Trends in Psychology
https://doi.org/10.1007/s43076-025-00504-5

REVIEW

Factors Related to Mean Score Group Differences on Cognitive Ability Tests: A Systematic Review

Stephen Cuppello[1] · **Lara D. Zibarras**[1] · **Philip J. Corr**[1]

Received: 30 August 2024 / Revised: 16 September 2025 / Accepted: 21 October 2025

Abstract

Commonly used in employee selection, cognitive ability tests (CATs) show mean score group differences by gender, ethnicity and socioeconomic status (SES). Prior research supports both substantive accounts, such as biological and environmental factors, and methodological and contextual accounts, such as stereotype threat and differential item functioning, related to group differences. This systematic review explores the evidence for the theoretical explanations that purport to explain these differences. Two hundred and twenty-five papers met inclusion criteria which explored potential explanations of group differences. These were grouped into 10 factors: biological, environmental, latent trait / measurement invariance, criterion validity, item bias, behaviour in test, anxiety, attitudes, experience, and stereotype threat. Results showed that whilst biological and environmental factors are related to group differences, so are several factors relevant to test design and administration, most notably stereotype threat, differences in domain experience and exposure and self-confidence. Theoretical and practical implications for the occupational use of CATs as well as limitations and recommendations are discussed.

Keywords Cognitive ability tests · Intelligence · Gender · Ethnicity · Socioeconomic status · Employee selection

Introduction

Cognitive ability tests (CATs) are used as part of selection and recruitment due to their utility in predicting workplace performance (Bertua et al., 2005; Sackett et al., 2022; Schmidt & Hunter, 1998), though widely reported mean score group differ-

✉ Stephen Cuppello
stephen.cuppello@city.ac.uk

[1] Department of Psychology and Neuroscience, City St George's, University of London, Northampton Square, London EC1V 0HB, UK

Published online: 11 November 2025

ences threaten to undermine their practical utility (e.g., Anastasi, 1958; Bates et al., 2013; Neisser et al., 1996; Roth et al., 2001). These differences have the potential to further entrench workplace inequalities (Woods & Patterson, 2024). Whilst influential research has suggested these differences are not a result of test bias (e.g., Jensen, 1980), subsequent research has found significant limitations in methods used (e.g., Dolan et al., 2004; Wicherts, 2017), opening up the possibility that group differences may be partially explained by biases in tests or testing processes.

Substantive accounts, such as genetic group differences (e.g., Warne, 2021) or socio-cultural differences (e.g., Hausdorf & Robie, 2018), are based on biological or environmental factors that are unlikely to be mitigated through changes to testing procedures. However, methodological and contextual accounts, such as stereotype threat (e.g., Steele & Aronson, 1995) or differential item functioning (DIF; e.g., Abad et al., 2004), suggest potential avenues through which changes to the administration of CATs could minimise mean score group differences. This paper aims systematically to review prior studies that look to explain mean score group differences in CATs and critically evaluate the evidence base for the various theoretical accounts of them. The ultimate aim is to identify methodological and contextual aspects of testing that could be improved to make tests fairer without undermining their utility.

Cognitive Ability and Mean Score Group Differences

Attempts to understanding the utility of measuring cognitive ability is not new. Binet and Simon (1907) looked to measure intelligence in children using such tests, and in doing so concluded that intelligence could be measured distinctly from culture and education. Spearman first argued that intelligence can be understood as a single general factor (*g*; Spearman, 1904, 1927). Since then, influential authors, such as Thurstone (1938) and more recently Carroll (1993), have proposed hierarchical models of intelligence with an overarching general intelligence factor at the top. In a review of 85 years of research into recruitment selection processes, Schmidt and Hunter (1998) concluded that scores on CATs were the strongest predictor of job performance when looking at candidates without prior experience in the role.

Researchers have explored mean score group differences from the early years of cognitive ability testing. In a seminal book on differential psychology, first published in 1937, Anastasi (1958) reported findings from studies demonstrating notable ability differences by socio-economic class as well as gender differences on specific cognitive domains. However, the author issued a caution with regards to the use of descriptive data to suggest ethnic differences on test scores *result* from ethnicity, rather than environmental factors. Hyde (1990) noted it that the first intelligence tests by Binet and later in the US by Terman assumed no gender differences in intelligence and so were constructed with the aim of having similar average scores for girls and boys.

A meta-analysis by Roth et al. (2001) looking at ethnic differences in occupational testing found support for previous findings that there is around a one standard deviation difference in mean test scores between Black and White samples, with smaller differences found when comparing Hispanic and White samples. Furthermore, the authors note important modifiers of these relationships, such as job complexity, study design, applicant vs incumbent samples, and constructs measured. A more recent

meta-analysis by Sackett and colleagues (2022) reported a slightly smaller but still large effect size when looking at Black-White CAT score differences in occupational samples (d=.79). Neisser et al. (1996) reported that most CATs are constructed so there are no overall gender differences; however, differences do exist within some cognitive domains, such as in spatial and numerical domains where men typically perform better, and in verbal domains where women typically perform better. It is difficult to consider gender differences in intelligence when tests are constructed to avoid such differences; however, this review will focus on differences observed in existing and popular intelligence testing. Adults from lower SES backgrounds typically score modestly lower, with childhood SES found to explain around 5% of adult intelligence (Bates et al., 2013). Overall, ethnic differences in mean intelligence scores are the largest, SES are modest, and gender differences are not typically seen in *g* but are found on specific cognitive domains.

The terms "gender" and "sex" have often been used interchangeably in the literature making distinctions between biological sex and socially constructed gender difficult. Similarly, the terms "race" and "ethnicity" are often used interchangeably, however in certain regions such as the USA these terms importantly have distinct meanings with "race" relating to biological difference between groups and "ethnicity" relating to cultural differences. This is further complicated when reviewing literature from multiple countries, where majority or culturally dominant ethnic groups are different. Given the intention to review all extant literature, this review explores papers about each of these terms, whilst acknowledging the inherent limitations. For the purposes of this review, the terms "gender" will be used to refer to both socially constructed gender and biological sex. No papers in this review discuss differences between participants' identified genders and the biological sex they were assigned at birth, and so these will be assumed to be the same. The term "ethnicity" will be used to cover both biological race and cultural ethnicity. Groups considered to be racially distinct are typically considered to be ethnically distinct, whereas the opposite is not always true, such as Hispanic-White and Non-Hispanic-White groups in the US.

A further point to note is that there are both similarities and differences that exist in the literature when considering different demographics. Some factors have a stronger evidence base for certain demographics than other. This is important to consider in this review. If for example several compelling studies were found supporting a particular factor's contribution to gender score differences on CATs, the lack of studies investigating ethnicity would not indicate that this factor is not relevant, only that there is a gap in the literature. It is for this reason that this paper will consider the case for different demographic groups for a factor together.

Factors Relating to Mean Score Demographic Group Differences

When looking at explanations for differences, there is little consensus in the literature, and so this forms the main focus of this review. In 2001, Hough et al. reviewed literature looking directly at factors related to group differences on tests used in occupational contexts. They found ethnicity, gender and age differences on cognitive domains in line with studies cited above; and they then looked to review the existent

literature to explore explanations of differences and potential strategies to overcome them.

They first explored the extent to which cultural differences may explain group differences. DIF, whereby two individuals possessing identical levels of a latent trait but belonging to different groups have different chances of getting a given question correct, was generally found to have a weak impact on overall test scores where it existed (Hough et al., 2001). They also found evidence suggesting culture-free, non-verbal ability tests still showed modest group differences and that attempts to address the confound of ethnicity vs culture by building tests with Black cultural components found little success in overcoming differences. They concluded that cultural differences have no more than a modest impact on test score differences.

In reviewing test coaching, designed to give test takers more familiarity and more favourable impressions of the testing process in an effort to increase motivation and reduce anxiety, the authors (Hough et al., 2001) also found less than compelling findings. Little research had been done in occupational contexts and whilst some studies demonstrated potential in addressing group differences, effects were modest and inconsistent. Findings did seem to be more consistent where test takers were less familiar with tests.

Test perceptions have an important influence on test scores (Hough et al., 2001). White people were typically more motivated to do well on tests, were more likely to believe tests would influence their future relationship with the company that asked them to take them and had more positive perceptions of the criterion validity of ability tests. Women and older people have greater test anxiety. Group differences in test performance lowered when motivation scores were partialled, test motivation had a moderating effect on test validity and validity perceptions positive correlated with test performance. The authors (Hough et al., 2001) presented only a few studies that had been published at the time; findings were sometimes inconsistent and typically suggested reducing differences in test perceptions would have, at best, a small effect on test score differences.

A further point made by Hough and colleagues (2001) was around stereotype threat whereby the existence of a negative stereotype around performance on a particular task can influence a person's ability to perform that task, through mechanisms such as changes in test anxiety or motivation. At the time, this was a newly emerging area of research, and whilst the authors presented some promising early studies, it was difficult to infer how big an impact stereotype threat could have on high-stakes test performance. Since then, whilst a sizable body of evidence has supported the effects of stereotype threat, it is recognised that it does not fully explain group mean score differences (e.g., Kirnan et al., 2009) and some research suggests effects are overstated (e.g., Finnigan & Corker, 2016).

Hough and colleagues (2001) end with a review of various methods for statistically evaluating potential adverse impact that goes beyond examining mean test score differences on which most previous research relies. They cite the importance of future research to explore group differences in criterion validity, such as prediction

of work performance, group differences in reliability, group differences in latent variable relationships, construct equivalence between groups and DIF.

A more recent review by Berry et al. (2011) specifically focused on understanding differential criterion validity for CATs between different ethnicities. Whilst not aiming to be exhaustive, the authors offered four potential explanations of differential validity. Two of which were accounted for by Hough et al. (2001), namely psychometric characteristics of tests or criteria such as test error and bias, and contextual factors such as stereotype threat. Range restriction was also posited to account for differential criterion validity although this is less relevant when considering mean score differences. Finally, the paper suggests true differences between how test scores predict performance, citing a well-established phenomenon whereby standardised test scores underpredict female college performance, as female grades are more driven by effort and conscientiousness than those of men.

There are further considerations not acknowledged by Hough et al. (2001). Hunter and Hunter (1984) cited three papers showing that Black people's work performance is similar, or lower, than levels which their cognitive ability would predict. They suggested these differences are likely due to poverty and hardship. Socioeconomic status (SES) and educational opportunities are likely to confound the relationship between ethnicity and cognitive ability. Test experience and prior domain experience are other potential avenues for exploration. Baenninger and Newcombe (1989) found some support that gender differences in spatial ability can be explained by prior experience. It merits research attention to consider the effect of prior exposure to testing and test domains on test performance.

Current Review

Exploring the factors influencing demographic group mean score differences on CATs holds profound implications for the development and administration of equitable and valid recruitment and selection tests. Hough et al.'s (2001) review highlighted emerging factors that had then garnered limited research attention or exhibited inconsistent findings. Over the past two decades, academic understanding has significantly evolved, emphasising the enduring importance of understanding factors contributing to test score differences for fairness in employee selection. Thus, there is a pressing need for a contemporary and comprehensive review. This paper aims to systematically examine existing literature on factors related to mean test score differences among demographic groups on CATs. The primary research question is therefore: what factors can be identified, with a robust evidence base, as potential explanations of group differences in occupational testing? In establishing this, we aim to pinpoint promising methodological and contextual approaches with compelling empirical support, offering potential pathways to reduce these differences. Implementation of such strategies in employee selection processes has the potential to foster fairer use of CATs.

Method

The current review was conducted according to the systematic review protocols described in the Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA-P; Moher et al., 2015). The review was not pre-registered with PROSPERO or similar.

Search Strategy

A broad search strategy was needed to gain an understanding of the factors related to mean score group differences on CATs, and to identify potential avenues to mitigate these differences. A literature search was conducted using the PsycINFO, PsychArticles, Web of Science and Business Source Ultimate databases. In order to identify multiple and competing viewpoints on potential factors, broad search terms were used due to the expansive, disparate and conflicting nature of extant literature. Search terms were chosen based on factors identified in an initial literature review summarised in the introduction to this paper. Additional, more generic search terms were included to capture factors not currently identified.

The following search terms were used, which were lightly amended based on each databases syntax: (("Racial and Ethnic Differences" OR “Age Differences” OR "Human Sex Differences" OR (race or ethni* or sex or gender or “socioeconomic status” or SES)) AND ("test motivatio*" or "test-taking motivatio*" or "test perceptio*" or “test experience” or "test-expectancy" or "test familia*" or “test coaching” or "test orientat*" or "test prepar*")) OR (((“Cognitive Assessment” or ((cognitive abilit* or intelligen* or IQ) AND (tes* or measur* or asses*))) not yout* not chil* not adoles* not health not clinical not impair* not aging not emotio* not physical not disab*) AND (“stereotype threat” or "test bia*" or "differential item functio*" or “DIF”)) OR (((“Cognitive Assessment” or ((cognitive abilit* or intelligen* or IQ) AND (tes* or measur* or asses*))) not yout* not chil* not adoles* not health not clinical not impair* not aging not emotio* not physical not disab*) AND ("Racial and Ethnic Differences" OR “Age Differences” OR "Human Sex Differences" OR (race or ethni* or sex or gender or “socioeconomic status” or SES)) AND (“construct invariance” or “construct equivalence” or “measurement invariance” or “measurement equivalence” or "self-efficacy" or “test validity” or “criterion validity” or “predictive validity”)).

Backward citation searching was conducted by reviewing identified papers’ reference list for any additional studies that may meet inclusion and exclusion criteria. Forward citation searching was conducted using Google Scholar. Given the large volume of papers reviewed, and given the large number of citations particularly influential papers had received, it was considered unfeasible to view all papers citing each identified paper. Therefore, the first 20 papers and the first 10 paper published within the last five years were reviewed for each paper. Whilst restricting the forward citation search in this way may have inadvertently led to some research meeting inclusion criteria being missed, this approach prioritises relevant papers while acknowledging that not all citations could be reviewed.

Inclusion and Exclusion Criteria

To meet inclusion criteria, articles needed to offer insights into explanations of group mean score differences on CATs, contain original research and be published in a peer-reviewed journal. Articles that were descriptive of differences but did not offer explanations for differences were excluded, as were papers offering approaches for measuring specific biases that did not offer explanations, because these papers were not considered helpful in answering this paper's primary research question. Studies with samples aged less than 18 years, clinical samples, non-human samples and studies exploring cognitive decline were also excluded. Whilst insights might be gained from such research, a decision was made to exclude these papers to narrow down results to healthy, working-aged people to better answer the research question. Articles were also excluded if they were exclusively based on tests of knowledge or crystalised intelligence, as well as those based on standardised academic tests such as SATs. Many CATs include both fluid and crystalised intelligence components, and studies using such tests were included. However, there is a large body of research concerning standardised education testing in the US as part of college admissions. The nature of these tests is significantly different from occupational CAT use. Test preparation and test practice are much more important, and access to these is likely to vary by demographic group. Inclusion would likely skew or conflate findings, especially considering that these papers would likely make up a majority of the papers included in this review. Finally, papers that were not peer-reviewed or not in English were excluded. Meta analyses and systematic reviews were not directly included but reference lists were reviewed for inclusion. Comment papers without novel evidence were excluded. Due to the long history of research in this field, restrictions were not placed on the time of publication. A summary of inclusion and exclusion criteria can be seen in Table 1.

Data Extraction and Synthesis

Paper selection, data extraction and data synthesis were completed by the first author. These processes were reviewed by the second and third authors. The initial search identified 1124 papers, 316 were excluded due to duplication, 541 were screened out based on abstracts and 140 were excluded for not meeting inclusion and exclusion criteria. The 127 remaining papers were fully reviewed. An additional 98 papers were identified as meeting criteria that were either cited in one or more of these papers or

Table 1 Outline of Inclusion and Exclusion Criteria for Literature Search

Inclusion criteria	Exclusion criteria
1. Population: adult humans (mean age 18+)	1. Articles merely descriptive of differences
2. Original research	2. Articles only offering approaches for measuring specific biases
3. Published in English in peer-reviewed journal	3. Clinical or cognitive decline samples
4. Time period: unrestricted	4. Articles primarily using tests of knowledge or standardised tests (e.g., SATs)

themselves cited one or more of these papers, totalling 225 papers. A full flow diagram of included papers can be seen in Fig. 1.

In order to group papers into factors, each was systematically tagged based on the specific account or accounts that potentially gave explanations for group mean score differences. The list of tags developed as more papers were reviewed, and theoretically similar accounts were merged in order to allow them to be critically reviewed together. This resulted in ten factors, namely: biological differences, environmental differences, latent trait differences/measurement invariance, criterion validity differences, item bias, test taking behavioural differences, test anxiety, test attitudes and perceptions, test experience and stereotype threat.

