.gov.uk/guidance/pupil-premium-information-for-schools-and-alternative-provision-settings).

Cadinu, Maass, Rosabianca and Kiesner (2005) used structural equation modelling to characterise ST explicitly as performance-mitigating intrusive thoughts that occur during task performance. Sixty female university students were divided into two groups and asked to complete a series of mathematics questions. The experimental group were informed that studies had demonstrated that males performed better at such tasks. The control group were informed that there was no measurable difference between males and females in task performance. Participants were asked to record any intrusive thoughts that occurred to them as they completed the task. The authors found that the experimental group performed significantly worse than the control group, and exhibited a higher incidence of negative thought intrusion. Their model suggested that negative thought intrusions related to mathematics mediate the effects of ST on performance. That is, awareness of stereotype threat leads to an increase in negative domain-specific thinking, which in turn leads to a decrease in performance.

The ST hypothesis has been challenged, however, as a meta-analysis of all stereotype threat studies conducted prior to 2012 revealed that only 55 per cent of studies with adequate experimental designs in fact replicated the effect of stereotype threat on females' performance (Stoet & Geary, 2012). Others have suggested that ST effects have been inflated due to publication bias (Flore & Wicherts, 2015). Nonetheless, a subsequent and more recent review of several meta-analyses suggests that ST effects are in fact robust (Spencer, Logel & Davies, 2016).

Test Anxiety

Academic anxieties are another potential explanation for gender differences in performance. Test anxiety describes negative emotional reactions to evaluative settings and encompasses physiological (bodily reactions) and cognitive (intrusive thoughts) components (Hembree, 1988). On the other hand, mathematics anxiety and science anxiety describe negative emotional reactions elicited specifically by mathematics and science, respectively (Hembree, 1990; Mallow, 1994). Test anxiety, mathematics anxiety and science anxiety are interrelated yet are considered distinct constructs (Mallow, 2006).

Importantly, an abundance of research has shown that girls and adult females report higher test anxiety, maths anxiety and science anxiety than do boys and adult males, and gender differences in test and maths anxiety emerge in the primary school years (e.g., reviewed in Devine, Hill, Carey, & Szűcs, 2018; Hembree, 1988; Hembree, 1990). Furthermore, many studies have revealed moderate negative associations between academic anxieties and academic achievement/test performance (Hembree, 1988; Hembree, 1990). However, research has not consistently shown that the academic anxiety-performance relationship differs by gender. For example, some studies have shown stronger correlations in girls than boys (e.g., Devine, Fawcett, Szűcs, & Dowker, 2012), yet other studies have revealed no gender differences in the correlation between academic anxiety and performance (e.g., Cassady & Johnson, 2002; Ma, 1999), or revealed stronger correlations in males than females (McCarthy & Goffin, 2005). Academic anxieties have also been linked to task avoidance and course drop-out (Hembree, 1990).

Participation in STEM Subjects

Compounding the problem of unequal participation in STEM courses and careers is the finding that women who are equally able to their male counterparts choose to leave mathematics and science majors at a higher rate than men (Stage & Maple, 1996). This is known as the 'leaky pipeline' phenomenon.

Hyde et al. (2008) concluded that the discrepancy between men and women at the top end of the distribution is unable to account for the imbalance in representation in some STEM academic pathways. For example, university engineering courses average only 15 per cent female participation in the United States. Thus, test scores only play part of the story in determining female representation in STEM subjects as those who possess the aptitude clearly get 'lost' somewhere along this trajectory. Additionally, those who do elect this pathway have a higher drop-out rate than their male peers. The *Women's Experiences in College Engineering* project, funded by the National Science Foundation and the Sloan Foundation, writes that the exit of many young women is not driven by ability, but rather by women negatively interpreting their grades and having low self-confidence. They also expressed dissatisfaction with the levels of competition, lack of support, and discouragement from faculty and peers (Goodman Research Group, 2002).

Pell (2014) identified four critical periods for the retention of women in science – early childhood, adolescence, college and the job entry period. From Cambridge Assessment Admissions Testing's perspective, it is the first three which will impact on take-up of science-based subjects and subsequent application for admission to STEM degrees. Active encouragement through education by parents, teachers and other mentors is crucial to both males and females deciding to pursue STEM careers. A lack of support for women in any of the critical periods identified will lead to under-representation at university.

Ceci, Ginther, Kahn and Williams (2014, p.75) argue that it is the second of these periods (adolescence) that is most pertinent with regard to female participation in academic science careers in the twenty-first century. The authors present evidence of increasing gender equality in representation in STEM subjects that, combined with improving test scores by girls in STEM subjects in school, can no longer explain under-representation in certain STEM programmes, at least in the United States context. Instead, they identify a trend that women are more likely to elect health and public-facing occupations instead of pursuing mathematics-intensive careers in academic science. The authors used the American Community Survey (2012), which provides census-based information regarding the United States population, to investigate degree attainment of women and men up to the ages of 30 to 35. They found that more women than men had master's degrees (27 per cent vs. 21.8 per cent) and professional degrees (9.1 per cent vs. 7.9 per cent) although of those employed, more than twice as many men were in science and engineering professions than women (37.4 per cent vs. 18.3 per cent). However, if health practitioners and educators are included, the proportion of men and women involved in science-based careers approaches equality (51.2 per cent vs. 45.2 per cent). The authors conclude that during adolescence, girls are more likely to make a conscious choice to pursue health-related professions even if their mathematics achievements match those of their male counterparts. Similarly, Morgan, Gelbgeiser and Weeden (2013) reported that gender differences in occupational plans expressed by high school seniors (from 2002 data) could not be explained by differences in math coursework in high school, although girls were more likely than boys to plan a biological/health occupation (27 per cent vs. 11 per cent).