Data on these factors, the groups to which the research pertains (gender, ethnicity, age, SES), country, sample and relevant findings were extracted directly from published papers. In addition, a variable labelled 'Impact' was extracted stating whether or not the article found that the factor may have more than a negligible impact on group test scores (yes, no, maybe). Risk of bias was conducted using the Mixed Methods Appraisal Tool (MMAT; Hong et al., 2018), as it provides a single framework for which to appraise the studies included in this review employing varying

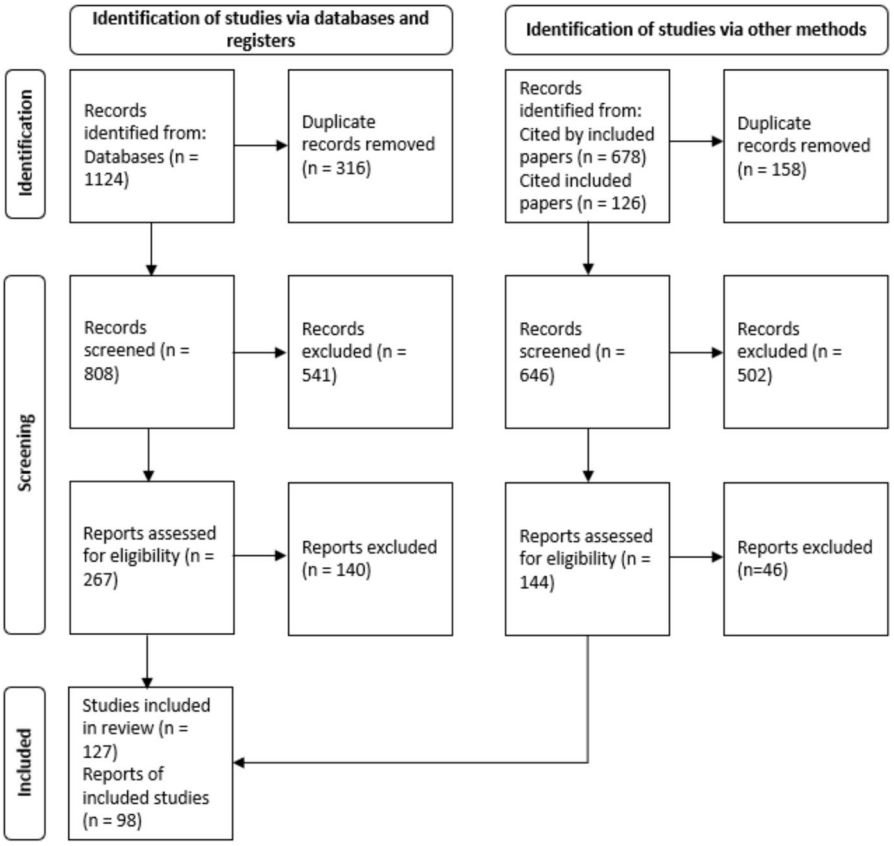

Fig. 1 Flow Diagram of Papers Considered in this Review

methodologies. Given the sizable heterogeneity of literature in this field, this tool was selected as it allowed for a consistent approach for evaluating the quality of papers. It is reliable and efficient (Pace et al., 2012) and has been previously employed to assess risk of bias in systematic reviews (e.g., Gledhill et al., 2018). Authors discourage calculating overall scores as potentially being reductionist, though these are presented for transparency. Finally, a variable labelled 'Ecological Validity' captured whether or not the paper was based on, or closely resembled, genuine test use such as high-stakes occupational testing (yes, no, partially). Papers were stored in reference management software, Mendeley, and extracted data were stored in a protected spreadsheet in Microsoft Excel. Narrative synthesis based on factor was chosen to analyse data (Popay et al., 2006). Given the heterogeneity of studies to be reviewed, quantitative synthesis was not considered appropriate. This approach was chosen to allow a critical evaluation of existing literature supporting or refuting the relationship between each factor and group mean score differences in CATs, in order to address the primary aims of this paper. For each factor, proposed explanations for differences will be described with supporting and countering evidence which will then be summarised. Increased weighting will be given to papers evaluated as having a lower risk of bias based on MMAT evaluation and accounts with replicated findings across multiple papers. Give the size of this review, each paper will not be individually critiqued, although MMAT summary scores for each paper can be found in Appendix A Table 4.

Results

Table 2 shows a summary of results, including the percentage of papers finding each factor has a meaningful impact on group differences, breakdowns of the demographics studied, evaluation of quality, ecological validity, student samples and US based samples. Of the 225 papers, 55% employed exclusively US samples, 8% Canadian, 6% German, 4% British, 20% from other single counties and a further 7% used samples from multiple countries. Seventy one percent of studies were based on majority or exclusively student samples. This proportion varied significantly by factor, with the most common use in Anxiety (94%) and Attitudes (91%) and the least common use in Criterion Validity (0%) and Environmental Differences (50%). Only 14% of studies met the criteria for ecological validity. This again varied by factor with Criterion Validity meeting these criteria most frequently (94%) followed by Latent Trait/Measurement invariance (23%). The two factors meeting these criteria the least were Biological (0%) and Stereotype Threat (4%).

The search strategy included search terms relating to age, and it was anticipated that this review would look to explore factors contributing to age differences in CAT scores. However, only two papers (Frisby & Osterlind, 2006; Reeve et al., 2008) measured age. In both papers, reliable evidence supporting causal factors for age differences were not found. Because of this, results will not be presented for age.

In terms of methodological quality, 147 studies were classified as quantitative non-randomized, 79 as quantitative randomized controlled trials and one as quantitative descriptive. The overall average MMAT score was 43% and this did not vary much

Table 2 Summary Information for Papers Relating to Each of the Ten Factors

		Impact			Demographics						
Factor	Papers	Yes	Maybe	No	Gender	Ethnicity	SES	MMAT	Ecological Validity	Student Sample	US Sample
Biological	63	67%	21%	13%	61	2	0	42%	0%	61%	35%
Environmental	15	93%	7%	7%	8	7	3	42%	6%	50%	38%
Latent Trait/Measurement invariance	13	31%	23%	46%	2	12	0	38%	23%	62%	23%
Criterion Validity	16	31%	44%	25%	5	15	0	41%	94%	0%	75%
Item Bias	23	48%	17%	35%	13	10	0	40%	13%	77%	57%
Behaviour in Test	14	79%	14%	7%	9	5	1	40%	14%	86%	57%
Anxiety	16	75%	13%	13%	12	6	1	39%	6%	94%	69%
Attitudes	23	52%	22%	26%	11	12	0	48%	13%	91%	74%
Experience	18	72%	28%	0%	16	2	0	42%	11%	83%	67%
Stereotype Threat	75	81%	8%	11%	51	27	7	45%	4%	90%	73%
Overall	225	82%	21%	20%	188	98	12	43%	14%	71%	56%

Notes: Papers = number of papers investigating that factor; Impact = proportion of papers finding that factor meaningfully impacted group differences; Demographics = number of papers investigating each demographic group; MMAT = average MMAT score for papers in that factor; Ecological Validity = percentage of papers in that factor meeting ecological validity requirements; Student Sample = proportion of papers for that factor using primarily student samples; US Sample = proportion of papers using US participants

by factor. Attitudes had the highest average (48%) and Latent Trait/Measurement invariance had the lowest average (38%). Evaluating the studies with the MMAT was useful in highlighting certain common potential risks of bias looking at the literature as a whole. In the quantitative nonrandomized studies, criteria 3.4 ("Are the confounders accounted for in the design and analysis?") was only met twice. Given the breadth of literature on group differences in CAT scores, it is very difficult for authors to effectively control for all potential confounders. Also, overreliance on student samples meant only 18% could be seen as representative of target populations. In quantitative randomized controlled trials, only 10% met condition 2.4 ("Are outcome assessors blinded to the intervention provided?") whilst only 11% met 2.2 ("Are the groups comparable at baseline?") as 80% did not provide this information. See the supplementary materials for details of authors, country, sample, demographic groups, factor, impact, relevant results, MMAT score and ecological validity score for each paper reviewed.

Below are summaries of papers by factor. Of the 225 papers, 181 only addressed one of the ten factors. Additional commentary from the 44 papers that addressed multiple factors is included at the end.

Biological

Sixty-three papers examined biological factors as an explanation for group differences in CAT scores. Several used brain scans during testing and found brain activation differences between women and men (e.g., Gur et al., 2000; Hugdahl et al., 2006; Jaušovec & Jaušovec, 2012; Jordan et al., 2002). However, studies did not universally support these difference (e.g., Halari et al., 2006; Unterrainer et al., 2000) and one paper found activation differences were inconsistent with cognitive domains that displayed gender differences (Bell et al., 2006). Research also provided some support for anatomical differences (Koscik et al., 2009) and differences in lateral differentiation between genders (Levy & Reid, 1978).

Two papers supported a genetic explanation for gender differences, with one finding modest support for a genetic basis of gender differences (Johnson & Bouchard, 2007) and a further study finding ethnic differences were more pronounced on items a twin comparison suggested were more heritable (Rushton et al., 2007a, 2007b). A further paper argued that replicating gender differences in spatial ability across 40 countries suggested an evolutionary basis (Silverman et al., 2007).

There are several accounts for developmental roots for gender differences in specific cognitive domains. Research has found that females with male twins outperform those with female twins on spatial tasks, supporting a prenatal masculinization hypothesis through hormonal transfer (Heil et al., 2011; Vuoksimaa et al., 2010) though this has not been supported in all papers (Toivainen et al., 2018). A further hormonal development account is inferred by 2D:4D ratios, which has been found to correlate negatively to prenatal testosterone and positively to prenatal oestrogen. Several papers have linked prenatal testosterone to numerical and spatial performance (e.g., Collaer et al., 2007; Kempel et al., 2005), which could suggest a cause for gender differences in these domains. Lastly, a longitudinal study found pubescent

testosterone levels were related to adult spatial performance in men which may partially explain group differences (Vuoksimaa et al., 2012).

A lot of research attention has been given to the role of sex hormones in gender differences in ability, although findings are often inconsistent. Testosterone levels have been linked to spatial performance in both genders (e.g., Burkitt et al., 2007; Silverman et al., 1999; Yang et al., 2007), although other studies have found no links with spatial, verbal or other gendered domains (e.g., Hirnstein et al., 2014; McKeever et al., 1987; Puts et al., 2010). Oestrogen has been linked to verbal performance (Rosenberg & Park, 2002), gonadotropins to verbal and spatial performance (Gordon & Lee, 1986) and androgens to spatial performance (Shute et al., 1983). Several experimental studies have manipulated hormone levels, such as administering testosterone resulting in increased spatial performance in women (Aleman et al., 2004), and administering luteinising hormone releasing hormone resulting in performance changes on several specific cognitive domains (Gordon et al., 1986). Finally, Griksiene and Ruksenas (2011) found verbal and spatial performance differences between women using different types of hormonal birth control suggesting a link between hormones and cognitive abilities.

Studies have also found compelling findings regarding monthly hormonal fluctuations, with most research focusing on women. Multiple papers found women's performance fluctuated throughout the menstrual cycle with better spatial performance in the menstrual phase and better verbal performance in the luteal phase (e.g., Hampson, 1990a, 1990b; McCormick & Teillon, 2001; Šimić & Santini, 2012), with some evidence suggesting female/male differences in spatial abilities were not found for women in the menstrual phase (Moody, 1997). Although not as pronounced, similar monthly effects have been found in men (Courvoisier et al., 2013). However, findings are not ubiquitous with some studies not finding a link between cycle time and performance (e.g., Epting & Overman, 1998; Gordon & Lee, 1993).

Out of the 63 papers in this factor, 61 related to gender providing strong support for biological accounts contributing to CAT score differences. Whilst sample sizes in these papers were often small, findings were frequently replicated across multiple studies. Effects were observed in papers scoring higher on the MMAT, suggesting research quality was not a factor influencing findings. However, no studies relating to this factor were conducted on high-stakes testing which may limit generalisability. Only 2 papers related to ethnicity and findings are less clear, and no papers included SES.

Environmental

Support for several environmental explanations of group differences were shown in 15 papers. Previous research has suggested that some of these factors may be genetically influenced (e.g., SES; Krapohl & Plomin, 2016), though research is conceptually distinct enough to consider these studies separately from biological accounts. Parental education level (McKay et al., 2003), individual education level (Dolan et al., 2006) and income (Díaz et al., 2012), which vary by ethnicity and SES, were found to be related to cognitive ability. Indeed, ethnic differences were significantly reduced when controlling for years of education, language proficiency and genera-

tional immigration status (Hausdorf & Robie, 2018). Ethnic differences on cognitive domains were more pronounced on verbal tasks, suggesting environmental factors contribute to differences (Razani et al., 2007). Additionally, multiple papers have used twin samples to support the role of environmental factors in group differences (Johnson & Bouchard, 2007; Rushton et al., 2007a).

Building on the links between hormone levels and gender differences in specific cognitive domains outlined above, several papers detail oral contraceptive use and type influences CAT performance on verbal and spatial task (Griksiene & Ruksenas, 2011; Hampson et al., 2022), with women using oral contraceptives performing worse on spatial tasks than other women and men (Griksiene et al., 2018). Finally, preference for gendered spatial toys as a child was related to spatial performance as an adult (Moè et al., 2018), as has spatial play as a child (Frenken et al., 2016).

Overall, 8 papers related to gender, 7 to ethnicity and 3 to SES, and strong support was found that various environmental factors contribute to differences for each group. Samples sizes were typically larger and findings were consistent across papers scoring higher and lower on the MMAT, so findings were less likely to be due to methodological limitations. Similarly to biological research, environmental research was rarely conducted on high-stakes occupational testing.

Latent Trait/Measurement Invariance

Thirteen papers related to latent trait differences or measurement invariance, with significant discordance in findings. These papers were included as findings of either measurement invariance or group difference on latent traits, support that group differences in mean test scores are due to differences in ability, not test bias. Several papers focused on Spearman's hypothesis, or the Jensen Effect, which states that if items with the greatest *g* loading highly correlate with items that also show the greatest group differences by using the method of correlated vectors, this suggests that differences are, indeed, on the latent trait and not due to test bias. This effect has been widely supported when looking at ethnic differences (e.g., Nyborg & Jensen, 2000; Rushton & Skuy, 2000; Rushton et al., 2004; Te Nijenhuis & Van der Flier, 1997). However, this theory has been criticised as being inappropriate, with one study finding that the hypothesis was erroneously supported when comparing CAT items with personality test items (Wicherts, 2017). A further study used multi-group confirmatory factor analysis (MGCFA) on a data set previously used to support the hypothesis (cited above: Te Nijenhuis & Van der Flier, 1997) and was unable to find factorial invariance, undermining original conclusions (Dolan et al., 2004). Research has used MGCFA elsewhere to suggest CATs do not necessarily have measurement invariance between cultures (Cockcroft et al., 2015). A similar approach to these studies concluded that there was measurement invariance with regard to women and men, suggesting group differences were at the latent trait level (Dolan et al., 2006).

Measurement invariance has also been supported by finding high correlations between item difficulties between different cultures, suggesting group differences are in *g* rather than culturally determined (Rushton et al., 2007b). Beyond this, authors have variously found support that group mean score differences are related to differences in processing speeds in more elementary cognitive tasks and so not test bias

(Vernon & Jensen, 1984), that neuropsychological predictors of CAT performance differed between cultures which suggests different constructs were being measured (Fasfous et al., 2013) and that Black/White test score differences were not replicated on learning tasks potentially suggesting cultural bias (Grubb & Ollendick, 1986).

Given such a conflicting body of literature, latent trait and measurement invariance research currently does not offer much to address the research question for this paper. In all, 12 papers examined ethnicity and 2 examined gender. Whilst sample sizes were typically larger, the average MMAT score for this factor was the lowest for all factors, suggesting this body of research is methodologically weaker than others.

Criterion Validity

Of the papers that met inclusion criteria, 16 referred to criterion validity or predictive validity with all but one simulation study using data from high-stakes testing. Whilst differences in criterion validity would not directly suggest explanatory factors for group mean score differences, these papers were considered to meet inclusion criteria as differences in criterion validity could support the relevance of methodological and contextual factors. The key concept is that if some form or forms of test bias were impacting the scores of women, ethnic minorities and lower SES people, we would expect tests to underpredict their work performance. Many papers do not find this by ethnicity (Sackett et al., 2023) or gender (Ispas et al., 2010) with some suggesting that range restriction for minority applicants explains any differences found (Roth et al., 2014) and several papers have found that CATs overpredict performance for women and ethnic minorities (e.g., Roberts & Skinner, 1996). In this regard, however, some papers' findings were more ambiguous (Harville, 1996), found no predictive bias but found tests were less valid for Black applicants (Rotundo & Sackett, 1999) or found that differential validity was removed by controlling for education (Reeb, 1976).

Contrary to these findings, other authors have found that CAT performance was less predictive of training performance for Black recruits (Carretta & Doub, 1998; De Meijer et al., 2008) and less predictive of job performance (Gardner & Deadrick, 2008, 2012). Some have found modest support for underprediction of Black training performance (Houston, 1987; Te Nijenhuis & Van der Flier, 2000). Countering range restriction concerns, Berry et al. (2014) suggested that CATs are seldom used in isolation in recruitment decision; however, they concluded that underpredictions of minority performance was unlikely. Other notable findings were that weighting subtests differently can potentially reduce adverse impact without sacrificing criterion validity (Wee et al., 2014). Finally, Lefkowitz and Battista (1995) found that supervisors gave higher performance ratings to same-ethnicity staff and when controlling for this, the criterion validity of tests reduced. Papers cited here mostly assume no bias in the criterion, which is often subjective manager ratings.

Criterion validation studies benefited from mostly being based on high-stakes testing. In all, 15 studies looked at ethnicity and 5 at gender. Higher scoring MMAT papers found effects as did lower scoring papers and many studies were conducted on very large, occupational samples. Despite this, taken as a whole, criterion validation studies do not offer much to this paper's research question. It is not clear as to

whether CATs predict job performance equally for different groups which then does not suggest whether or not differences may be the result of test bias.

Item Bias

Twenty-three papers examined potential item bias or DIF. Generally, studies either did not identify biased items, or biased items only negligibly impacted overall test scores (e.g., Abad et al., 2004; Jensen & McGurk, 1987; Vanderpool & Catano, 2008). There were some other notable findings though. Prior knowledge can lead to ethnic biases in items (Fagan & Holland, 2002, 2007), although other research suggests there is potential to overcome this with practice items (Tanzer et al., 1995). Some research showed no ethnic differences on culture-free tests (Domino & Morales, 2000) although other research has found modest gender differences on similar tests (Arendasy & Sommer, 2012).

Item content has received some attention in the literature. Multiple studies found using human figures as item content reduced the gender gap in mental rotation performance (e.g., Alexander & Evardone, 2008; Doyle & Voyer, 2018; Jansen & Lehmann, 2013) though not consistently (Voyer & Jansen, 2016). Similarly, studies found that including gender stereotyped item content created bias (Rahe & Jansen, 2022; Rahe et al., 2020), though again not consistently (Walsh et al., 1999). There has also been some support that using realistic stimuli reduces gender bias (Fisher et al., 2018), though providing social context does not seem to reduce ethnic bias (DeShon et al., 1998).

Item bias research is roughly evenly split between gender (13 papers) and ethnicity (10). MMAT scores were on the lower end, suggesting methodological limitations in this area, however high quality papers with large samples typically found that whilst DIF exists on test items, it does not meaningfully contribute to group mean score differences.

Behaviour in Test

Many of the 14 papers in this factor examined differences in test strategy. Considering gender, Stenlund et al. (2017) found that women and men employ different test-taking strategies. Several papers have found women use different, often piecemeal, strategies in spatial tasks (Cherney & Collaer, 2005; Heil & Jansen-Osmann, 2008; Picucci et al., 2011) and that by preventing more male-favoured holistic strategies, gender differences can be modestly reduced (Hirnstein et al., 2009).

The use of ineffective test strategies and test preparation have been found to mediate the relationship between ethnicity and performance (Ellis & Ryan, 2003), as has regulation of cognition, defined as the way people plan, implement strategies, self-evaluate and error check (Nguyen et al., 2003). Interestingly, a paper investigating stereotype threat found that under threat, people are more likely to persevere with an ineffective problem-solving strategy (Carr & Steele, 2009) perhaps suggesting the relationship between strategy use and performance is best understood through other factors such as stereotype threat.

Beyond strategy use, there has been some suggestion of group differences in effort. Campbell et al. (2018) found women showed greater pupil dilation in a spatial task which the authors suggested was a proxy for cognitive effort although this link is perhaps less clear. Harrison et al. (2006) found no differences in effort between SES or ethnic groups.

In a study exploring a range of testing behaviours, large differences were found between Black and White people in high-stakes testing, even when ability was controlled for (Frisby & Osterlind, 2006). Separately, in a study in the USA and Austria, differences in test behaviour were related to motivational differences in completing the experimental task (Tanzer et al., 1995), which has interesting implications for generalising findings from experimental studies to high-stakes testing.

Studies related to test behaviour have been conducted with gender (9), ethnicity (5) and SES (1). Sample sizes were on the lower end, as were average MMAT scores and 86% of samples were primarily students, suggesting this body of work is methodologically weaker than others. There were no compelling accounts of variations in test behaviour that contribute to group differences, beyond those that may be explained though other accounts such as stereotype threat.

Anxiety

Of 16 papers relating to anxiety, most were based on gender. Trait and state anxiety are related to performance (Kumari & Corr, 1998) and despite gender differences on general CATs not typically being found, women report being more anxious in high-stakes testing (Reeve et al., 2008; Stenlund et al., 2017). Additionally, Gabriel et al. (2011) found stress impacted spatial performance differently for women and men and Sokolowski et al. (2019) found women display more maths anxiety which is related to performance. Regarding ethnicity, Thames et al. (2015) found Black test takers experienced greater anxiety related to negative performance evaluation than White test takers, which in turn was related with lower performance, suggesting differences in anxiety may contribute to group score differences on CATs.

It is important to consider state anxiety as it relates to mechanisms for stereotype threat. Many authors specifically address this, and these findings will be reported in the stereotype threat section below (e.g., Harrison et al., 2006; Nguyen et al., 2003; Schmader et al., 2009).

Whilst most of this research looks at gender, ethnicity and SES have also been studied. MMAT scores were amongst the lowest and 94% of research involved student samples, though research typically benefitted from large sample sizes. Despite women typically scoring higher in trait anxiety, they perform similarly to men on overall CATs. State anxiety does appear to be related to test performance, however the mechanism seems to be better explained through stereotype threat.

Attitudes

In all, 23 papers examined attitudinal causes for group differences, though support for theories was often mixed and ambiguous. Self-confidence has received support in mediating the relationship between gender and performance on spatial tests (Arrighi

& Hausmann, 2022; Cooke-Simpson & Voyer, 2007; Picucci et al., 2011). Maertz et al. (2005) found self-efficacy differences between genders but not ethnicities. In perhaps the most compelling paper in self-confidence, Estes and Felker (2012) found women were less confident in a spatial task, that confidence was related to performance and manipulating confidence increased performance.

Studies have shown that Black people have lower motivation and face validity perceptions in testing (Arvey et al., 1990), though there is support that these differences both do (Chan et al., 1997) and do not (Chan, 1997) mediate relationship between ethnicity and performance. Ployhart and Ehrhart (2002) found that reducing ethnic differences in test motivation could reduce performance differences by 5–30%. Stenlund et al. (2017) did not find gender differences in motivation. There is some support that face valid tests reduce group differences in performance (Grand et al., 2010).

Several additional findings are notable. Moè et al. (2009) found that a belief that ability can be improved was related to higher spatial performance in women. Black people's beliefs that CATs are measures of knowledge rather than ability was related to lower test scores (Palumbo & Steele-Johnson, 2014). Finally, Forbes and Schmader (2010) found that training a more positive attitude towards the testing domain increased motivation but not performance.

Studies relating to attitudes were roughly split between gender (11) and ethnicity (12). Papers scored the highest on the MMAT on average and sample sizes were typically large. However, 91% employed student samples and 74% of studies were conducted in the US, which may limit generalisability to working populations in other regions. Within these papers, differences in self-confidence is modestly supported as a potential contributing factor to group score differences in CATs, although the number of studies in this area is very low and few studies have been conducted on high-stakes testing.

Experience

Eighteen papers met inclusion criteria relating to experience of the testing situation or domain. Many of these papers examined gender differences in spatial ability, which has received a lot of attention due to female underrepresentation in STEM disciplines which often involve spatial skills. There were many compelling findings that suggest experience contributes significantly to gender and ethnic differences in performance. Several studies found training, practice tests and practice items improved performance more for women than men, and even eliminate the gender gap on spatial performance (e.g., Cherney et al., 2003; Kass et al., 1998; Saccuzzo et al., 1996). These improvements have also led to a reduction in differences in brain activation between men and women (Jaušovec & Jaušovec, 2012). However, findings are not ubiquitous (McGee, 1978) and there is some suggestion that female gains do not transfer to spatial tasks outside of the trained task (Vasta et al., 1996). Interestingly, multiple studies found playing video games reduced the gender gap in spatial performance with women improving more than men (e.g., Cherney, 2008; Feng et al., 2007; Terlecki et al., 2008). However, practicing wrestling improved male and female performance equally (Moreau et al., 2012).

Outside of spatial abilities, Cherney and Collaer (2005) found prior maths experience explained a significant but very modest amount of gender differences in maths performance. With ethnicity, multiple studies found that practice and training improved Black performance more than White performance (Skuy et al., 2002; Campion et al., 2019).

Most papers in this domain looked at gender and spatial performance, specifically to explore gender differences in STEM domains. Despite this, findings are similar exploring other cognitive domains, and exploring ethnicity, though there is far less research in this area. Taken as a whole, these papers suggest differences in experience do seem to be a contributing factor to group mean score differences. Papers scoring higher on the MMAT were more likely to find support for experiential difference, suggesting methodological limitations may have been a factor in some papers not finding similar support. This being said, few papers were conducted on high-stakes testing and sample sizes in this area were on the lower end.

Stereotype Threat

This factor has attracted the most research attention with 75 papers meeting inclusion criteria, mostly with robust findings that threat impacts performance by gender (e.g., Çetinkaya et al., 2020; Hausmann, 2014; Quinn & Spencer, 2001), ethnicity (e.g., Scherbaum et al., 2011; Shapiro et al., 2013; Steele & Aronson, 1995) and SES (e.g., Harrison et al., 2006; Konan et al., 2011; Spencer & Castano, 2007). The impact of threat was worse for individuals with multiple threatened identities (e.g., Gonzales et al., 2002; Tine & Gotlieb, 2013). Most studies rely on priming threat, but effects occur even when not primed (Smith & White, 2002). Stereotype threat has been considered a source of measurement bias (Wicherts et al., 2005). However, some papers have struggled to replicate experimental findings in real world settings (e.g., Gillespie et al., 2010).

Many individual differences are moderators and mediators of the relationship between threat and performance such as identifying with the threatened domain, identifying with the stereotyped group and test difficulty (Ployhart et al., 2003). Research has also found similar links with the strength of implicit stereotypes held (Kiefer & Sekaquaptewa, 2007), self-confidence (Sanchis-Segura et al., 2018), testosterone levels (Josephs et al., 2003) and locus of control (Cadinu et al., 2006).

Several mechanisms by which threat impacts performance have been proposed by researchers, with the two most supported being the depletion of working memory and executive resources (e.g., Beilock et al., 2007; Johns et al., 2008; Schmader & Johns, 2003) and the interpretation of increased anxious arousal (e.g., O'Brien & Crandall, 2003; Penner & Willer, 2011; Schmader et al., 2009) with the two accounts not being mutually incompatible.

Studies have supported several interventions aimed at overcoming the effects of threat, the simplest of which being stating the test does not show group differences in performance (e.g., Campbell & Collaer, 2009; Spencer et al., 1999). The presence of a competent role model from the threatened identity has also been shown to disrupt threat (e.g., Marx & Goff, 2005; Marx & Roman, 2002; McIntyre et al., 2003), as has priming an alternative, competent stereotype (Ortner & Sieverding, 2008) or

priming perceptions of a more powerful status (Harada et al., 2013). Finally, studies showed approaches such as teaching about threat (Johns et al., 2005), prompting participants to reappraise mind wandering as normal (Schuster et al., 2015), mindfulness (Weger et al., 2012) and self-affirmation (Martens et al., 2006) disrupted established mechanisms for stereotype threat. These studies also provide support for the impact of threat.

Overall, there is a large body of research conducted in this area with 51 papers looking at gender, 27 at ethnicity and 7 at SES. MMAT scores are on the higher end and sample sizes are mixed with a range of smaller and larger samples. However, very few studies were conducted on high-stakes testing with papers such as Gillespie et al. (2010) struggling to replicated stereotype threat effects found in laboratory studies with student samples in real life occupational settings. This appears to be the greatest limitation of this research. Many studies demonstrate that stereotype threat can influence group differences, but few show that it actually does in a recruitment context.

Multiple Factor Research

Only 44 of 225 papers addressed multiple factors. Table 3 shows the number of papers that addressed each combination of factors. Generally, these papers did not

Table 3 Counts of the Number of Papers that Address Multiple Factors

Number of Papers	Factors Included
7	Biological, Stereotype Threat
5	Attitudes, Stereotype Threat
4	Biological, Environmental
3	Item Bias, Stereotype Threat
3	Behaviour in Test, Attitudes
2	Anxiety, Attitudes, Stereotype Threat
2	Environmental, Item Bias
2	Latent Trait, Item Bias
2	Behaviour in Test, Stereotype Threat
1	Behaviour in Test, Anxiety, Stereotype Threat
1	Behaviour in Test, Anxiety, Attitudes, Stereotype Threat
1	Item Bias, Behaviour in Test, Attitudes
1	Behaviour in Test, Anxiety, Attitudes
1	Biological, Anxiety
1	Biological, Experience
1	Environmental, Latent Trait
1	Environmental, Stereotype Threat
1	Latent Trait, Stereotype Threat
1	Item Bias, Behaviour in Test
1	Behaviour in Test, Experience
1	Behaviour in Test, Stereotype Threat
1	Anxiety, Attitudes
1	Attitudes, Experience

provide strong evidence for consistent interactions between factors. A majority of these papers (24) addressed stereotype threat, with seven investigating how biological factors influenced the mechanisms for the impact of threat. For example, Josephs et al. (2003) demonstrated testosterone moderated the effects of threat. None of these papers suggested that accounts described in the biology section above could be explained by stereotype threat. Five papers examined stereotype threat and attitude, for example exploring whether perceptions that tests are valid disrupts stereotype threat effects (Hollis-Sawyer & Sawyer, 2008), although findings across these papers were not strong enough to establish a relationship between stereotype threat and test attitudes. Other than stereotype threat potentially being an explanation of group differences in both test anxiety and test behaviour as described above, other papers did not produce a strong case for a consistent relationship between stereotype threat and factors.

Outside of stereotype threat, four papers looked at the interplay between biological and environmental factors, for example Johnson and Bouchard (2007) finding support that both factors predict CAT scores. Other combinations of factors accounted for an additional 16 papers, although again strong support for the interactions between factors was not supported.

Taken together, whilst these 44 papers offer interesting insights, they do not go far enough to provide reliable insights as to how varying accounts interact with one another. Despite the lack of evidence, it is likely that further research could support interactions between factors.

Discussion

This review aimed to explore all factors related to mean score group differences in CATs, in order to create a better understanding of these differences whilst identifying any potential avenues for mitigating them in employee selection through changes to test design and use. It is clear there is a huge volume of research that has been conducted to this end, and the field is highly disparate. Many of the factors explored have a body of compelling evidence to support their contribution, both those related to substantive accounts and methodological/contextual accounts of group differences.

Considering biological factors, the role of sex hormones and monthly hormonal fluctuations in gender differences in stereotypically gendered cognitive domains is clear. What is less clear is the mechanisms by which they influence performance. These may be biological, but it is also possible that sex hormones influence gender identification, which in turn leads to increases in stereotype threat in specific domains. However, a threat-related account becomes less relevant when considering developmental hormone levels on subsequent cognitive performance. Brain activation differences are compelling, but again may be related to non-biological factors (e.g., Jaušovec & Jaušovec, 2012). Finally, genetic evidence is currently lacking though promising, and so the area warrants further exploration. Taken together, whilst the current review finds little to suggest ethnic or SES differences have specified biological causes, there are irrefutable gender implications for the use of CATs in selection

contexts. Recruiting using specific domains that are not related to job performance may inadvertently introduce adverse impact.

The impact of environmental factors on group mean score differences is unambiguous. Parental and individual education, income, immigration status, and language proficiency have been found to relate to ethnic and SES differences, whilst gendered play and contraceptive use and type has been found to be related to gender differences. These are very important considerations for occupational testing and, as with the biological differences above, suggest that there is clear evidence that changes to test design and implementation will not eliminate group mean score differences in CATs, only potentially reduce them.

Latent trait and measurement invariance findings are inconsistent and there is sizable disagreement between studies. Given the lack of consensus, there is a need for more research and more theoretical developments to help build a greater understanding. In terms of the current paper, this factor offers little in terms of implications for the use of CATs in selection.

Some of the criterion validation studies reported provide a sound account that CATs do not underpredict performance of women or minorities and may overpredict it. They are also highly relevant in an occupational context. This suggests that group differences are not based on test biases. However, there is little to explain potential overprediction, and little to address potential bias in criteria used. These studies are clearly very important, but even when looking at objective criteria, the workplace is not experienced in the same way by different groups of people. As such, it is not possible to infer whether these criterion validation studies effectively support measurement invariance, especially when results are often inconsistent.

Many CATs have been found to show DIF, and test publishers should be mindful of this and endeavour to remove biased items from their tests. This being said, from the evidence explored here, item bias is highly unlikely to contribute in any meaningful way to group mean score differences in occupational testing.