On the other hand, Wang, Eccles and Kenny (2013) longitudinal research indicated that secondary school students' career choice was related to their *relative* mathematics and verbal skills in 12th grade. That is, students who performed highly in mathematics and verbal abilities were less likely to pursue STEM careers than were students with high mathematics and average verbal abilities; that is, those with higher mathematics abilities relative to other skills. The authors speculated that students with high abilities across both domains may have a wider range of career opportunities available to them than students who excel only in mathematics. As there were more female students in the group with high abilities across both domains, the authors suggested that the wider choice of

careers across both STEM and non-STEM fields available to females may contribute to females' underrepresentation in STEM careers.

In their review, Ceci et al. (2014) provide little evidence to support the idea that sex discrimination in employment and application procedures contributes to females' underrepresentation in STEM fields. Instead, their review suggests that gender differences in subject preferences and attitudes that emerge earlier on in learning may make a larger contribution to the differential uptake of STEM careers in males and females. However, as noted by Cheryan, Ziegler, Montoya and Jiang (2017), Ceci et al.'s (2014) review only touches on research addressing why gender differences in subject preferences and attitudes emerge in the first place. They suggest that such differences may relate to ability perceptions and to gender stereotyped messages about mathematics and science careers (Perez-Felkner et al., 2012 as cited by Ceci et al., 2014).

Indeed, research by Elwood (2005) suggests that teachers' perceptions of students influence tiering decisions in GCSE mathematics, which may contribute to the lower rates of girls studying mathematics beyond GCSE in the U.K. When GCSE mathematics had three tier options, more girls were entered for the intermediate-level examination (for which the maximum grade is B) than the higher tier, due to teachers perceiving girls as being more anxious and less confident in mathematics (Stobart, White, Elwood, Hayden, and Mason, 1992). Thus, entry decisions were based on perceived emotional factors rather than girls' competence. As schools often have minimum grade requirements for studying A-level mathematics (e.g., A grades), entry of more girls into the intermediate tier may have resulted in a disproportionate number of girls being unable to continue with mathematics to A-level. Although Bramley et al. (2015) reported equal proportions of girls and boys entering mathematics and other STEM subject GCSEs in 2015, mathematics entries were not reported separately for the foundation and higher tiers.

More recently, research by Vitello and Crawford (2018) investigated gender differences in tier entry for GCSE subjects in student records in the National Pupil Database, a database containing exam results and demographic data for pupils in England. Vitello and Crawford found that the odds of female students being entered into the foundation tier was 1.3 times the odds for males for science GCSEs, and over 2.6 times the odds for males for mathematics GCSEs. These authors noted that the gender effect was reversed for language GCSEs, with males being more likely to be entered to the foundation tier than females for language GCSEs. Further research is needed to investigate the impact of tiering on girls' uptake of STEM subjects beyond GCSE. Smithers (2014) notes that at A-level, males outnumber females in mathematics, further maths and physics entries, however the proportions of males and females entering some other STEM subjects such as chemistry and geography is more balanced, while females outnumber males in biology entries.

Another potential influencing factor in girls' participation in STEM fields is the lack of self-esteem that women are claimed to feel during adolescence (Eccles, 1998). In 1992, the American Association of University Women (AAUW) published a report entitled *How Schools Short-change Girls.* This report collated the findings of more than 1,300 published studies and synthesised the findings with a national survey (*Short-changing Girls, Short-changing America*) which explicitly linked girls' self-esteem with their school environment and the content of what they learned in the classroom. This directly led to the Goals 2000: Educate America Act, a reform passed in 1994. The authors claimed that this was unintentionally caused by a competitive environment in which boys dominate discussion (boys are eight times more likely to call out in class than girls) and by devaluing classroom participation by girls when it occurred (boys are more likely to receive precise comments on their contributions than girls) (Sadker & Sadker, 1994). The AAUW report concluded that lower self-esteem would result in lower confidence, lower academic achievement and therefore different choices regarding academic pathways.

However, this picture is not universally accepted. Although Maccoby and Jacklin's (1974) study of gender differences identified males as more assertive, aggressive and less anxious, the effect size of the difference between the genders regarding *self-esteem* was very small (Feingold, 1994, p.449). Subsequent meta-analyses have confirmed that gender differences in self-esteem range from small to medium and emerge across cultures (e.g., Bleidorn et al., 2016; Kling, Shibley Hyde, Showers, & Buswell, 1999). Similarly, Harter, Waters and Whitesell (1997) and Harter, Waters, Whitesell and Kastelic (1998) analysed 'voice' in adolescent girls and boys to determine which group felt more able to express their opinions openly. Boys and girls aged 12 to 17 were asked to complete a questionnaire measuring voice. This questionnaire tapped the extent to which teenagers were able to express their opinions, share their thoughts, and express their values and feelings with peers, parents, and teachers. The authors found no significant differences between girls' and boys' scores in terms of 'voice' and found that girls and boys became more open and willing to express their voice as they got older. The authors concluded that both girls and boys felt they had a stronger voice when supported by parents, teachers and friends.

Self-competence beliefs such as academic self-efficacy (belief in one's ability to successfully carry out a task, Bandura, 1997) and self-concept (appraisal of oneself, which can be general or specific e.g., as a mathematician, Wills-Herrera, 2014) have also been linked to academic motivation and subject choice (Betz & Hackett, 1983; Eccles, 2011; Pajares, 2002). Research has shown that both constructs are related to academic achievement and that domain-specific self-efficacy is more strongly related to academic achievement than self-esteem (Mone, Baker, & Jeffries, 1995; Multon, Brown, & Lent, 1991).