Differences in test strategy seems to be well supported with groups performing worse on specific domains often employing less effective strategies. In a study looking not at strategy but test experience, Cherney et al. (2003) found exposure to test items or providing detailed instructions reduced the gender gap in spatial performance. It is possible that effective test strategies are developed through experience in the given domain, and that greater exposure to test content and more elaborated instructions in selection testing could reduce group differences without sacrificing validity. Other differences in test behaviour such as differences in effort did not appear especially relevant.

Anxiety differences are likely better explained in terms of stereotype threat; however, it is notable that women typically display higher trait anxiety and report greater test anxiety. Meaningful gender differences are seldom found on *g*, but where they are, there is a possible that anxiety is playing a part. Efforts to reduce experience of test anxiety in an occupational context could be worth exploring to protect against gender differences in CATs, though evidence is less than convincing.

Self-confidence was found to be an important mediator between gender and spatial performance and is the strongest attitudinal account that may related to group differences on CATs. Given the success of manipulating confidence (Estes & Felker,

2012), this is an area that warrants further exploration in occupational testing because currently there are only a few studies supporting this. Regarding motivational differences, support was less clear cut. Research regarding stereotype threat has found that individuals under threat are more likely to withdraw or put in less effort as a means of self-preservation (e.g., Spencer et al., 2016), and so this could explain why motivational differences were found in some studies.

Experiential difference findings were almost exclusively related to gender differences on spatial tests from papers exploring why women are underrepresented in STEM. There is great potential here looking at other domains. By extension, it would follow that men might perform worse than women in verbal tasks if they spend less time as children in verbal play. There are also likely important differences in activities performed in child play and adult leisure time by different ethnic and SES groups, which could contribute to domain experience and group differences. Building a greater understanding of the mechanisms by which past experience gives certain groups advantages in specific domains, it is possible that interventions aimed at mitigating those advantages may reduce group differences in occupational testing.

Finally, stereotype threat has a large body of evidence to support that it impacts performance for women and men, lower SES people and ethnic minorities across many cognitive domains. As well as evidence for the mechanisms by which threat impacts performance, a lot of research has focused on interventions aimed at mitigating threat. However, there have been challenges in replicating studies in high stakes testing, which may suggest the nature of the experimental approach could be contributing to the magnitude of effects seen.

Implications

There are significant theoretical and practical implications of this review. Perhaps most significantly, finding that stereotype threat, test experience and self-confidence moderate the relationship between certain group identities and scores on CATs has implications for the understanding of group differences and the ways we measure ability. Whilst this review found biological and environmental factors to be important, these artifacts of the testing process need to be considered when exploring group differences.

From a theoretical perspective, limitations in this body of research taken as a whole have implications for our understanding of CATs, and specifically for this review, our ability to understand the causes of mean score group difference on occupational CAT use. From the reviewed research, it is clear that occupational testing has received much less research attention than educational testing, presumably as private occupational test publishers are less motivated than public education departments to explore differences. There is also likely more funding in specific areas, such as in promoting women in STEM fields (Kahn & Ginther, 2017). Whilst much can be learned for this body of research, it does raise questions about how complete our understanding of occupational testing is, and what gaps exist in our knowledge. A further comment, which is true for much psychology literature (e.g., Hanel & Vione, 2016), is that most of this research depends on student samples, most commonly psychology student samples. Concerns around demand characteristics aside, gener-

alising findings from student samples who have little or no employment experience to occupational testing is problematic. Indeed, the overuse of student samples is concerning as student and nonstudent samples typically produce differing results (Shen et al., 2011). University students are typically young, educated and more likely to be from higher SES backgrounds. As such, they are not representative of the wider working population and so may bias conclusions. Group differences in CAT scores may not be the same as those seen in occupational samples and the extent to which different accounts explain those differences may vary. As an example, Shewach et al. (2019) stated that stereotype threat effects are typically more pronounced in individuals who are domain-identified, that is, performance on the domain is important to them. Students have pursued education and so are potentially more likely to be identified with cognitive domains, therefore stereotype threat research conducted on students may find the effects are larger than on working population samples. Shewach et al. (2019) also commented that student samples typically used in this research have vastly different motivation for completing tests as part of research, often for course credit where performance is irrelevant, when compared with people taking CATs as part of a recruitment process who are motivated by the potential for a new role where performance is important. Indeed only 14% of studies reviewed met the criteria for ecological validity. Recruitment tests are high stakes, applicants are aware they are being evaluated, and they are aware of the competitive context. The majority of studies reviewed were not conducted in these settings. There are likely to be significant differences in terms of factors such as motivation and anxiety when taking a test for an experiment compared with taking a test for a job (e.g., Barry et al., 2010) and text anxiety is found to negatively impact test scores in occupational testing (McCarthy & Goffin, 2005). Spencer et al. (2016) describe anxiety as a supported mechanism for stereotype threat and so it logically follows that stereotype threat may impact test performance differently on high-stakes testing when compared with laboratory studies.

A further critique of this literature comes from the fact that 56% of studies were conducted in the US, which may limit cross-cultural generalisability. No notable trends were observed when comparing US studies with non-US studies, however only 8% of studies involved at least some participants outside of North America and Europe with only 4% using participants exclusively outside of those regions. There may be significant cross-cultural difference in group differences that the research currently does not account for. Gaps in the literature, overuse of student samples, a lack of research on high stakes testing and an overreliance on Western samples all limit our ability to understand the causes of group differences in a recruitment context. Whilst conclusions drawn from this review suggest multiple potential factors that may be introducing bias into occupational CAT use, more research is needed on high-stakes recruitment testing in order to confirm that these effects generalise to this context.

Additionally, a majority of the papers reviewed looked at factors in isolation. Those papers that did consider multiple factors rarely aimed to consider how factors interact to explain group differences. This makes it incredibly difficult to tease apart potentially confounding accounts. For example, anxious arousal has been found to be a mechanism for stereotype threat (Ben-Zeev et al., 2005) and yet no papers identified in this review investigating test anxiety examined stereotype threat. Threat also

was not considered in the one paper finding that manipulating self-confidence influenced test scores (Estes & Felker, 2012). The interactions between stereotype threat, test anxiety and self-confidence are perhaps easier to hypothesise, however there are likely relationships beyond these that existing literature does not fully account for. Thomsen et al. (2000) found support that some brain activation differences between women and men when taking tests was actually a result in different test strategy adoption, which could suggest some biological accounts for group differences are actually behavioural ones. Without significantly more research in this area, it is not possible to reliably construct a conceptual model to represent the interactions between factors, which would be valuable in understanding how to potentially mitigate group differences. Future research should consider exploring multiple factors in the same studies to work towards a greater understanding of the interplay between factors.

Originally, this review intended to explore age differences in CAT scores. However, no papers were found that sought to explain age difference in CAT scores. There were multiple papers found in the search process that looked specifically at cognitive decline in retirement ages, but these did not meet inclusion and exclusion criteria as they were not based on working age participants. This is an important gap in the literature that future research should look to investigate. There are likely biological and environmental difference throughout people's lives that impact their performance on CATs, and different age groups might also be susceptible to factors such as stereotype threat and item bias.

From a practical perceptive, given stereotype threat, domain experience and self-confidence are found to have a moderating effect on CAT scores, researchers exploring the measurement of intelligence, and not only those investigating group differences, need to control for these factors. Most significantly though, occupational test developers and test users also need to be aware of these relationships, and explore ways to mitigate them as much as is possible. Fostering workplace diversity has become a key objective for organisation (Ones et al., 2017), with multiple studies highlighting the commercial advantages it brings to them (e.g., Gomez & Bernet, 2019; Nathan & Lee, 2013; Wee et al., 2014). Whist CATs are highly predictive of work success, their use will result in differing selection rates for groups that exhibit mean score differences. This has been termed the diversity-validity dilemma (Ployhart & Holtz, 2008). In outlining specific biases present in the occupational use of CAT, this review brings focus to the most promising areas to address this dilemma, though developing efforts to mitigate biases. This is especially important when considering that organisations lack highly valid selection methods that don't show group differences (Sackett et al., 2022).

Limitations

This review is not without its limitations. In order to focus on studies relevant for occupational testing, we did not include studies primarily involving child samples. Many child studies, often regarding environmental or biological factors related to group differences, were rejected for this review due to concerns of generalisation of findings. However, it is possible that something could be gained from exploring these

studies in future reviews. For example, there is a body of evidence exploring a potential genetic influence of ethnic differences on CAT scores that did not meet inclusion criteria (e.g., Osborne, 1980), as well as literature highlighting methodological concerns limiting conclusions drawn from such research (e.g., Suzuki & Valencia, 1997).

This review also excluded studies that were solely based on performance on high-stakes standardised education testing, due to assumed confounding factors such as differences in parental and school support. There are huge volumes of research here, largely based in the US. Whilst there are insights to gain, including these papers would have meant they disproportionately influenced results, and findings may not have been generalisable. However, this decision may have inadvertently excluded insightful research investigating group difference that was not picked up elsewhere in this review, especially considering much of the research will have been conducted on high-stakes, something lacking in much of the research reviewed. It is also worth addressing how broad the scope for this review was and the significant heterogeneity of the literature reviewed. This was needed to address the research question, especially in considering findings from each factor in relation to others, but this approach sacrificed the specificity of a narrower focus. Important concepts may have not received as much focus as they potentially warranted in a narrower review. Finally, in selecting a search strategy including the terms "gender" and "sex", "ethnicity" and "race" as well as including papers from anywhere in the world, important nuances and distinctions in group differences may have been overlooked.

Additionally, there were several methodological decisions that were made due to practical constraints when conducting a systematic review of this size. The application of the inclusion and exclusion criteria, as well as data extraction and synthesis were all completed by one person. Whilst the other authors reviewed this work, this could potentially have introduced bias. Also, forward citation searching was not as expansive as it could have been, which could have led to insightful research not being identified.

Conclusion

Whilst substantive explanations, such as biological and environmental factors, clearly contribute to group mean score differences on CATs, this systematic review outlines several potential methodological and contextual avenues for further exploration to change test design and administration in efforts to make widespread occupational testing fairer, namely mitigating stereotype threat, mitigating advantages gained by domain experience and exposure, and reducing differences in self-confidence. There is some support that exploring differences in state anxiety may be insightful, and whilst only impacting group differences negligibly, exploring DIF in CATs is still something that should be recommended. However, much of the evidence has not been conducted in high-stakes occupational testing and is overly reliant on US student samples. Additional, there is a paucity of research exploring the interaction of these explanatory factors. Future research should focus on understanding these factors together in an occupational context.

Appendix A

Table 4 Table Containing Data Extracted from Each Paper Included in the Review

Paper	Country	Sample	Groups	Factor	Impact	Relevant Results	MMAT Score	Ecological Validity
Abad et al. (2004)	USA	1820 students	Gender	Item Bias	Yes	Some items were found to have gender bias, although this did not fully account for groups score differences	80%	Yes
Aleman et al. (2004)	Netherlands	26 women	Gender	Biological	Yes	Administering testosterone was related to an increase in spatial performance in small sample of women	40%	No
Alexander and Evardone (2008)	USA	228 students	Gender	Item Bias	Yes	Gender differences on mental rotation tasks were halved when items featured humans rather than abstract shapes	60%	No
Alvarez-Vargas et al. (2020)	USA	517 students	Gender	Anxiety	Yes	Spatial anxiety partially mediated relationship between gender and spatial performance	40%	No
Arendasy and Sommer (2012)	Austria	1780 adults	Gender	Item Bias	Maybe	Aspects of item design of matrix-style questions can lead to DIF between genders	40%	No
Aronson et al. (2002)	USA	79 students	Ethnicity	Stereotype Threat	Yes	Encouraging students to see intelligence as malleable improved GPA arguably by removing impact of threat	40%	Yes
Arrighi and Hausmann (2022)	UK	269 students	Gender	Anxiety Attitudes	Yes Yes	Some support that self-confidence and spatial anxiety mediate relationship between gender and spatial performance, though sample too small	40%	No
Arvey et al. (1990)	USA	263 applicants	Ethnicity	Attitudes	Yes	Black applicants had lower motivation, lower face validity and higher preparation for CATs; gender and age both related to anxiety and test ease; these factors were all related to test performance	60%	Yes
Ashcraft and Faust (1994)	USA	80 students	Gender	Anxiety	Yes	Small study showing women experienced greater maths anxiety, and anxiety was linked to performance in a non-linear way	20%	No

Table 4 (continued)

Paper	Country	Sample	Groups	Factor	Impact	Relevant Results	MMAT Score	Ecological Validity
Bauer et al. (2021)	Germany	107 students	Gender	Stereotype Threat	No	Investigated potential for mindfulness to overcome threat, however manipulation was unsuccessful, and analysis mostly grouped genders	40%	No
Beilock et al. (2007)	USA	190 participants	Gender	Stereotype Threat	Yes	Six studies demonstrating how threat depleted working memory resources and can spill over beyond threatened task	20%	No
Bell et al. (2006)	Canada	33 students	Gender	Biological	Maybe	Men and women differed in brain activation on several cognitive domains but not consistently ones with performance differences	40%	No
Ben-Zeev et al. (2005)	USA	39 female students	Gender	Stereotype Threat	Yes	Women under threat performed better on an easy unrelated task and worse on a difficult one; threat effects reduced when women misdirected their arousal; supports arousal account of threat	40%	No
Berry et al. (2014)	USA	1,000,000+people	Ethnicity	Criterion Validity	Maybe	Criterion to test SD ratios account for a modest proportion of differential validity, though suggested underprediction of minorities unlikely	40%	Yes
Brown and Day (2006)	USA	136 students	Ethnicity	Stereotype Threat	Yes	Using high threat, low threat and standard test instructions, demonstrated that threat has an impact even on a culture-free test using standard instructions	60%	No
Brown and Josephs (1999)	USA	269 students	Ethnicity	Stereotype Threat	Yes	Women and men performed worse when they believe a test would be diagnostic of weak or exceptional ability respectively; an external handicap mitigated performance concerns	40%	No
Burkitt et al. (2007)	Canada	75 students	Gender	Biological	Yes	Men outperformed women, as did higher testosterone people from both genders, on a mental rotation task	60%	No
Burton et al. (2005)	USA	134 students	Gender	Biological	Yes	Less gender-typical digit ratios were associated with better verbal and spatial performance	20%	No
Butler et al. (2006)	USA	25 people	Gender	Biological	Yes	Brain activation gender differences were found when completing a spatial task	40%	No

Table 4 (continued)

Paper	Country	Sample	Groups	Factor	Impact	Relevant Results	MMAT Score	Ecological Validity
Cadinu et al. (2006)	Italy	60 female students	Gender	Stereotype Threat	Yes	Despite typically performing better, women with an internal locus of control were more susceptible to threat	40%	No
Campbell and Collaer (2009)	USA	194 students	Ethnicity	Stereotype Threat	Yes	Stating that a test did not show gender differences reduced gender differences	60%	No
Campbell et al. (2018)	Ireland	99 students	Gender	Behaviour	Yes	In spatial task, women's pupils were more dilated suggesting more effort though causes unclear. Speed and sensitivity were similar	60%	No
Campion et al.(2019)	USA	25,548 adults	Ethnicity	Experience	Yes	Practice tests increased performance on actual tests and increase was higher for ethnic minorities. Those with higher scores were more likely to go on to take actual tests	60%	Yes
Carr and Steele (2009)	USA	129 students	Gender	Stereotype Threat	Yes	Priming threat led to an increase in persevering with an ineffective problem-solving strategy	60%	No
Carretta and Doub (1998)	USA	41,976 military recruits	Gender, Ethnicity	Criterion Validity	Yes	Test performance was less predictive of training performance for Black recruits	20%	Yes
Çetinkaya et al. (2020)	Turkey	156 female students	Gender	Stereotype Threat	Yes	Threat impacted performance and was moderated by self vs group threat, gender identification and type of affirmation strategy used	40%	No
Chan (1997)	USA	241 students	Ethnicity	Attitudes	Maybe	Ethnicity, validity perceptions and performance correlated, though author not able to show a mediating relationship	60%	No
Chan et al. (1997)	USA	180 students	Ethnicity	Attitudes	Yes	Black/White score differences were partially explained by face validity and motivation, though differences in face validity and motivation may be explained by prior performance	60%	No

Table 4 (continued)

Paper	Country	Sample	Groups	Factor	Impact	Relevant Results	MMAT Score	Ecological Validity
Chan et al. (1998)	USA	197 students	Ethnicity	Attitudes	No	No racial differences found between relationships between belief in tests, pretest reactions and test scores	60%	No
Cherney and Collaer (2005)	USA	198 students	Gender	Behaviour Experience	Maybe Maybe	Prior maths experience or test strategy did not fully account for gender differences in spatial task	40%	No
Cherney (2008)	USA	61 students	Gender	Experience	Yes	Playing computer games reduced gender gap in spatial test	60%	No
Cherney et al. (2003)	USA	113 students	Gender	Experience	Yes	Gender gap in spatial task reduced by either exposure or detailed instructions	40%	No
Cherney et al. (2014)	USA	60 students	Gender	Experience	Yes	Brief use of video games improved spatial performance more so in women than men	0%	No
Clark et al. (2011)	USA	236 students	Region	Stereotype Threat	Yes	Replication of previous findings that performance is lower when threat is activated in people from Southern USA	40%	No
Cockcroft et al. (2015)	South Africa; UK	456 students	Ethnicity	Latent Trait/ Measurement invariance	Yes	No measurement invariance found between Black South African and White British groups	40%	No
Collaer et al. (2007)	International	255,116 people	Gender	Biological	Yes	Prenatal sex hormones, inferred from digit ratios, predicted spatial performance	80%	No
Cooke-Simpson and Voyer (2007)	Canada	80 students	Gender	Attitudes	Maybe	Men were more confident and outperformed women, but unclear if gender differences in confidence contributed to score differences	40%	No
Courvoisier et al. (2013)	Switzerland	17 people	Gender	Biological	Yes	Monthly hormonal changes influenced mental rotation performance in both genders and was more pronounced in women	40%	No
Croizet and Claire (1998)	France	128 students	SES	Stereotype Threat	Yes	Stating a test was diagnostic of ability was related to performance only in low SES participants	40%	No

Table 4 (continued)

Paper	Country	Sample	Groups	Factor	Impact	Relevant Results	MMAT Score	Ecological Validity
De Meijer et al. (2008)	Netherlands	3,047 police trainees	Ethnicity	Criterion Validity	Yes	Modest differential validity found by ethnicity on predicting training performance	80%	Yes
Dennehy et al. (2014)	USA	83 students	SES	Stereotype Threat	Yes	Priming low SES people with an implemental mindset increased test performance	60%	No
DeShon et al. (1998)	USA	558 students	Ethnicity	Item Bias	No	Found no support that test items including social context reduced ethnic differences	0%	No
Díaz et al. (2012)	Spain, Morocco	460 students and staff	SES	Environmental	Yes	Education and income found to be related to scores in Moroccan but not Spanish sample, suggesting some environmental cause	40%	No
Dolan et al. (2004)	Netherlands	1,322 blue-collar workers	Ethnicity	Latent Trait/ Measurement invariance	Yes	MGCFA of sample from previous study demonstrating Spearman's hypothesis unable to find factorial invariance undermining initial conclusions	60%	Yes
Dolan et al. (2006)	Spain	588 people	Gender	Environmental Latent Trait/ Measurement invariance	Yes No	No gender differences found in g, though childhood education was found to predict g	20%	No
Domino and Morales (2000)	USA	250 students	Ethnicity	Item Bias	Maybe	Mexican- and Anglo-American students did not differ in scores on culture-free test items	40%	Yes
Doyle and Voyer (2018)	USA	108 students	Gender	Item Bias	Yes	Gender differences on mental rotation tasks were reduced when items featured humans rather than abstract shapes	60%	No
Ellis and Ryan (2003)	USA	170 students	Ethnicity	Behaviour Attitudes	Yes Maybe	Test preparation and use of ineffective test strategy both mediated relationship between ethnicity and performance; self-efficacy, test preparation and use of ineffective strategies all higher in Black participants	40%	No

Table 4 (continued)

Paper	Country	Sample	Groups	Factor	Impact	Relevant Results	MMAT Score	Ecological Validity
Epting and Overman (1998)	USA	47 students	Gender	Biological	No	Cycle time did not impact performance in women	40%	No
Estes and Felker (2012)	USA	524 students	Gender	Attitudes	Yes	Across four studies, women were less confident, confidence was related to performance and manipulating confidence increased performance in rotation task	60%	No
Fagan and Holland (2002)	USA	97, 157, 79, 36, 93 students	Ethnicity	Environmental Item Bias	Yes Yes	Five studies showing that prior knowledge accounts for Black/White differences in vocabulary tests	20%	No
Fagan and Holland (2007)	USA	77, 65, 86, 223, students	Ethnicity	Item Bias	Yes	Four studies in which they demonstrate prior knowledge explains test scores on certain cognitive domains, though items were selected that were expected to show bias	0%	No
Falter et al. (2006)	UK	46 students	Gender	Biological	Yes	2D:4D ratios supported the influence of prenatal testosterone on cognitive ability	40%	No
Fasfous et al. (2013)	Spain, Morocco	54 adults	Ethnicity	Latent Trait/ Measurement invariance	Yes	Neuropsychological predictors of CAT performance differed between cultures, suggesting tests were measuring different constructs	20%	No
Feng et al. (2007)	Canada	20 students	Gender	Experience	Yes	Playing an action game improved spatial performance and reduced gender gap	40%	No
Fisher et al. (2018)	Canada	470 students	Gender	Item Bias	Yes	More realistic stimuli significantly reduced gender gap in spatial task	40%	No
Forbes and Schmader (2010)	USA	498 students	Gender	Attitudes Stereotype Threat	No Yes	Retraining attitudes led to increased motivation but not performance, whereas countering threat led to increased working memory and performance in four studies with small samples	60%	No
Forbes et al. (2015)	USA	58 students	Ethnicity	Biological Stereotype Threat	Yes Yes	EEGs suggested a potential biological default mode network marker for vulnerability to threat	40%	No

Table 4 (continued)

Paper	Country	Sample	Groups	Factor	Impact	Relevant Results	MMAT Score	Ecological Validity
Frenken et al. (2016)	UK, Russia	659 students	Gender	Environmental	No	Women with brothers did not outperform women with sisters on a spatial test, though spatial play was related to spatial ability	40%	No
Frisby and Osterlind (2006)	USA	5769 people	Gender, Ethnicity	Behaviour	Yes	Test behaviour differed by ethnicity and gender, ethnicity differences remained when ability was controlled for	20%	Yes
Gabriel et al. (2011)	USA	156 students	Gender	Anxiety	Yes	Stress impacted spatial performance differently between men and women	40%	No
Gardner and Deadrick (2008)	USA	719 machine operators	Ethnicity	Criterion Validity	Yes	Ability test underpredicted Black employee performance	20%	Yes
Gardner and Deadrick (2012)	USA	626 machine operators	Ethnicity	Criterion Validity	Yes	Ethnicity moderated the criterion validity of ability	20%	Yes
Gillespie et al. (2010)	USA	229 applicants	Ethnicity	Stereotype Threat	No	In a real work setting, efforts to reduce threat in testing were not successful, with low sample sizes	40%	Yes
Gizewski et al. (2006)	Germany	24 people	Gender	Biological	Yes	Gender differences found in brain activation in spatial and verbal tasks	20%	No
Gonzales et al. (2002)	USA	115 students	Gender, Ethnicity	Stereotype Threat	Yes	Latino women experienced increased threat by belonging to two stereotyped groups	60%	No
Gordon and Lee (1986)	USA	62 students	Gender	Biological	Yes	Sex hormones correlated with performance on spatial and verbal tests	40%	No
Gordon and Lee (1993)	USA	82 female students	Gender	Biological	No	No support for performance differences at different points of menstrual cycle	40%	No
Gordon et al. (1986)	USA	56 students	Gender	Biological	Yes	Administering hormones (LHRH) was related to performance changes on specific cognitive domains	80%	No

Table 4 (continued)

Paper	Country	Sample	Groups	Factor	Impact	Relevant Results	MMAT Score	Ecological Validity
Gouchie and Kimura (1991)	Canada	88 students	Gender	Biological	Yes	Higher testosterone led to an increase in spatial/maths performance in women and decrease in men; no differences on tests that favour women	20%	No
Grand et al. (2011)	USA	345 students	Gender	Attitudes Stereotype Threat	Yes No	Some support that face valid tests can reduce group differences and modestly increase general performance	80%	No
Gresky et al. (2005)	USA	129 students	Gender	Stereotype Threat	Yes	Threatened women performed better after completing an exercise considering their multiple social identities	40%	No
Griksiene and Ruksenas (2011)	Lithuania	43 students	Gender	Biological Environmental	Yes Yes	Verbal and spatial performance differences found between women using different birth control suggest hormonal relationship	40%	No
Griksiene et al. (2018)	Lithuania	99 students	Gender	Environmental	Yes	Women using oral contraceptives performed worse in spatial task than other women and men, likely due to employing different strategies	60%	No
Griksiene et al. (2019)	Lithuania	63 adults	Gender	Biological	Yes	Hormonal and brain activation differences linked to gender differences in spatial performance	60%	No
Grubb and Ollendick (1986)	USA	80 students	Ethnicity	Latent Trait/ Measurement invariance	Maybe	Black/White test score differences on CATs not replicated on learning tasks potentially suggesting cultural bias	40%	No
Gur et al. (2000)	USA	27 people	Gender	Biological	Yes	Men and women showed differing activation in regions of the brain when taking gendered tests	60%	No
Halari et al. (2005)	UK	84 people	Gender	Biological	No	Hormone levels were not related to performance in tests with gender score differences	80%	No
Halari et al. (2006)	UK	19 people	Gender	Biological	Maybe	Small study was unable to replicate previous findings that brain activation varies by gender on cognitive tasks	40%	No

Table 4 (continued)

Paper	Country	Sample	Groups	Factor	Impact	Relevant Results	MMAT Score	Ecological Validity
Halpern and Tan (2001)	Turkey	158 + medical students	Gender	Biological Stereotype Threat	Maybe No	Some support of the influence of hormones and menstrual cycle on ability, though threat a potential confounder in some studies	20%	No
Hampson (1990a)	Canada	50 female students	Gender	Biological	Yes	Hormonal differences at different cycle stages were associated with differing performance levels in gendered tests	40%	No
Hampson (1990b)	Canada	45 female students	Gender	Biological	Yes	Hormonal differences at different cycle stages were associated with differing performance levels in gendered tests	40%	No
Hampson et al. (2022)	Canada	264 female students	Gender	Environmental	Yes	Contraceptive use and contraceptive type was related to spatial performance	40%	No
Harada et al. (2013)	USA	22 women	Gender	Stereotype Threat	Yes	Priming higher power reduced cognitive interference and improved maths performance in women	80%	No
Harrison et al. (2006)	USA	260 students	Ethnicity, SES	Behaviour Anxiety Stereotype Threat	No Yes Yes	Lower SES participants performed worse when threat was primed, were more anxious and less academically identified, though there were no differences in effort or self-esteem	60%	No
Harville (1996)	USA	478 military staff	Gender, Ethnicity	Criterion Validity	Maybe	No evidence of gender bias, no evidence of racial bias in predictive validity found on two of three comparisons	40%	Yes
Hausdorf and Robie (2018)	Canada	3,159 applicants	Ethnicity, SES	Environmental	Yes	Ethnic differences were significantly reduced after controlling for years of education, language proficiency and generational immigration status	80%	Yes
Hausmann (2014)	UK	184 students	Gender	Stereotype Threat	Yes	Supported findings that stereotype priming influences performance on gendered tests	40%	No
Hausmann et al. (2000)	Germany	Eight female students	Gender	Biological	Maybe	Modest support for hormonal influence on spatial ability in very small sample	20%	No

Table 4 (continued)

Paper	Country	Sample	Groups	Factor	Impact	Relevant Results	MMAT Score	Ecological Validity
Hausmann et al. (2009)	Germany	114 adults	Gender	Biological Stereotype Threat	Maybe Maybe	Some evidence that priming leads to stereotype threat. Potential relationship between priming and hormone levels	40%	No
Heil and Jansen-Osmann (2008)	Germany	72 adults	Gender	Behaviour	Yes	Gender differences on rotation task more pronounced for more complex problems suggesting differences in strategy	40%	No
Heil et al. (2011)	Germany	400 female students	Gender	Biological	Yes	Women with a male twin outperformed women with female twin on spatial test supporting organisational effects hypothesis	40%	No
Hirnstein et al. (2009)	UK	34 students	Gender	Behaviour	Maybe	Gender difference on spatial test modestly reduced in format that prohibited specific strategies employed more by men	20%	No
Hirnstein et al. (2015)	Germany	75 mostly students	Gender	Stereotype Threat	Maybe	Threat priming increased male and female performance on a verbal test in which women typically excel	20%	No
Hirnstein et al. (2014)	Germany	136 adults	Gender	Biological Stereotype Threat	No Maybe	Testosterone unrelated to performance on stereotyped tasks; some support for threat though findings inconsistent	40%	No
Hollis-Sawyer and Sawyer (2008)	USA	189 students	Ethnicity	Attitudes Stereotype Threat	Maybe Yes	Priming threat led to lower performance in ethnic minority groups, though face validity manipulation findings were ambiguous	60%	No
Hooven et al. (2004)	USA	28 male students	Gender	Biological	Maybe	Testosterone related to performance on some elements of spatial tasks but not the rotation itself	60%	No
Houston (1987)	USA	14,427 military recruits	Ethnicity	Criterion Validity	Maybe	Test performance was less predictive of training performance in Black recruits	20%	Yes

Table 4 (continued)

Paper	Country	Sample	Groups	Factor	Impact	Relevant Results	MMAT Score	Ecological Validity
Howard et al. (1992)	UK	65 people	Gender	Biological	Yes	EEGs revealed gender differences in brain CNV asymmetries when completing cognitive tests suggesting a biological cause	60%	No
Hugdahl et al. (2006)	Norway	11 people	Gender	Biological	Yes	fMRI study showed gender differences in brain activation during spatial task	40%	No
Ispas et al. (2010)	Romania	527 employees	Gender	Criterion Validity	No	No evidence of differential prediction by gender or age as interaction between ability and group insignificant	40%	Yes
Jansen and Lehmann (2013)	Germany	120 athletes or non-athletes	Gender	Item Bias	Yes	Gender differences on mental rotation tasks were halved in conditions using human figures rather than abstract shapes	40%	No
Jaušovec and Jaušovec (2012)	Slovenia	165 students	Gender	Biological Experience	Maybe Yes	Gender differences in brain activation found in spatial tests, however training improved women's performance and led to similar activation patterns as males, suggesting difference in experience	40%	No
Jensen and McGurk (1987)	USA	426 students	Ethnicity	Environmental Item Bias	Yes No	No evidence of item bias on cultural or non-cultural items, though relationship between SES and ability found	40%	No
Jensen (1977)	USA	1496 job applicants	Ethnicity	Item Bias	No	Several analyses of Wonderlic Personnel Test showed little evidence of item bias impacting scores	60%	Yes
Johns et al. (2005)	USA	117 students	Gender	Stereotype Threat	Yes	Teaching women about stereotype threat diminished its effects	60%	No
Johns et al. (2008)	USA	273 students	Gender	Stereotype Threat	Yes	Four studies demonstrating how threat limits executive resources as people attempt to control their increased anxiety	40%	No

Table 4 (continued)

Paper	Country	Sample	Groups	Factor	Impact	Relevant Results	MMAT Score	Ecological Validity
Johnson and Bouchard (2007)	North America, UK, Australia	436 twins and families	Gender	Biological Environmental Latent Trait/ Measurement invariance	Maybe Maybe Maybe	Results supported that *g* masks gender differences in residual abilities. Suggested levels of intelligence similar but men and women have different approaches to solving problems. Modest support that ability is predicted by both genetics and environment	60%	No
Jordan et al. (2002)	Germany	24 students	Gender	Biological	Maybe	Women and men performing equally well in a spatial test displayed activation in differing brain regions	60%	No
Josephs et al. (2003)	USA	202 students	Gender	Biological Stereotype Threat	Yes Yes	Suggested threat effects occur due to desire to maintain or enhance status and demonstrated testosterone moderated the effects of threat	80%	No
Kass et al. (1998)	USA	41 naval staff	Gender	Experience	Yes	Practice session with feedback eliminated the gender gap on spatial task performance	60%	No
Kempel et al. (2005)	Germany	40 students	Gender	Biological	Yes	Some support of relationship between hormone levels in development and numerical/spatial ability found by comparing performance with 2D:4D ratio	20%	No
Kiefer and Sekaquaptewa (2007)	USA	63 female students	Gender	Stereotype Threat	Yes	Some evidence that holding stronger implicit stereotypes and having greater gender identification led to reduced performance on high-stakes maths exams in women	40%	Yes
Kirnan et al. (2009)	USA	49,150 applicants; 434 students	Gender, Ethnicity	Stereotype Threat	Maybe	Several studies showing inconsistently that priming threat led to lower performance for all groups. Authors accept design weaknesses	0%	Partially
Klein et al. (2007)	Belgium	69 Africans in Belgium	Ethnicity	Stereotype Threat	Yes	Performance was lower for Africans living in Belgium when threat was primed in a staged selection context. Anxiety was not found to mediate	60%	No
Konan et al. (2011)	Switzerland	202 students	Ethnicity, SES	Stereotype Threat	Yes	Stereotype threat on women and low SES people was reduced when stereotypes about immigrants was primed	80%	No

Table 4 (continued)