Gender differences in self-beliefs have been mixed, with different patterns emerging depending on the subject of interest (Fan, 2011; Huang, 2013). A recent cross-national meta-analysis revealed that boys had higher mathematics self-efficacy than girls (Huang, 2013). Importantly, this difference was most marked at the upper secondary school level, supporting the conjecture that girls and boys have similar levels of mathematics self-efficacy during early schooling, but boys develop higher levels of self-efficacy by early adolescence (Pajares, 2002). Similarly, OECD's summary of the 2012 PISA assessment revealed that the gender gap was much wider in mathematics self-efficacy than in science self-efficacy, and was widest for items in which students had to rate their efficacy in solving problems with gender-stereotypical content (OECD, 2015a). OECD also reported that girls had significantly lower maths and science self-concept beliefs than boys. As mentioned earlier in this review there is evidence to suggest that children's self-concept in mathematics may be influenced by socialisers' beliefs and expectations (Eccles, 1982a).

Eccles' expectancy-value theory explains gender gaps in STEM subject uptake and performance in terms of psychological, social and cultural forces (Eccles et al., 1983; Eccles, 2011). This theory links two sets of beliefs to academic choices: an individual's expectancies (e.g., expectations for success) and their subjective task values (e.g., interest in a task, perceived usefulness). However, an individual's beliefs are influenced by various other factors including the cultural milieu (e.g., gender roles, gender stereotypes), a person's achievement-related experiences, affective reactions (e.g., anxiety), a person's goals and self-perceptions (e.g., self-concept), as well as socialisers' beliefs and expectations. Thus, the expectancy-value theory encompasses many of the ideas discussed throughout this review. This theory has received a lot of empirical support, particularly through longitudinal research (reviewed in Eccles, 1994; 2011; and Fan, 2011) and mathematics task values have been found to be linked to gender differences in STEM career attainment (Wang, Degol & Ye, 2015).

In relation to the abovementioned review by Ceci et al., (2014), Halpern (2014) argued that the gender gap debate should focus on gender gaps in specific disciplines on the basis that unequal progress has been made. For example, the magnitude of gender differences varies when comparing fields that are spatial-mathematics intensive (e.g., engineering, mathematics/computer science and physical sciences) and fields that are not (e.g., life science, psychology, and social science). Each STEM discipline now seems to face unique challenges. In a similar vein, more recent reviews of gender disparities in STEM participation have disaggregated STEM fields (e.g., Cheryan, Master, & Meltzoff, 2015; Cheryan et al., 2017).

Method Effect

Despite large gains made by female students, in many cases now surpassing males in test scores, this review has presented evidence of a common theme of males outperforming females in large-scale standardised tests which are predominantly formed of multiple-choice questions. This section will examine whether males and females respond differently to multiple-choice questions and what explanations are offered if this is the case.

Research in the 1980s indicated that gender bias was more pronounced in multiple-choice questions than in essay-style questions (Ferber, Birnbaum and Green, 1983; Heath, 1989), although these findings coincide with broader trends of gender differences which were prevalent during that decade. Ferber et al. (1983) found that males outperformed females on both multiple-choice and essay-style questions, whereas more recent test evidence has shown a marked improvement in female performance in all areas, most noticeably in essay-style questions, in which they now routinely outscore males.

Research has sought to address the method effect associated with multiple-choice questions in different subject areas. In the UK, Elwood (1999) found gender differences on GCSE coursework (which typically includes a range of question styles such as written papers, multiple-choice tests and practicals), and revealed that girls outperformed boys across several subjects. However, boys showed greater score variation than girls, and coursework performance had more of an impact on boys' overall exam scores than girls'. Elwood also noted that in general, coursework had less of an impact on final grades than was intended. More recently, Bramley et al. (2015) found that gender differences on written papers with multiple-choice questions varied by subject; however, only two specifications with multiple-choice questions were included in their analysis for comparison to written papers with other question types. Comparing a larger selection of written papers with short and long-answer questions, they reported that girls and boys performed equally on written papers comprised of short-answer questions, but girls performed better than boys on papers with long or essay questions.

Arthur and Everaert (2012) investigated the impact of examination format on performance in accounting and found that females outperformed males in constructed-response items, and to a lesser extent in multiple-choice items. The authors presented this as evidence that multiple-choice items favour males on the basis that the score difference between females and males would be greater if multiple-choice items had not been used in the test. However, assuming that the genders should perform equally by default, the results instead suggest that there is no gender bias in multiple-choice questions and that constructed response items appear to be biased in favour of females.

To explore the impact of different methodological characteristics, Buck, Kostin and Morgan (2002) investigated differential item functioning of gender in advanced placement exams for United States

history. The authors found differences between male and female test scores were based more on item content than item type. Differences between genders were found to be consistent for topic across item types (free response or multiple choice). The authors concluded that multiple-choice items were not intrinsically biased against female test takers. Additionally, Du Plessis and Du Plessis (2009) performed an experiment whereby multiple-choice items were paired with free-response item types in a first-year economics test for undergraduate students enrolled on an economics course. Similar to Buck et al, the authors found no evidence of differences between test scores for males and females on three different first-year undergraduate economics tests when conducted using multiple-choice items.

Various authors have noted the impact of penalising incorrect responses in multiple choice tests on males' and females' behaviour. Du Plessis and Du Plessis (2009) explored this question, although found no evidence of a guessing differential between males and females in multiple choice items, even in situations when incorrect guesses were penalised. Kelly and Dennick (2009) found the opposite for undergraduate medicine course performance though, noting substantial gender bias relating to 'true-false-abstain' (TFA) item types in which incorrect responses are penalised. The authors present longitudinal data to support their claims. Across eight years of data for the first two years of an undergraduate medical curriculum, multiple-choice questions were found to contain the highest disparity between genders. Males were found to be 16.7 times more likely to outscore females in tests containing this item type. The proposed explanation for this large discrepancy is the higher propensity among females to abstain, with a corresponding result that males outscore females by three per cent. This trend extended to true/false questions in anatomy and physiology. Female advantage was recorded in one year in which short-response items were included. The common argument is that multiple-choice items favour greater male risk-taking behaviour and more cautious response-solving strategies by female test takers (Kelly & Dennick, 2009).

Hirschfeld, Moore and Brown (1995) cites 'willingness to guess' as a source of gender bias in multiple-choice item types. The authors argued that tests which do not penalise test takers for incorrect responses show less gender bias than those which do. The authors concluded that this was because males were more confident than females in high-stakes testing environments.

This general finding was replicated by Baldiga (2014) who found that in situations in which test takers are penalised for incorrect responses, females respond to significantly fewer questions than males. However, Baldiga found no difference between males and females in terms of confidence (a self-reported measure for each item) or in participants' knowledge of the materials presented to them (United States and world history SAT). Given a male and a female with similar self-reported probabilities of getting a question correct, the female is one-third more likely to omit the question than the male (Baldiga, 2014, p.443) if incorrect responses are penalised. This difference is explained via different risk preferences associated with the high-pressure environment, although Baldiga concedes that this accounts for only 40 per cent of the observed gender gap.

Some sociological explanations are offered for these observations, including males' and females' differing conception of success in competitive environments (score maximisation versus incurring few penalties respectively), and socialised passivity in females which feeds through to exam performance. Suggested explanations remain conjectural and require further exploration. Arthur and Everaert (2012, p.473) note that evidence of multiple-choice item types favouring males over females has been demonstrated in English-speaking countries, although they note that studies examining this question tend not to control for other variables that may impact on performance, such as prior university experience or cultural and socio-economic variables.

Generally speaking, the pattern of gender differences that is often observed across the three BMAT

sections described in the introduction correspond with claims made about gender bias in relation to specific testing methodologies (e.g., an advantage for males on multiple-choice questions; an advantage for females on essay-style questions). However, if the inclusion of multiple-choice questions in high-stakes standardised tests favours males over females, then this trend would be observable across all instruments and subject areas that employ multiple-choice questions, although may be mitigated (or magnified) by already observed trends such as male relative advantage in spatial and numerical skills, and female advantage in verbal skills. The research literature rather suggests that multiple-choice questions themselves do not necessarily discriminate against female test takers, although penalising incorrect responses has a clear impact on how both males and females interact with these items.

Conclusions

The report has investigated current educational literature to identify the extent to which gender differences identified in performance in STEM admissions tests may reflect the performance of males and females in STEM subjects more generally. The conclusions are summarised with respect to the research questions.

1. To what extent do differences in test scores represent some truth in the domains of interest in terms of:
    a. Sampling?
    b. Ability?
    c. Test score variance?
    d. Developmental differences?

Prior research suggests that males show performance advantages in mathematics, science and spatial abilities, whereas females have higher verbal and reading abilities. Recent international meta-analyses have shown that mean gender differences in maths and science have diminished over time, and that effect sizes tend to be small or medium. Recent UK research indicates that girls and boys have similar performance in STEM subjects at GCSE level.

Larger gender differences may exist at the upper end of the performance distribution. In the UK, boys are not consistently overrepresented at the high end of the distribution in STEM subjects at GCSE level. With regard to A-levels, there is a slightly higher percentage of boys gaining A* in mathematics, further mathematics and chemistry than girls, but a slightly higher percentage of girls gaining A* in geography, biology and physics.

STEM admissions test gender differences may potentially reflect gender differences at the upper end of the performance distribution in maths and chemistry, as candidates opting to take the STEM admissions tests are likely to be within the highest performing students.

A male advantage in spatial abilities and males' tendency to use spatial problem-solving strategies is also relevant to some STEM admissions tests. Males show an advantage on problems which lend themselves to spatial strategies, yet this gender difference is reduced if spatial problem solving is encouraged in all students.

Some STEM admissions test problems require spatial problem solving, thus, spatial problem-solving strategies should be encouraged in preparation material for tests such as BMAT.

The gender difference in BMAT Section 3 likely reflects the pervasive female performance advantage in reading and verbal abilities and the tendency for females to perform better on essay questions than males.

2. To what extent do differences in STEM admissions test scores reflect a general societal bias against females in terms of:
    a. Participation in STEM subjects?
    b. Method effect?

Research suggests that many socio-cultural and developmental factors such as teacher gender, single-sex schooling, attitudes towards competition, academic anxiety, stereotype threat, self-beliefs and the gender-stereotyped beliefs of parents and teachers may influence performance and decisions to pursue scientific careers.

Gender differences in STEM admissions test performance may reflect the long-term influence of these socio-cultural and personal factors on females' STEM education. For example, potential, high-calibre female students may have dropped out of STEM subjects earlier in their education and/or opted not to apply for medical or scientific courses of study. Gender differences could also reflect the consequences of test anxiety (or other academic anxieties), or stereotype threat on performance during admissions test administration.

Collectively, the research therefore suggests that gender differences in STEM admissions test performance are likely to reflect gender differences in factors outside of the tests.

Caveats – Publication Bias and Replicability

The report concludes with two caveats that emerged during the course of the literature review.

1. One caveat associated with the literature review is the sampling of papers. As the mean difference between males and females overall is small, then a large number of studies would be expected to return negative findings. However, in the literature search, the overwhelming majority reported significant differences between males and females. Studies reporting no differences between males and females tend to go unreported in the literature.