Paper	Country	Sample	Groups	Factor	Impact	Relevant Results	MMAT Score	Ecological Validity
Koscik et al. (2009)	USA	76 adults	Gender	Biological	Yes	Several anatomical brain differences found between men and women related to spatial performance	60%	No
Kumari and Corr (1998)	India	128 female students	Gender	Biological Anxiety	Yes Yes	Trait anxiety, stress and menstrual phase all found to be related to performance	20%	No
Lefkowitz and Battista (1995)	USA	369 employees	Gender, Ethnicity	Criterion Validity	Yes	Supervisors gave higher performance ratings to same-ethnicity staff; when controlling for this, criterion validity of tests reduced	60%	Yes
Levy and Reid (1978)	USA	73 students	Gender	Biological	Maybe	Some evidence that women were less laterally differentiated than men	40%	No
Maertz et al. (2005)	USA	287 applicants	Gender, Ethnicity	Attitudes	Yes	Self-efficacy was related to test performance and differed by gender but not ethnicity. Impact of test on self-efficacy differed between genders and ethnicities	60%	Yes
Maloney et al. (2012)	Canada	367 students	Gender	Anxiety	Maybe	Suggestion gender differences in anxiety can be accounted for by ability but data presented don't support this	40%	No
Martens et al. (2006)	USA	164 students	Gender	Stereotype Threat	Yes	Self-affirmation reduced impact of threat on women's performance on maths and spatial tests	40%	No
Marx and Goff (2005)	USA	59 students	Ethnicity	Stereotype Threat	Yes	Marginally significant increase in test scores when threat reduced by presence of competent Black experimenter	60%	No
Marx and Roman (2002)	USA	43 students	Gender	Stereotype Threat	Yes	The present of a competent role model diminishes effect of threat	60%	No
Mayer and Hanges (2003)	USA	152 students	Ethnicity	Stereotype Threat	Yes	Distinguished between general and specific threat and found both related to performance; unable to evidence of mediation of anxiety, self-efficacy, interference, evaluation apprehension	20%	No

Table 4 (continued)

Paper	Country	Sample	Groups	Factor	Impact	Relevant Results	MMAT Score	Ecological Validity
McCormick and Teillon (2001)	USA	82 students	Gender	Biological	Yes	Cycle phase explained 15% variance in spatial performance, no impact of stress, mood or performance perceptions	40%	No
McFarland et al. (2003)	USA	241 students	Ethnicity	Attitudes Stereotype Threat	No No	Prior findings that threat influences test scores were not replicated	20%	No
McGee (1978)	USA	347 students	Gender	Experience	Maybe	Modest practice and training led to increase in spatial ability scores for both genders	20%	No
McGlone and Aronson (2006)	USA	90 students	Gender	Stereotype Threat	Yes	Women performed worse and men better when gender identity was primed rather than private student identity	0%	No
McIntyre et al. (2003)	USA	268 students	Gender	Stereotype Threat	Yes	Modest support that presenting competent role models mitigated the impact of threat	80%	No
McKay et al. (2003)	USA	87 students	Ethnicity, SES	Environmental Stereotype Threat	Yes Yes	Parental education and influence of threat partially mediated relationship between ethnicity and ability	80%	No
McKeever et al. (1987)	USA	83 students	Gender	Biological	No	Testosterone was not found to be linked to spatial or verbal ability in men or women	40%	No
Moè (2016)	Italy	205 students	Gender	Attitudes Experience	Yes Yes	Training and practice improved mental rotation scores	40%	No
Moè (2018)	Italy	144 students	Gender	Stereotype Threat	Maybe	Large practice effect meant evaluating results was difficult	40%	No
Moè et al. (2009)	Italy	120 female students	Gender	Behaviour Attitudes	Yes Yes	Belief that ability can be improved and adopting spatial strategies were related to higher performance in spatial task in women	40%	No
Moè et al. (2018)	Italy, Germany	176 students	Gender	Environmental	Yes	Preference for spatial toys as a child related to spatial performance	40%	No

Table 4 (continued)

Paper	Country	Sample	Groups	Factor	Impact	Relevant Results	MMAT Score	Ecological Validity
Moffat and Hampson (1996)	Canada	80 students	Gender	Biological	Yes	Relationship between testosterone and spatial ability found, suggesting hormonal influence on score differences	20%	No
Moody (1997)	USA	46 students	Gender	Biological	Yes	Men scored higher than women in luteal phase but same as women in menstrual phase in spatial task	60%	No
Moreau et al. (2012)	France	62 students	Gender	Experience	Maybe	10 month wrestling training improved spatial performance equally for men and women	80%	No
Nguyen et al. (2003)	USA	172 students	Ethnicity	Behaviour Anxiety Attitudes Stereotype Threat	Yes Yes No No	Black participants under threat reported more anxiety; regulation of cognition partially mediated ethnic effects on performance	60%	No
Nyborg and Jensen (2000)	USA	4,037 male veterans	Ethnicity	Latent Trait/ Measurement invariance	No	Spearman's hypothesis supported in that magnitude of Black/White score differences strongly correlated with items' g-loading	40%	Yes
O'Brien and Crandall (2003)	USA	164 students	Gender	Anxiety Attitudes Stereotype Threat	No No Yes	Arousal-based explanation for threat supported as women under threat performed better at an easy task and worse at a difficult one; women were found to be more anxious, but this was unrelated to performance; no differences found for motivation	20%	No
Ortner and Sieverding (2008)	Austria	161 adults	Gender	Stereotype Threat	Yes	Priming male gender stereotypes led to increased spatial performance in women but not men	20%	No
Palumbo and Steele-Johnson (2014)	USA	320 students	Ethnicity	Attitudes Stereotype Threat	Maybe No	No support for stereotype threat, some support for Black people believing CATs to be more a measure of knowledge which was related to lower test scores	0%	No
Peña et al. (2008)	Spain	1,763 applicants	Gender	Experience	Yes	Gender differences found on strategies used on spatial test for applicants for air traffic control job	60%	Yes

Table 4 (continued)

Paper	Country	Sample	Groups	Factor	Impact	Relevant Results	MMAT Score	Ecological Validity
Penner and Willer (2011)	USA	84 students	Gender	Biological Stereotype Threat	Yes Yes	Manipulating glucose levels suggested impact of threat was more likely caused by arousal than ego depletion	40%	No
Pennington and Heim (2016)	UK	144 female students	Gender	Stereotype Threat	Yes	Testing in the presence of other women reduced the impact of threat	40%	No
Phillips and Silverman (1997)	Canada	60 female students	Gender	Biological	Yes	Cycle phase was related to performance on difficult spatial tasks	20%	No
Picucci et al. (2011)	Italy	91 students	Gender	Behaviour Attitudes	Yes Yes	Controlling for performance, women were less confident and employed different strategies than men on spatial task	60%	No
Pletzer et al. (2019)	Austria	82 adults	Gender	Biological Environmental	Maybe Maybe	Some support for sex hormones and gender roles being related to spatial performance	60%	No
Ployhart and Ehrhart (2002)	USA	500	Ethnicity	Attitudes	Yes	Monte Carlo simulation informed by prior research found reducing ethnicity differences in test motivation could reduce performance differences by 5–30%	80%	No
Ployhart et al. (2003)	USA	394 students	Ethnicity	Anxiety Attitudes Stereotype Threat	Yes Yes Yes	Some support for impact of threat in selection context; perceptions of threat were related to perceptions of face validity, motivation and anxiety; motivation and anxiety were related to performance	40%	No
Puts et al. (2010)	USA	337 students	Gender	Biological	No	Testosterone levels were not related to spatial test performance in men or women	40%	No
Quinn and Spencer (2001)	USA	144 students	Gender	Stereotype Threat	Yes	In two studies, women under threat were less able to formulate effective problem-solving strategies	40%	No

Table 4 (continued)

Paper	Country	Sample	Groups	Factor	Impact	Relevant Results	MMAT Score	Ecological Validity
Rahe and Jansen (2022)	Germany	117 adults	Gender	Item Bias Stereotype Threat	Yes Maybe	Men outperformed women on male stereotyped mental rotation items, but not female stereotyped items, potentially due to stereotype threat or object familiarity	40%	No
Rahe et al. (2020)	Germany	323 students	Gender	Biological, Item Bias, Experience	Yes	Gender differences on a mental rotation task were found for male-stereotyped objects but not female-stereotyped objects	60%	No
Razani et al. (2007)	USA	86 adults	Ethnicity	Environmental	Yes	Ethnic differences more pronounced on verbal tasks suggesting environmental effect	40%	No
Reeb (1976)	Israel	Large sample of military recruits	Ethnicity	Criterion Validity	Maybe	Differential validity by ethnicity largely eliminated when controlling for education	40%	Yes
Reeve et al. (2008)	USA	104 psychology students	Gender	Anxiety	Yes	Women more anxious in high-stakes testing, which may influence test scores	20%	No
Rilea (2008)	USA	73 students	Gender	Item Bias Behaviour	Yes Yes	Gender differences on spatial performance reduced when tasks included human figures suggesting holistic strategy use	40%	No
Roberts and Bell (2000)	USA	44 students	Gender	Experience	Yes	Familiarisation task to computer resulted in no gender difference on spatial task	40%	No
Roberts and Skinner (1996)	USA	13,559 military cadets	Gender, Ethnicity	Criterion Validity	No	Regression analysis found small overprediction for women and Black candidates	20%	Yes
Rosenberg and Park (2002)	USA	18 female students	Gender	Biological	Maybe	Some support for link between oestrogen and verbal ability in very small study	20%	No
Roth et al. (2014)	USA	Simulation	Ethnicity	Criterion Validity	No	Range restriction accounts for differential validity on cognitive tests	80%	No
Rotundo and Sackett (1999)	USA	23,806 employees	Ethnicity	Criterion Validity	Maybe	Evidence for predictive bias not found when comparing Black and White supervisor ratings; though test showed significantly lower validity for Black ratees	80%	Yes

Table 4 (continued)

Paper	Country	Sample	Groups	Factor	Impact	Relevant Results	MMAT Score	Ecological Validity
Rushton and Skuy (2000)	South Africa	309 students	Ethnicity	Latent Trait/ Measurement invariance Item Bias	No Maybe	Black/White differences more pronounced on g-saturated items suggesting difference in latent trait; gender differences not found on g potentially suggesting item bias	40%	No
Rushton et al. (2004)	South Africa	306 students	Ethnicity	Latent Trait/ Measurement invariance	No	Supported latent trait account for group differences as score differences most pronounced on highly g-loaded items and factor structure equivalent between groups	40%	No
Rushton et al. (2007a)	USA, Serbia, South Africa	152 USA twin pairs; 1045 comparion sample	Ethnicity	Biological Environmental	Yes Yes	Moderate correlations found in twin study between genetic and environmental influence on test items and group differences on those items, suggesting genetic and environmental causes	20%	No
Rushton et al. (2007b)	Serbia, South Africa	323 Serbian Roma,	Ethnicity	Latent Trait/ Measurement invariance	No	Correlating item difficulties between Serbian Roma and Black South Africans potentially supports measurement invariance	20%	No
Ryan et al. (2015)	USA	67 students	Gender	Stereotype Threat	Maybe	Priming positive or negative sports words had inconsistent results on athletes and non-athletes	0%	No
Saccuzzo et al. (1996)	USA	240 students	Gender	Experience	Yes	Found practice effect on spatial test greater for women and gender gap wasn't significant in second session	60%	No
Sackett et al. (2023)	USA	19,294 applicants	Ethnicity	Criterion Validity	No	Using data from previous studies suggest that ability tests don't underpredict Hispanic work performance	40%	Yes
Sanchis-Segura et al. (2018)	Spain	105 students	Gender	Stereotype Threat	Yes	Stereotypic association correlated with spatial test performance in men and women; priming led to gender differences, self-confidence mediated effects	60%	No
Scherbaum et al. (2011)	USA	173 students	Ethnicity	Behaviour Stereotype Threat	Yes Yes	Threat reduced Black people's performance with some evidence that they employed less advantageous test behaviour	40%	No
Schmader and Johns (2003)	USA	158 students	Gender, Ethnicity	Anxiety Stereotype Threat	No Yes	Three studies showing threat impaired working memory but was not related to anxiety	60%	No

Table 4 (continued)

Paper	Country	Sample	Groups	Factor	Impact	Relevant Results	MMAT Score	Ecological Validity
Schmader (2002)	USA	65 students	Gender	Stereotype Threat	Yes	Only women's gender identity moderated effects of threat on performance	60%	No
Schmader et al. (2009)	USA	231 students	Gender, Ethnicity	Anxiety Stereotype Threat	Yes Yes	Three studies showing that interpretation of anxious arousal under threat impacts performance	80%	No
Schöning et al. (2007)	Germany	24 people	Gender	Biological	Yes	fMRI evidence suggests gender differences and menstrual cycle differences when completing cognitive tests	40%	No
Schuster et al. (2015)	Europe	113 students	Gender	Stereotype Threat	Yes	Reappraising mind-wandering as normal reduced impact of threat in women	60%	No
Shapiro et al. (2013)	USA	245 students	Ethnicity	Stereotype Threat	Yes	Demonstrated effect of threat on tests scores in four studies and highlighted distinction between self-as-target and group-as-target of threat in efficacy of interventions	40%	No
Sharps et al. (1993)	USA	124 students	Gender	Item Bias Stereotype Threat	Yes Yes	Small samples finding gender differences on spatial tasks could be reduced by altering the context or instructions of items	30%	No
Sharps et al. (1994)	USA	112 students	Gender	Stereotype Threat	Yes	Gender differences on difficult spatial tasks were reduced when instructions did not highlight spatial nature of test	0%	No
Shih et al. (1999)	North America	46 female Asian students	Gender	Stereotype Threat	Yes	Asian women performed better on a numerical test when ethnicity was primed and worse when gender was primed	60%	No
Shih et al. (2002)	USA	152 students	Gender	Stereotype Threat	Yes	Self-relevance and the subtlety of activation were found to impact the effects of threat	80%	No
Shute et al. (1983)	USA	115 students	Gender	Biological	Yes	Androgen levels were positively related to spatial performance in females and negatively in males	20%	No

Table 4 (continued)

Paper	Country	Sample	Groups	Factor	Impact	Relevant Results	MMAT Score	Ecological Validity
Silverman and Phillips (1993)	Canada	475 + students	Gender	Biological	Yes	Women performed better at spatial test in menstrual phase	20%	No
Silverman et al. (1999)	Canada	84 students	Gender	Biological	Yes	Support for hormonal cause of gender differences on spatial tests in positive correlation between T and mental rotation performance	40%	No
Silverman et al. (2007)	International	244,893 people	Gender	Biological	Yes	Gender differences on spatial ability replicated in 40 countries supporting an evolutionary cause	40%	No
Šimić and Santini (2012)	Croatia	19 students	Gender	Biological	Yes	Performance on spatial and verbal tasks fluctuated throughout the cycle as did masculinity but not anxiety or femininity	40%	No
Skuy et al. (2002)	South Africa	98 students	Ethnicity	Experience	Yes	Practice and education improved performance of African participants more than non-African participants suggesting environmental factors	60%	No
Smith and White (2002)	USA	143 students	Gender, Ethnicity	Stereotype Threat	Yes	Demonstrated effect of threat even when not explicitly primed	40%	No
Sokolowski et al. (2019)	Canada	175 students	Gender	Anxiety	Yes	Women showed more maths anxiety and anxiety influences performance	40%	No
Spencer and Castano (2007)	USA	46 students	SES	Stereotype Threat	Yes	Lower SES participants performed worse when SES was salient and when test presented as diagnostic of ability	20%	No
Spencer et al. (1999)	USA	177 students	Gender	Stereotype Threat	Yes	Suggesting tests showed no gender differences removed threat and subsequent gender differences in test scores	60%	No

Table 4 (continued)

Paper	Country	Sample	Groups	Factor	Impact	Relevant Results	MMAT Score	Ecological Validity
Steele and Aronson (1995)	USA	269 students	Ethnicity	Stereotype Threat	Yes	Four studies demonstrating that priming threat in tests positioned as diagnostic of ability was related to performance in Black students	60%	No
Stenlund et al. (2017)	Sweden	1129 students	Gender	Behaviour Anxiety Attitudes	Yes Yes No	Gender differences in test strategy and anxiety but not motivation in high-stakes education testing	20%	Yes
Tanzer et al. (1995)	USA, Austria	691 students and military	Nation	Item Bias Behaviour Attitudes	No Yes Yes	LLTM useful in detecting bias, no DIF found in spatial test, practice items can potentially reduce bias, test taking motivation and behaviour varied across regions	20%	No
Te Nijenhuis and van der Flier (1997)	Netherlands	2,128 blue collar applicants	Ethnicity	Latent Trait/ Measurement invariance	Maybe	Spearman's hypothesis supported though difference in construct validity, modest DIF and greater mean score differences on more language-dependent tasks comparing immigrants with Dutch people	20%	Yes
Te Nijenhuis and Van der Flier (2000)	Netherlands	156 truck driver	Ethnicity	Criterion Validity	Maybe	Some evidence of differential prediction found in some criteria	20%	Yes
Terlecki and Newcombe (2005)	USA	180 students	Gender	Experience	Yes	Computer experience partially mediated the relationship between gender and spatial performance	20%	No
Terlecki et al. (2008)	USA	180 students	Gender	Experience	Maybe	Video game training and practice improved spatial performance, and women improved more over time	0%	No
Thames et al. (2015)	USA	76 adults	Ethnicity	Anxiety	Maybe	Relationship between anxiety and test performance was difference between Black and White participants	20%	No
Thilers et al. (2006)	Sweden	2383 adults	Gender	Biological	Maybe	Large study showing small relationships between testosterone and various cognitive abilities in men and women	60%	No