2. There is also the problem of replicability; studies in specific locations with particular instruments are rarely replicated across different samples. Studies tend to stand in isolation. There may be sufficient differences in data collection practices to undermine claims that different studies support each other in terms of their findings if there are substantial differences in the conditions in which the research was conducted.

One final caveat is that this review was concluded in 2018 and has not been updated to include more recent studies. Therefore, it excludes any additional findings that address questions about gender gaps in STEM subjects, and other possible contributing factors which have been added to the growing body of research. Gender differences in STEM continues to be a topic that generates a lot of interest and debate, as policies move towards widening participation in higher education.

References

American Association of University Women. (1992). *The American Association of University Women Report: How Schools Short-change Girls*. Washington, DC: The AAUW Educational Foundation and National Association.

Aronson, J., Lustina, M. J., Good, C., Keough, K., Steele, C. M., & Brown, J. (1999). When white men can't do math: necessary and sufficient factors in stereotype threat. *Journal of Experimental Social Psychology, 35*, 29–46.

Arthur, N., & Everaert, P. (2012). Gender and Performance in Accounting Examinations: Exploring the Impact of Examination Format. *Accounting Education: An International Journal, 21*(5), 471–487.

Baldiga, K. (2014). Gender differences in willingness to guess. *Management Science, 60*, 434–448.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: Freeman.

Benbow, C. P., & Stanley, J. C. (1980). Sex differences in mathematical ability: Fact or artifact? *Science, 210*, 1,262–1,264.

Benbow, C. P., & Stanley, J. C. (1983). Sex differences in mathematical reasoning ability: More facts. *Science, 222*, 1,029–1,031.

Betz, N., & Hackett, G. (1983). The relationship of mathematics self-efficacy expectations to the selection of science-based college majors. *Journal of Vocational Behavior, 23*, 329¬345.

Bleidorn, W., Arslan, R. C., Denissen, J. J., Rentfrow, P. J., Gebauer, J. E., Potter, J., & Gosling, S. D. (2016). Age and gender differences in self-esteem – A cross-cultural window. *Journal of Personality and Social Psychology, 111*(3), 396–410.

Bramley, T., Vidal Rodeiro, C.L., & Vitello, S. (2015). *Gender differences in GCSE*. Cambridge Assessment Research Report. Retrieved from: www.cambridgeassessment.org.uk/Images/gender-differences-in-gcse.pdf

Buck, G., Kostin, I., & Morgan, R. (2002). *Examining the relationship of content to gender-based performance differences in advanced placement exams*. College Board Research Report 2002-12, ETS RR-02-25. Princeton: Educational Testing Service.

Cadinu, M., Maass, A., Rosabianca, A., & Kiesner, J. (2005). Why do women underperform under stereotype threat?. *Psychological Science*, *16*(7), 572–578.

Carrell, S, E., Page, M. E., & West, J. E. (2009). *Sex and Science: How Professor Gender Perpetuates the Gender Gap*. NBER Working Paper 14959. Retrieved from: www.nber.org/papers/w14959

Cassady, J. C., & Johnson, R. (2002). Cognitive test anxiety and academic performance. *Contemporary Educational Psychology*, *27,* 270–295.

Ceci, S. J., Ginther, D., Kahn, S., & Williams, W. M. (2014). Women in Academic Science: A Changing Landscape. *Psychological Science in the Public Interest, 15* (3), 75–141 (whole-issue monograph).

Cheryan, S., Master, A., & Meltzoff, A. N. (2015). Cultural stereotypes as gatekeepers: increasing girls' interest in computer science and engineering by diversifying stereotypes. *Frontiers in Psychology, 6* , Article number 49.

Cheryan, S., Ziegler, S. A., Montoya, A. K., & Jiang, L. (2017). Why are some STEM fields more gender balanced than others?. *Psychological Bulletin, 143*(1), 1–35.

Cohen, J. (1977). *Statistical Power Analysis for the Behavioral Sciences.* New York: Academic Press.

Dee, T. S. (2006). The why chromosome: How a teacher's gender affects boys and girls. *Education Next, 6*(4), 68–75.

Dee, T. S. (2007). Teachers and the gender gaps in student achievement. *Journal of Human Resources*, *42*(3), 528–554.

Devine, A., Fawcett, K., Szűcs, D., & Dowker, A. (2012). Gender differences in mathematics anxiety while controlling for test anxiety. *Behavioral and Brain Functions*, 8. doi: www.behavioralandbrainfunctions.com/content/8/1/33

Devine, A., Hill, F., Carey, E., & Szűcs, D. (2018). Cognitive and emotional math learning problems largely dissociate: Prevalence of developmental dyscalculia and mathematics anxiety. *Journal of Educational Psychology*, *110*(3), 431–444.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology, 56*, 5–18.

Du Plessis, S., & Du Plessis, S. (2009). *A new and direct test of the 'gender bias' in multiple-choice questions*. Stellenbosch Economic Working Papers 23/09. Stellenbosch: Stellenbosch University, Department of Economics.

Eccles [Parsons], J. S., Adler, T. F., & Kaczala, C. M. (1982a). Socialization of achievement attitudes and beliefs: Parental influences. *Child Development, 53*, 310–321.

Eccles [Parsons], J. S., Kaczala, C. M., & Meece, J. L. (1982b). Socialization of achievement attitudes and beliefs: Classroom influences. *Child Development, 53*, 322–339.

Eccles [Parsons], J. S., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1983). Expectations, values, and academic behaviors. In J. T. Spence (Ed)., *Perspective on achievement and achievement motivation* (pp. 75–146). San Francisco: W. H. Feeman.

Eccles, J. S. (1994). Understanding women's educational and occupational choices: applying the Eccles et al. model of achievement-related choices. *Psychology of Women Quarterly, 18*, 585–610.

Eccles, J. S. (1998). Perceived control and the development of academic motivation: Commentary. *Monographs of the Society for Research in Child Development, 63*(2/3), 221–231.

Eccles, J. S. (2011). Gendered educational and occupational choices: Applying the Eccles et al. model of achievement-related choices. *International Journal of Behavioral Development, 35*, 195–201.

Ellison, G., & Swanson, A. (2010). The gender gap in secondary school mathematics at high achievement levels: Evidence from the American mathematics competitions. *Journal of Economic Perspectives, 24*(2), 109–128.

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.

Elwood, J. (1999). Equity issues in performance assessment: The contribution of teacher-assessed coursework to gender-related differences in examination performance. *Educational Research and Evaluation, 5*(4), 321–344.

Elwood, J. (2005). Gender and achievement: what have exams got to do with it?. *Oxford Review of Education, 31*, 373–393.

Emery, J. (2013). *BMAT test-taker characteristics and the performance of different groups 2003–2012*. Cambridge Assessment Internal Report 1442.

European Commission. (2013). *The current situation of gender equality in Bulgaria – Country Profile*. Retrieved from: eige.europa.eu/gender-equality-index/2013/country/BG

Fan, W. (2011). Social influences, school motivation and gender differences: an application of the expectancy-value theory. *Educational Psychology, 31*, 157–175.

Feingold, A. (1988). Cognitive gender differences are disappearing. *American Psychologist, 43*, 95–103.

Feingold, A. (1992). Sex differences in variability in intellectual abilities: A new look at an old controversy. *Review of Educational Research, 62*, 61–84.

Feingold, A. (1994). Gender difference in personality: A meta-analysis. *Psychological Bulletin, 116*(3), 429–456.

Ferber, M. A., Birnbaum, B. G., & Green, C. A. (1983). Gender differences in economic knowledge: A reevaluation of the evidence. *The Journal of Economic Education, 14* (Spring), 24–37.

Fernando, J. (2022). *Opportunity Cost.* Retrieved from: www.investopedia.com/terms/o/opportunitycost.asp#:~:text=Opportunity%20cost%20is%20the%20forgone,and%20weighed%20against%20the%20others

Flore, P. C., & Wicherts, J. M. (2015). Does stereotype threat influence performance of girls in stereotyped domains? A meta-analysis*. Journal of School Psychology, 53*, 25–44.

Frome, P. M., & Eccles, J. S. (1998). Parents' influence on children's achievement-related perceptions. *Journal of Personality and Social Psychology, 74*, 435–452.

Gallagher, A. M., Levin, J. Y., & Cahalan, C. (2002). *Cognitive patterns of gender differences on mathematics admissions tests*. ETS Research Report No. 02-19. Princeton: Educational Testing Service.

Gneezy, U., & Rustichini, A. (2000). Pay enough or don't pay at all. *Quarterly Journal of Economics, 115*(3), 791–810.

Goldstein, H. (2004). International comparisons of student attainment: some issues arising from the PISA study. *Assessment in Education: Principles, Policy and Practice, 11*(3), 319–330.

González de San Román, A., & de la Rica, S. (2012). *Gender gaps in PISA test scores: The impact of social norms and the mother's transmission of role attitudes*. IZA Discussion Papers 6338, Institute for the Study of Labor (IZA).

Goodman Research Group. (2002). *Final Report of the women's experiences in college engineering (WECE) project – IWITTS*. Retrieved from: http://www.iwitts.org/proven-practices/topics/retention/271-final-report-of-the-womens-experiences-in-college-engineering-wece-project

Gortazar, L., Herrera-Sosa, K., Kutner, D., Moreno, M., & Gautam, A. (2012). *How can Bulgaria improve its education system?: An analysis of PISA 2012 and past results*. Washington, DC: World Bank Group. Retrieved from: documents.worldbank.org/curated/en/2012/09/20278281/can-bulgaria-improve-education-system-analysis-pisa-2012-past-results

Guiso, L., Monte, F., Sapienza, P., & Zingales, L. (2008). Culture, gender and math. *Science, 30*, 320(5880), 1164–1165.

Halpern, D. F. (1992). *Sex Differences in Cognitive Abilities* (2nd ed.). Hillsdale: Erlbaum.

Halpern, D. F. (2014). It's complicated – in fact, it's complex: explaining the gender gap in academic achievement in science and mathematics. *Psychological Science in the Public Interest, 15*(3), 72–74.

Halpern, D. F., Eliot L., Bigler, R. S., Fabes, R. A., Hanish, L. D., Hyde J., Liben, L. S., & Martin C. L. (2011). The pseudoscience of single-sex schooling, *Science, 333*(6050), 1706–1707.

Hannay, T. (2016). *Reducing the attainment gap: good ways and bad*. Retrieved from: www.schooldash.com/blog-1606.html

Harter, S., Waters, P. L., & Whitesell, N. R. (1997). Lack of voice as a manifestation of false self behavior among adolescents: The school setting as a stage upon which the drama of authenticity is enacted. *Educational Psychologist, 32*, 135–173.

Harter, S., Waters, P. L., Whitesell, N. R., & Kastelic, D. (1998). Level of voice among female and male high school students: Relational context, support, and gender orientation. *Developmental Psychology, 54*, 892–901.

Heath, J. A. (1989). An econometric model of the role of gender in economic education. *The American Economic Review, 79*, 226–235.

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

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research, 58*, 47–77.

Hembree, R. (1990). The nature, effects, and relief of mathematics anxiety. *Journal for Research in Mathematics Education, 21*, 33–46.

Higher Education Statistics Agency Limited. (2014). *Introduction – Students 2012/13*. Retrieved from: www.hesa.ac.uk/component/content/article?id=3129

Hines, M. (2011). Gender development and the human brain. *Annual Review of Neuroscience, 34*, 69–88.