Table 4 (continued)

Paper	Country	Sample	Groups	Factor	Impact	Relevant Results	MMAT Score	Ecological Validity
Thomsen et al. (2000)	Norway	11 people	Gender	Biological	Yes	Small study providing limited support for gender differences in brain activation when completing spatial task indicates differences in strategies	40%	No
Tine and Gotlieb (2013)	USA	71 students	Gender, Ethnicity, SES	Stereotype Threat	Yes	Supported effect of threat. Participants with multiple stigmatised identities were impacted by threat more and reported putting in more effort	60%	No
Toivainen et al. (2018)	UK	1,464 pairs of twins	Gender	Biological	No	No support found that females with male twins outperform those with female twins in spatial tasks	40%	No
Unterrainer et al. (2000)	Austria	13 students	Gender	Biological	No	Small study finding no brain activation differences between men and women when completing spatial task	60%	No
Vanderpool and Catano (2008)	Canada		Ethnicity	Environmental Item Bias	Yes No	Verbal test scores are lower when English isn't first language; little evidence of DIF was found across multiple tests	40%	No
Vasta et al. (1996)	USA	180 students	Gender	Experience	Maybe	Training eliminated gender differences in specific spatial task though gains did not transfer to other spatial tasks	40%	No
Vernon and Jensen (1984)	USA	106 students	Ethnicity	Latent Trait/ Measurement invariance	No	Findings suggest mean score differences are related to differences in processing speed in elementary cognitive tasks	20%	No
Voyer and Jansen (2016)	Canada	120 students	Gender	Item Bias	No	Spatial gender differences not reduced when human figures used rather than abstract shapes	40%	No
Vuoksimaa et al. (2010)	Finland	804 twins	Gender	Biological	Yes	Females with male twins outperformed those with female twins, providing support for prenatal masculinization hypothesis	20%	No

Table 4 (continued)

Paper	Country	Sample	Groups	Factor	Impact	Relevant Results	MMAT Score	Ecological Validity
Vuoksimaa et al. (2012)	Finland	308 twins	Gender	Biological	Yes	Pubescent testosterone levels were negatively correlated with adult spatial performance in men and not women in longitudinal study	40%	No
Walsh et al. (1999)	Canada	174 students	Gender	Item Bias Stereotype Threat	No Yes	Gendered items did not account for performance differences, but evidence that threat impacted women's scores	20%	No
Wee et al. (2014)	USA	34,569 employees	Ethnicity	Criterion Validity	Maybe	Weighting subtests differently can reduce adverse impact without sacrificing criterion validity	40%	Yes
Weger et al. (2012)	UK	71 female students	Gender	Stereotype Threat	Yes	Performance detriments under threat were eliminated after a mindfulness exercise	40%	No
Weiss et al. (2003)	Austria	20 students	Gender	Biological	Yes	Controlling for ability, men and women showed differing brain activation in spatial tasks suggesting biological differences	40%	No
Wicherts (2017)	International	5,547 mostly students	Ethnicity	Latent Trait/ Measurement invariance Item Bias	Maybe Maybe	Demonstrates that common use of CTT-based Method of Correlated Vectors is inappropriate to detect Jenson Effect and suggests using IRT methods	40%	No
Wicherts et al. (2005)	Netherlands, USA	442 students	Gender, Ethnicity	Latent Trait/ Measurement invariance Stereotype Threat	Yes Yes	Two studies meeting inclusion criteria found evidence that stereotype threat was a source of measurement bias	60%	No
Wister et al. (2013)	USA	75 female students	Gender	Biological Stereotype Threat	Maybe No	Limited support for menstruation threat priming, some evidence for impact of time of cycle on cognitive ability	40%	No
Wraga et al. (2006)	USA	70 students	Gender	Stereotype Threat	Yes	Three studies with small samples showed how inducing threat impacted performance on mental rotation task	40%	No

Table 4 (continued)

Paper	Country	Sample	Groups	Factor	Impact	Relevant Results	MMAT Score	Ecological Validity
Wraga et al. (2007)	USA	54 student/graduate women	Gender	Stereotype Threat	Yes	Using fMRI, positive priming led to better processing efficiency and better performance; negative priming led to increased social/emotional activation and poorer performance	40%	No
Yang et al. (2007)	China	95 male students	Gender	Biological	Maybe	Potential support for testosterone influencing gender differences on spatial tests, in that high T men performed better only on certain difficult tasks	60%	No
Zoref and Williams (1980)	USA	Six CATs	Gender, Ethnicity	Item Bias	Maybe	Tests items tend to feature male/White people and stereotypical behaviour more	80%	No

Authors' Contributions All authors contributed to the study conception and design. Article searching, data collection and analysis were performed by Stephen Cuppello. The first draft of the manuscript was written by Stephen Cuppello and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding There is no funding source.

Data Availability (Data Transparency) The authors confirm that the data supporting the findings of this research are available within the article's appendix.

Declarations

Conflict of Interests The authors declare that they have no conflict of interests.

Informed Consent Not applicable.

References

Articles Marked with an Asterisk Were Included in the Systematic Review.

* Abad, F. J., Colom, R., Rebollo, I., & Escorial, S. (2004). Sex differential item functioning in the Raven's advanced progressive matrices: Evidence for bias. *Personality and Individual Differences, 36*(6), 1459–1470. https://doi.org/10.1016/s0191-8869(03)00241-1

* Aleman, A., Bronk, E., Kessels, R. P., Koppeschaar, H. P., & van Honk, J. (2004). A single administration of testosterone improves visuospatial ability in young women. *Psychoneuroendocrinology, 29*(5), 612–617. https://doi.org/10.1016/s0306-4530(03)00089-1

* Alexander, G. M., & Evardone, M. (2008). Blocks and bodies: Sex differences in a novel version of the mental rotations test. *Hormones and Behavior, 53*(1), 177–184. https://doi.org/10.1016/j.yhbeh.2007.09.014

* Alvarez-Vargas, D., Abad, C., & Pruden, S. M. (2020). Spatial anxiety mediates the sex difference in adult mental rotation test performance. *Cognitive Research: Principles and Implications, 5*(1), 1–17. https://doi.org/10.1186/s41235-020-00231-8

Anastasi, A. (1958). *Differential psychology* (3rd ed.). Macmillan. https://doi.org/10.1177/001316445901900431

* Arendasy, M. E., & Sommer, M. (2012). Gender differences in figural matrices: The moderating role of item design features. *Intelligence, 40*(6), 584–597. https://doi.org/10.1016/j.intell.2012.08.003

* Aronson, J., Fried, C. B., & Good, C. (2002). Reducing effects of stereotype threat on African American college students by shaping theories of intelligence. *Journal of Experimental Social Psychology, 38*(2), 113–125. https://doi.org/10.1006/jesp.2001.1491

* Arrighi, L., & Hausmann, M. (2022). Spatial anxiety and self-confidence mediate sex/gender differences in mental rotation. *Learning & Memory, 29*(9), 312–320. https://doi.org/10.1101/lm.053596.122

* Arvey, R. D., Strickland, W., Drauden, G., & Martin, C. (1990). Motivational components of test taking. *Personnel Psychology, 43*(4), 695–716. https://doi.org/10.1111/j.1744-6570.1990.tb00679.x

* Ashcraft, M. H., & Faust, M. W. (1994). Mathematics anxiety and mental arithmetic: An exploratory investigation. *Cognition & Emotion, 8*(2), 97–125. https://doi.org/10.1080/02699939408408931

Baenninger, M., & Newcombe, N. (1989). The role of experience in spatial test performance: A meta-analysis. *Sex Roles, 20*(5–6), 327–344. https://doi.org/10.1007/bf00287729

Barry, C. L., Horst, S. J., Finney, S. J., Brown, A. R., & Kopp, J. P. (2010). Do examinees have similar test-taking effort? A high-stakes question for low-stakes testing. *International Journal of Testing, 10*(4), 342–363. https://doi.org/10.1080/15305058.2010.508569

Bates, T. C., Lewis, G. J., & Weiss, A. (2013). Childhood socioeconomic status amplifies genetic effects on adult intelligence. *Psychological Science, 24*(10), 2111–2116. https://doi.org/10.1177/0956797613488394

* Bauer, R., Jost, L., & Jansen, P. (2021). The effect of mindfulness and stereotype threat in mental rotation: A pupillometry study. *Journal of Cognitive Psychology, 33*(8), 861–876. https://doi.org/10.1080/20445911.2021.1967366

* Beilock, S. L., Rydell, R. J., & McConnell, A. R. (2007). Stereotype threat and working memory: Mechanisms, alleviation, and spillover. *Journal of Experimental Psychology: General, 136*(2), 256–276. https://doi.org/10.1037/0096-3445.136.2.256

* Bell, E. C., Willson, M. C., Wilman, A. H., Dave, S., & Silverstone, P. H. (2006). Males and females differ in brain activation during cognitive tasks. *NeuroImage, 30*(2), 529–538. https://doi.org/10.1016/j.neuroimage.2005.09.049

* Ben-Zeev, T., Fein, S., & Inzlicht, B. (2005). Arousal and stereotype threat. *Journal of Experimental Social Psychology, 41*(2), 174–181. https://doi.org/10.1016/j.jesp.2003.11.007

* Berry, C. M., Barratt, C. L., Dovalina, C. L., & Zhao, P. (2014). Can racial/ethnic subgroup criterion-to-test standard deviation ratios account for conflicting differential validity and differential prediction evidence for cognitive ability tests? *Journal of Occupational and Organizational Psychology, 87*(1), 208–220. https://doi.org/10.1111/joop.12036

Berry, C. M., Clark, M. A., & McClure, T. K. (2011). Racial/ethnic differences in the criterion-related validity of cognitive ability tests: A qualitative and quantitative review. *Journal of Applied Psychology, 96*(5), 881–906. https://doi.org/10.1037/a0023222

Bertua, C., Anderson, N., & Salgado, J. F. (2005). The predictive validity of cognitive ability tests: A UK meta-analysis. *Journal of Occupational and Organizational Psychology, 78*(3), 387–409. https://doi.org/10.1348/096317905x26994

Binet, A., & Simon, T. (1907). Le développement de l'intelligence chez les enfants. *L'Année Psychologique, 14*(1), 1–94. https://doi.org/10.3406/psy.1907.3737

* Brown, R. P., & Day, E. A. (2006). The difference isn't black and white: Stereotype threat and the race gap on Raven's Advanced Progressive Matrices. *Journal of Applied Psychology, 91*(4), 979. https://doi.org/10.1037/0021-9010.91.4.979

* Brown, R. P., & Josephs, R. A. (1999). A burden of proof: Stereotype relevance and gender differences in math performance. *Journal of Personality and Social Psychology, 76*(2), 246–257. https://doi.org/10.1037/0022-3514.76.2.246

* Burkitt, J., Widman, D., & Saucier, D. M. (2007). Evidence for the influence of testosterone in the performance of spatial navigation in a virtual water maze in women but not in men. *Hormones and Behavior, 51*(5), 649–654. https://doi.org/10.1016/j.yhbeh.2007.03.007

* Burton, L., Henninger, D., & Hafetz, J. (2005). Gender differences in mental rotation, verbal fluency and SAT scores to finger length ratios as hormonal indexes. *Developmental Neuropsychology, 28*(1), 493–505. https://doi.org/10.1207/s15326942dn2801_3

* Butler, T., Imperato-McGinley, J., Pan, H., Voyer, D., Cordero, J., Zhu, Y. S., et al. (2006). Sex differences in mental rotation: Top-down versus bottom-up processing. *NeuroImage, 32*(1), 445–456. https://doi.org/10.1016/j.neuroimage.2006.03.030

* Cadinu, M., Maass, A., Lombardo, M., & Frigerio, S. (2006). Stereotype threat: The moderating role of locus of control beliefs. *European Journal of Social Psychology, 36*(2), 183–197. https://doi.org/10.1002/ejsp.303

* Campbell, M. J., Toth, A. J., & Brady, N. (2018). Illuminating sex differences in mental rotation using pupillometry. *Biological Psychology, 138*, 19–26. https://doi.org/10.1016/j.biopsycho.2018.08.003

* Campbell, S. M., & Collaer, M. L. (2009). Stereotype threat and gender differences in performance on a novel visuospatial task. *Psychology of Women Quarterly, 33*(4), 437–444. https://doi.org/10.1111/j.1471-6402.2009.01521.x

* Campion, M. C., Campion, E. D., & Campion, M. A. (2019). Using practice employment tests to improve recruitment and personnel selection outcomes for organizations and job seekers. *Journal of Applied Psychology, 104*(9), 1089. https://doi.org/10.1037/apl0000401

* Carr, P. B., & Steele, C. M. (2009). Stereotype threat and inflexible perseverance in problem solving. *Journal of Experimental Social Psychology, 45*(4), 853–859. https://doi.org/10.1016/j.jesp.2009.03.003

* Carretta, T. R., & Doub, T. W. (1998). Group differences in the role of g and prior job knowledge in the acquisition of subsequent job knowledge. *Personality and Individual Differences, 24*(5), 585–593. https://doi.org/10.1016/s0191-8869(97)00210-9
Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*, 1st edn. Cambridge University Press. https://doi.org/10.1017/cbo9780511571312
* Çetinkaya, E., Herrmann, S. D., & Kisbu-Sakarya, Y. (2020). Adapting the values affirmation intervention to a multi-stereotype threat framework for female students in STEM. *Social Psychology of Education, 23*(6), 1587–1607. https://doi.org/10.1007/s11218-020-09594-8
* Chan, D. (1997). Racial subgroup differences in predictive validity perceptions on personality and cognitive ability tests. *Journal of Applied Psychology, 82*(2), 311. https://doi.org/10.1037/0021-9010.82.2.311
* Chan, D., Schmitt, N., DeShon, R. P., Clause, C. S., & Delbridge, K. (1997). Reactions to cognitive ability tests: The relationships between race, test performance, face validity perceptions, and test-taking motivation. *Journal of Applied Psychology, 82*(2), 300. https://doi.org/10.1037/0021-9010.82.2.300
* Chan, D., Schmitt, N., Sacco, J. M., & DeShon, R. P. (1998). Understanding pretest and posttest reactions to cognitive ability and personality tests: Performance-reactions relationships and their structural invariance across racial groups. *Journal of Applied Psychology, 83*(3), 471–485. https://doi.org/10.1037/0021-9010.83.3.471
* Cherney, I. D. (2008). Mom, let me play more computer games: They improve my mental rotation skills. *Sex Roles, 59*(11–12), 776–786. https://doi.org/10.1007/s11199-008-9498-z
* Cherney, I. D., & Collaer, M. (2005). Sex differences in line judgment: Relation to mathematics preparation and strategy use. *Perceptual and Motor Skills, 100*(3), 615–627. https://doi.org/10.2466/pms.100.3.615-627
* Cherney, I. D., Bersted, K., & Smetter, J. (2014). Training spatial skills in men and women. *Perceptual and Motor Skills, 119*(1), 82–99. https://doi.org/10.2466/23.25.pms.119c12z0
* Cherney, I. D., Jagarlamudi, K., Lawrence, E., & Shimabuku, N. (2003). Experiential factors on sex differences in mental rotation. *Perceptual and Motor Skills, 96*(3), 1062–1070. https://doi.org/10.2466/pms.2003.96.3c.1062
* Clark, J. K., Eno, C. A., & Guadagno, R. E. (2011). Southern discomfort: The effects of stereotype threat on the intellectual performance of US southerners. *Self and Identity, 10*(2), 248–262. https://doi.org/10.1080/15298861003771080
* Cockcroft, K., Alloway, T., Copello, E., & Milligan, R. (2015). A cross-cultural comparison between South African and British students on the Wechsler Adult Intelligence Scales Third Edition (WAIS-III). *Frontiers in Psychology, 6*, Article 297. https://doi.org/10.3389/fpsyg.2015.00297
* Collaer, M. L., Reimers, S., & Manning, J. T. (2007). Visuospatial performance on an internet line judgment task and potential hormonal markers: Sex, sexual orientation, and 2D: 4D. *Archives of Sexual Behavior, 36*(2), 177–192. https://doi.org/10.1007/s10508-006-9152-1
* Cooke-Simpson, A., & Voyer, D. (2007). Confidence and gender differences on the mental rotations test. *Learning and Individual Differences, 17*(2), 181–186. https://doi.org/10.1016/j.lindif.2007.03.009
* Courvoisier, D. S., Renaud, O., Geiser, C., Paschke, K., Gaudy, K., & Jordan, K. (2013). Sex hormones and mental rotation: An intensive longitudinal investigation. *Hormones and Behavior, 63*(2), 345–351. https://doi.org/10.1016/j.yhbeh.2012.12.007
* Croizet, J. C., & Claire, T. (1998). Extending the concept of ST to a social class: The intellectual underperformance of students from low socioeconomic backgrounds. *Personality and Social Psychology Bulletin, 24*(6), 588–594. https://doi.org/10.1177/0146167298246003
* De Meijer, L. A., Born, M. P., Terlouw, G., & Van Der Molen, H. T. (2008). Criterion-related validity of Dutch police-selection measures and differences between ethnic groups. *International Journal of Selection and Assessment, 16*(4), 321–332. https://doi.org/10.1111/j.1468-2389.2008.00438.x
* Dennehy, T. C., Ben-Zeev, A., & Tanigawa, N. (2014). 'Be prepared': An implemental mindset for alleviating social-identity threat. *British Journal of Social Psychology, 53*(3), 585–594. https://doi.org/10.1111/bjso.12071
* DeShon, R. P., Smith, M. R., Chan, D., & Schmitt, N. (1998). Can racial differences in cognitive test performance be reduced by presenting problems in a social context? *Journal of Applied Psychology, 83*(3), 438. https://doi.org/10.1037/0021-9010.83.3.438
* Díaz, A., Sellami, K., Infanzón, E., Lanzón, T., & Lynn, R. (2012). A comparative study of general intelligence in Spanish and Moroccan samples. *The Spanish Journal of Psychology, 15*(2), 526–532. https://doi.org/10.5209/rev_sjop.2012.v15.n2.38863