Hirschfeld, M., Moore, R. L., & Brown, E. (1995). Exploring the gender gap on the GRE subject test in economics. *Journal of Economic Education, 26*(1), 3–15.

Huang, C. (2013). Gender differences in academic self-efficacy: a meta-analysis. *European Journal of Psychology of Education, 28*, 1–35.

Huguet, P., & Régner, I. (2007). Stereotype threat among schoolgirls in quasi-ordinary classroom circumstances. *Journal of Educational Psychology, 99*(3), 545–560.

Huguet, P., & Régner, I. (2009). Counter-stereotypic beliefs in math do not protect schoolgirls from stereotype threat. *Journal of Experimental Social Psychology, 45*, 1,024–1,027.

Hyde, J. S. (1981). How large are cognitive gender differences? A meta-analysis using omega-squared and d. *American Psychologist, 36*(8), 892–901.

Hyde, J. S., & Linn, M. C. (2006). Gender similarities in mathematics and science. *Science*, *314*, 599–600.

Hyde, J. S., & Mertz, J. E. (2009). Gender, culture and mathematics performance. *PNAS, 106*(22), 8,801–8,807.

Hyde, J. S., Lindberg, S. M., Linn, M. C., Ellis, A. B., & Williams, C. C. (2008). Gender similarities characterize math performance. *Science, 321*, 494–495.

IEGE. (2013). *Gender Equality Index – Country Profiles, European Institute for Gender Equality*. Retrieved from: eige.europa.eu/rdc/eige-publications/gender-equality-index-country-profiles

Jacobs, J. (1991). Influence of gender stereotypes on parent and child mathematics attitudes. *Journal of Educational Psychology, 83*, 518–527.

Kelly, S., & Dennick, R. (2009). Evidence of gender bias in true-false-abstain medical examinations. *BMC Medical Education, 9*, Article number 32.

Kling, K. C., Shibley Hyde, J., Showers, C. J., & Buswell, B. N. (1999). Gender differences in self-esteem: A meta-analysis. *Psychological Bulletin, 125*(4), 470–500.

Johnson, S. P. (2004). Development of perceptual completion in infancy. *Psychological Science, 15*, 769–775.

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.

Kreiner, S., & Christensen, K. B. (2014). Analyses of model fit and robustness. A new look at the PISA scaling model underlying ranking of countries according to reading literacy. *Psychometrika, 79*(2), 210–231.

Kronholz, J. (2004). Economic time bomb: U.S. teens are among the worst at math. *Wall Street Journal*. Retrieved from: www.wsj.com/news/articles/SB110236760101392346

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), 1,065–1,073.

Ma, X. (1999). A meta-analysis of the relationship between anxiety toward mathematics and achievement in mathematics. *Journal for Research in Mathematics Education, 30*, 520–540.

Maccoby, E. E., & Jacklin, C. N. (1974). *The Psychology of Sex Differences*. Stanford: Stanford University Press.

Machin, S., & Pekkarinen, T. (2008). Global sex differences in TES score variability. *Science, 322* (5906), 1,331–1,332.

Mallow, J. V (1994). Gender-related science anxiety: A first binational study. *Journal of Science Education and Technology, 3*, 227–238.

Mallow, J. V. (2006). Science anxiety: research and action. In J. J. Mintzes & W. H. Leonard (Eds.), *Handbook of College Science Teaching* (pp. 3–14). Virginia: National Science Teachers Association.

Matyas, M. L. (1985). Factors affecting female achievement and interest in science and in scientific careers. In: J. B. Kahle (Ed.) *Women in Science: A Report From the Field* (pp. 27–48). Philadelphia: The Falmer Press.

McCarthy, J. M., & Goffin, R. D. (2005). Selection test anxiety: Exploring tension and fear of failure across the sexes in simulated selection scenarios. *International Journal of Selection and Assessment, 13*, 282–295.

McFarlane, T. (2014). *UAE Economic Vision: Women in science, technology and engineering*. Retrieved from: dokumen.tips/science/uae-economic-vision-women-in-science-technology-and-engineering.html?page=1

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry, 1*, 108–141.

Meyer, H-D., Zahedi, K., & signatories. (2014). An Open Letter to Andreas Schleicher, OECD, Paris. *Global Policy Journal.* Retrieved from: journals.sagepub.com/doi/pdf/10.2304/pfie.2014.12.7.872

Mone, M. A., Baker, D. D., & Jeffries, F. (1995). Predictive validity and time dependency of self-efficacy, self-esteem, personal goals, and academic performance. *Educational and Psychological Measurement, 55*, 716–727.

Morgan, S. L., Gelbgeiser, D., & Weeden, K. (2013). Feeding the pipeline: Gender, occupational plans, and college major selection. *Social Science Research, 42*(4), 989–1,005.

Multon, K. D., Brown, S. D., & Lent, R. W. (1991). Relation of self-efficacy beliefs to academic outcomes: A meta-analytic investigation. *Journal of Counseling Psychology, 38*, 30–38.

Niederle, M., & Vesterlund, L. (2010). Explaining the gender gap in math test scores: the role of competition. *Journal of Economic Perspectives, 24*(2), 129–144.

Nowell, A., & Hedges, L. V. (1998). Trends in gender differences in academic achievement from 1960 to 1994: An analysis of differences in mean, variance, and extreme scores. *Sex Roles, 39* (1/2), 21–43.

OECD. (2012). *PISA 2009 Technical Report*. Retrieved from: dx.doi.org/10.1787/9789264167872-en

OECD. (2015a). *The ABC of Gender Equality in Education: Aptitude, Behaviour, Confidence*. Retrieve from: dx.doi.org/10.1787/9789264229945-en

OECD. (2015b). *Social Institutions and Gender Index*. Retrieved from: www.genderindex.org/country/bulgaria

Pajares, F. (2002). Gender and perceived self-efficacy in self-regulated learning. *Theory Into Practice, 41*, 116–125.

Pell, A. N. (1996). Fixing the leaky pipeline: women scientists in academia. *Journal of Animal Science, 74*(11), 2,843–2,848.

Pinker, S. (2002). *The Blank Slate: The Modern Denial of Human Nature*. New York: Viking.

Pope, D. G., & Sydnor, J. R. (2010). Geographic variation in the gender differences in test scores. *Journal of Economic Perspectives, 24*(2), 95–108.

Preckel, F., Goetz, T., Pekrun, R., & Kleine, M. (2008). Gender differences in gifted and average-ability students: Comparing girls' and boys' achievement, self-concept, interest, and motivation in mathematics. *Gifted Child Quarterly, 52*(2), 146–159.

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*, 645–662.

Rioux, C., Paré, A., London-Nadeau, K., Juster, R.-P., Weedon, S., Levasseur-Puhach, S., Freeman, M., Roos, L. E., & Tomfohr-Madsen, L. M. (2022). Sex and gender terminology: a glossary for gender-inclusive epidemiology. *Journal of Epidemiology and Community Health*, *76*(8). Retrieved from: doi.org/10.1136/jech-2022-219171

Sadker, M., & Sadker, D. (1994). *Failing at Fairness: How America's Schools Cheat Girls*. New York: Scribner.
Smithers, A. (2014). *A-Levels 1951–2014*. Report for the Centre for Education and Employment Research, University of Buckingham. Retrieved from: www.alansmithers.com
Soh, K. (2014). Test language effect in international achievement comparisons: An example from PISA 2009. *Cogent Education, 1*(1). Article 955247.
Sommers, C. H. (2000). The war against boys. *The Atlantic Monthly*. Retrieved from: www.theatlantic.com/magazine/archive/2000/05/the-war-against-boys/304659/
Spelke, E. (2005). Sex differences in intrinsic aptitude for mathematics and science?: A critical review. *American Psychologist, 60*, 950–958.
Spencer, S. J., Logel, C., & Davies, P. G. (2016). Stereotype threat. *Annual Review of Psychology, 67*, 415–437.
Stage, F., & Maple, S. (1996). Incompatible goals: Narratives of graduate women in the mathematics pipeline. *American Educational Research Journal, 33* (1), 23–51.
Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist, 52* (6), 613–629.
Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69*, 797–811.
Stevenson, H., & Newman, R. (1986). Long-term prediction of achievement and attitudes in mathematics and reading. *Child Development, 57*, 646–659.
Stobart, G., White, J., Elwood, J., Hayden, M., & Mason, K. (1992). *Differential Performance in Examinations at 16+: English and Mathematics.* London: SEAC.
Stoet, G., & Geary, D.C. (2012). Can stereotype threat explain the gender gap in mathematics performance and achievement?. *Review of General Psychology, 16*, 93–102.
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*, e57988.
Stoet, G., & Geary, D.C. (2015). Sex differences in academic achievement are not related to political, economic, or social equality. *Intelligence, 48*, 137–151.
Sullivan, A., Joshi, H., & Leonard, D. (2010). Single-sex schooling and academic attainment at school and through the lifecourse. *American Educational Research Journal, 47*(1), 6–36.
Summers, L. (2005). *Remarks at NBER conference on diversifying the science and engineering workforce*. Retrieved from: www.harvard.edu/president/news-speeches-summers/2005/remarks-at-nber-conference-on-diversifying-the-science-engineering-workforce/
Thorndike, E.L. (1906). Sex in education. *The Bookman*, *23*, 211–214.
Vitello, S., & Crawford, C. (2018). Which tier? Effects of linear assessment and student characteristics on GCSE entry decisions. *British Educational Research Journal, 44*(1), 94–118.
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*, 412–423.
Wang, M. T., Degol, J., & Ye, F. (2015). Math achievement is important, but task values are critical too: examining the intellectual and motivational factors leading to gender disparities in STEM careers. *Frontiers in Psychology, 6,* Article number 36.
Wang, M. T., Eccles, J. S., & Kenny, S. (2013). Not lack of ability but more choice: individual and gender differences in choice of careers in science, technology, engineering, and mathematics. *Psychological Science, 24*(5), 770–775.
White, P. E. (1992). *Women and Minorities in Science and Engineering: An Update*. Washington, DC: National Science Foundation.
Willingham, W. W., & Cole, N. S (2010). *Gender and Fair Assessment*. New York: Routledge. (Originally published in 1997 by Erlbaum).

Wills-Herrera, E. (2014). Self-esteem. In A. C. Michalos (Ed.), *Encyclopedia of Quality of Life and Well-Being Research* (pp. 5,772–5,774). Dordrecht: Springer Netherlands.

Yu, X., & Shauman, K. A. (2003). *Women in Science: Career Processes and Outcomes*. Cambridge: Harvard University Press.

Zahidi, S. (2013). *What makes the Nordic countries gender equality winners?* Retrieved from: www.huffingtonpost.com/saadia-zahidi/what-makes-the-nordic-cou_b_4159555.html

We are Cambridge Assessment Admissions Testing, part of the University of Cambridge. Our research-based tests provide a fair measure of skills and aptitude to help you make informed decisions. As a trusted partner, we work closely with universities, governments and employers to enhance their selection processes.

Cambridge Assessment
Admissions Testing
The Triangle Building
Shaftesbury Road
Cambridge
CB2 8EA
United Kingdom

Admissions tests support:
admissionstesting.org/help