* Dolan, C. V., Colom, R., Abad, F. J., Wicherts, J. M., Hessen, D. J., & Van De Sluis, S. (2006). Multi-group covariance and mean structure modelling of the relationship between the WAIS-III common factors and sex and educational attainment in Spain. *Intelligence, 34*(2), 193–210. https://doi.org/10.1016/j.intell.2005.09.003
* Dolan, C. V., Roorda, W., & Wicherts, J. M. (2004). Two failures of spearman's hypothesis: The GATB in Holland and the JAT in South Africa. *Intelligence, 32*(2), 155–173. https://doi.org/10.1016/j.intell.2003.09.001
* Domino, G., & Morales, A. (2000). Reliability and validity of the D-48 with Mexican American college students. *Hispanic Journal of Behavioral Sciences, 22*(3), 382–389. https://doi.org/10.1177/0739986300223007
* Doyle, R. A., & Voyer, D. (2018). Photographs of real human figures: Item types and persistent sex differences in mental rotation. *Quarterly Journal of Experimental Psychology, 71*(11), 2411–2420. https://doi.org/10.1177/1747021817742079
* Ellis, A. P., & Ryan, A. M. (2003). Race and cognitive-ability test performance: The mediating effects of test preparation, test-taking strategy use and self-efficacy. *Journal of Applied Social Psychology, 33*(12), 2607–2629. https://doi.org/10.1111/j.1559-1816.2003.tb02783.x
* Epting, L. K., & Overman, W. H. (1998). Sex-sensitive tasks in men and women: A search for performance fluctuations across the menstrual cycle. *Behavioral Neuroscience, 112*(6), 1304. https://doi.org/10.1037/0735-7044.112.6.1304
* Estes, Z., & Felker, S. (2012). Confidence mediates the sex difference in mental rotation performance. *Archives of Sexual Behavior, 41*(3), 557–570. https://doi.org/10.1007/s10508-011-9875-5
* Fagan, J. F., & Holland, C. R. (2002). Equal opportunity and racial differences in IQ. *Intelligence, 30*(4), 361–387. https://doi.org/10.1016/s0160-2896(02)00080-6
* Fagan, J. F., & Holland, C. R. (2007). Racial equality in intelligence: Predictions from a theory of intelligence as processing. *Intelligence, 35*(4), 319–334. https://doi.org/10.1016/j.intell.2006.08.009
* Falter, C. M., Arroyo, M., & Davis, G. (2006). Testosterone: Activation or organization of spatial cognition? *Biological Psychology, 73*(2), 132–140. https://doi.org/10.1016/j.biopsycho.2006.01.011
* Fasfous, A. F., Hidalgo-Ruzzante, N., Vilar-López, R., Catena-Martínez, A., & Pérez-García, M. (2013). Cultural differences in neuropsychological abilities required to perform intelligence tasks. *Archives of Clinical Neuropsychology, 28*(8), 784–790. https://doi.org/10.1093/arclin/act074
* Feng, J., Spence, I., & Pratt, J. (2007). Playing an action video game reduces gender differences in spatial cognition. *Psychological Science, 18*(10), 850–855. https://doi.org/10.1111/j.1467-9280.2007.01990.x
Finnigan, K. M., & Corker, K. S. (2016). Do performance avoidance goals moderate the effect of different types of stereotype threat on women's math performance? *Journal of Research in Personality, 63*, 36–43. https://doi.org/10.1016/j.jrp.2016.05.009
* Fisher, M. L., Meredith, T., & Gray, M. (2018). Sex differences in mental rotation ability are a consequence of procedure and artificiality of stimuli. *Evolutionary Psychological Science, 4*(2), 124–133. https://doi.org/10.1007/s40806-017-0120-x
* Forbes, C. E., & Schmader, T. (2010). Retraining attitudes and stereotypes to affect motivation and cognitive capacity under stereotype threat. *Journal of Personality and Social Psychology, 99*(5), 740. https://doi.org/10.1037/a0020971
* Forbes, C. E., Leitner, J. B., Duran-Jordan, K., Magerman, A. B., Schmader, T., & Allen, J. J. (2015). Spontaneous default mode network phase-locking moderates performance perceptions under stereotype threat. *Social Cognitive and Affective Neuroscience, 10*(7), 994–1002. https://doi.org/10.1093/scan/nsu145
* Frenken, H., Papageorgiou, K. A., Tikhomirova, T., Malykh, S., Tosto, M. G., & Kovas, Y. (2016). Siblings' sex is linked to mental rotation performance in males but not females. *Intelligence, 55*, 38–43. https://doi.org/10.1016/j.intell.2016.01.005
* Frisby, C. L., & Osterlind, S. J. (2006). A descriptive analysis of test session observation checklist ratings from the Woodcock Johnson III standardization sample. *Journal of Psychoeducational Assessment, 24*(4), 342–357. https://doi.org/10.1177/0734282906288525
* Gabriel, K. I., Hong, S. M., Chandra, M., Lonborg, S. D., & Barkley, C. L. (2011). Gender differences in the effects of acute stress on spatial ability. *Sex Roles, 64*(1), 81–89. https://doi.org/10.1007/s11199-010-9877-0
* Gardner, D. G., & Deadrick, D. L. (2012). Moderation of selection procedure validity by employee race. *Journal of Managerial Psychology, 27*(4), 365–382. https://doi.org/10.1108/02683941211220180

* Gardner, D., & Deadrick, D. L. (2008). Underprediction of performance for US minorities using cognitive ability measures. *Equal Opportunities International, 27*(5), 455–464. https://doi.org/10.1108/02610150810882305
* Gillespie, J. Z., Converse, P. D., & Kriska, S. D. (2010). Applying recommendations from the literature on stereotype threat: Two field studies. *Journal of Business and Psychology, 25*, 493–504. https://doi.org/10.1007/s10869-010-9178-1
* Gizewski, E. R., Krause, E., Wanke, I., Forsting, M., & Senf, W. (2006). Genderspecific cerebral activation during cognitive tasks using functional MRI: Comparison of women in mid-luteal phase and men. *Neuroradiology, 48*(1), 14–20. https://doi.org/10.1007/s00234-005-0004-9
Gledhill, A., Forsdyke, D., & Murray, E. (2018). Psychological interventions used to reduce sports injuries: A systematic review of real-world effectiveness. *British Journal of Sports Medicine, 52*(15), 967–971. https://doi.org/10.1136/bjsports-2017-097694
Gomez, L. E., & Bernet, P. (2019). Diversity improves performance and outcomes. *Journal of the National Medical Association, 111*(4), 383–392. https://doi.org/10.1016/j.jnma.2019.01.006
* Gonzales, P. M., Blanton, H., & Williams, K. J. (2002). The effects of stereotype threat and double-minority status on the test performance of Latino women. *Personality and Social Psychology Bulletin, 28*(5), 659–670. https://doi.org/10.1177/0146167202288010
* Gordon, H. W., & Lee, P. A. (1986). A relationship between gonadotropins and visuospatial function. *Neuropsychologia, 24*(4), 563–576. https://doi.org/10.1016/0028-3932(86)90100-4
* Gordon, H. W., & Lee, P. A. (1993). No difference in cognitive performance between phases of the menstrual cycle. *Psychoneuroendocrinology, 18*(7), 521–531. https://doi.org/10.1016/0306-4530(93)90045-m
* Gordon, H. W., Corbin, E. D., & Lee, P. A. (1986). Changes in specialized cognitive function following changes in hormone levels. *Cortex, 22*(3), 399–415. https://doi.org/10.1016/s0010-9452(86)80004-1
* Gouchie, C., & Kimura, D. (1991). The relationship between testosterone levels and cognitive ability patterns. *Psychoneuroendocrinology, 16*(4), 323–334. https://doi.org/10.1016/0306-4530(91)90018-o
* Grand, J. A., Ryan, A. M., Schmitt, N., & Hmurovic, J. (2010). How far does stereotype threat reach? The potential detriment of face validity in cognitive ability testing. *Human Performance, 24*(1), 1–28. https://doi.org/10.1080/08959285.2010.518184
* Gresky, D. M., Ten Eyck, L. L., Lord, C. G., & McIntyre, R. B. (2005). Effects of Salient Multiple Identities on Women's Performance under Mathematics Stereotype Threat. *Sex Roles, 53*, 703–715. https://doi.org/10.1007/s11199-005-7735-2
* Griksiene, R., & Ruksenas, O. (2011). Effects of hormonal contraceptives on mental rotation and verbal fluency. *Psychoneuroendocrinology, 36*(8), 1239–1248. https://doi.org/10.1016/j.psyneuen.2011.03.001
* Griksiene, R., Arnatkeviciute, A., Monciunskaite, R., Koenig, T., & Ruksenas, O. (2019). Mental rotation of sequentially presented 3D figures: Sex and sex hormones related differences in behavioural and ERP measures. *Scientific Reports, 9*(1), 1–18. https://doi.org/10.1038/s41598-019-55433-y
* Griksiene, R., Monciunskaite, R., Arnatkeviciute, A., & Ruksenas, O. (2018). Does the use of hormonal contraceptives affect the mental rotation performance? *Hormones and Behavior, 100*, 29–38. https://doi.org/10.1016/j.yhbeh.2018.03.004
* Grubb, H. J., & Ollendick, T. H. (1986). Cultural-distance perspective: An exploratory analysis of its effect on learning and intelligence. *International Journal of Intercultural Relations, 10*(4), 399–414. https://doi.org/10.1016/0147-1767(86)90042-8
* Gur, R. C., Alsop, D., Glahn, D., Petty, R., Swanson, C. L., Maldjian, J. A., Turetsky, B. I., Detre, J. A., Gee, J., & Gur, R. E. (2000). An fMRI study of sex differences in regional activation to a verbal and a spatial task. *Brain and Language, 74*(2), 157–170. https://doi.org/10.1006/brln.2000.2325
* Halari, R., Hines, M., Kumari, V., Mehrotra, R., Wheeler, M., Ng, V., & Sharma, T. (2005). Sex differences and individual differences in cognitive performance and their relationship to endogenous gonadal hormones and gonadotropins. *Behavioral Neuroscience, 119*(1), 104. https://doi.org/10.1037/0735-7044.119.1.104
* Halari, R., Sharma, T., Hines, M., Andrew, C., Simmons, A., & Kumari, V. (2006). Comparable fMRI activity with differential behavioural performance on mental rotation and overt verbal fluency tasks in healthy men and women. *Experimental Brain Research, 169*, 1–14. https://doi.org/10.1007/s00221-005-0118-7
* Halpern, D. F., & Tan, U. (2001). Stereotypes and steroids: Using a psychobiosocial model to understand cognitive sex differences. *Brain and Cognition, 45*(3), 392–414. https://doi.org/10.1006/brcg.2001.1287

* Hampson, E. (1990a). Estrogen-related variations in human spatial and articulatory-motor skills. *Psychoneuroendocrinology, 15*(2), 97–111. https://doi.org/10.1016/0306-4530(90)90018-5
* Hampson, E. (1990b). Variations in sex-related cognitive abilities across the menstrual cycle. *Brain and Cognition, 14*(1), 26–43. https://doi.org/10.1016/0278-2626(90)90058-v
* Hampson, E., Morley, E. E., Evans, K. L., & Fleury, C. (2022). Effects of oral contraceptives on spatial cognition depend on pharmacological properties and phase of the contraceptive cycle. *Frontiers in Endocrinology*. https://doi.org/10.3389/fendo.2022.888510
Hanel, P. H., & Vione, K. C. (2016). Do student samples provide an accurate estimate of the general public? *PLoS One, 11*(12), Article e0168354. https://doi.org/10.1371/journal.pone.0168354
* Harada, T., Bridge, D. J., & Chiao, J. Y. (2013). Dynamic social power modulates neural basis of math calculation. *Frontiers in Human Neuroscience, 6*(350), 115–127. https://doi.org/10.3389/fnhum.2012.00350
* Harrison, L. A., Stevens, C. M., Monty, A. N., & Coakley, C. A. (2006). The consequences of stereotype threat on the academic performances of White and non-White lower income college students. *Social Psychology of Education, 9*(3), 341–357. https://doi.org/10.1007/s11218-005-5456-6
* Harville, D. L. (1996). Ability test equity in predicting job performance work samples. *Educational and Psychological Measurement, 56*(2), 344–348. https://doi.org/10.1177/0013164496056002015
* Hausdorf, P. A., & Robie, C. (2018). Understanding subgroup differences with general mental ability tests in employment selection: Exploring socio-cultural factors across inter-generational groups. *International Journal of Selection and Assessment, 26*(2–4), 176–190. https://doi.org/10.1111/ijsa.12226
* Hausmann, M. (2014). Arts versus science—Academic background implicitly activates gender stereotypes on cognitive abilities with threat raising men's (but lowering women's) performance. *Intelligence, 46*, 235–245. https://doi.org/10.1016/j.intell.2014.07.004
* Hausmann, M., Schoofs, D., Rosenthal, H. E., & Jordan, K. (2009). Interactive effects of sex hormones and gender stereotypes on cognitive sex differences—A psychobiosocial approach. *Psychoneuroendocrinology, 34*(3), 389–401. https://doi.org/10.1016/j.psyneuen.2008.09.019
* Hausmann, M., Slabbekoorn, D., Van Goozen, S. H., Cohen-Kettenis, P. T., & Güntürkün, O. (2000). Sex hormones affect spatial abilities during the menstrual cycle. *Behavioral Neuroscience, 114*(6), 1245. https://doi.org/10.1037/0735-7044.114.6.1245
* Heil, M., & Jansen-Osmann, P. (2008). Sex differences in mental rotation with polygons of different complexity: Do men utilize holistic processes whereas women prefer piecemeal ones? *Quarterly Journal of Experimental Psychology, 61*(5), 683–689. https://doi.org/10.1080/17470210701822967
* Heil, M., Kavšek, M., Rolke, B., Beste, C., & Jansen, P. (2011). Mental rotation in female fraternal twins: Evidence for intra-uterine hormone transfer? *Biological Psychology, 86*(1), 90–93. https://doi.org/10.1016/j.biopsycho.2010.11.002
* Hirnstein, M., Bayer, U., & Hausmann, M. (2009). Sex-specific response strategies in mental rotation. *Learning and Individual Differences, 19*(2), 225–228. https://doi.org/10.1016/j.lindif.2008.11.006
* Hirnstein, M., Coloma Andrews, L., & Hausmann, M. (2014). Gender-stereotyping and cognitive sex differences in mixed-and same-sex groups. *Archives of Sexual Behavior, 43*, 1663–1673. https://doi.org/10.1007/s10508-014-0311-5
* Hirnstein, M., Freund, N., & Hausmann, M. (2015). Gender stereotyping enhances verbal fluency performance in men (and women). *Zeitschrift Für Psychologie, 220*, 70–77. https://doi.org/10.1027/2151-2604/a000098
* Hollis-Sawyer, L. A., & Sawyer, T. P., Jr. (2008). Potential stereotype threat and face validity effects on cognitive-based test performance in the classroom. *Educational Psychology, 28*(3), 291–304. https://doi.org/10.1080/01443410701532313
Hong, Q. N., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M. P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M. C., Vedel, I., & Pluye, P. (2018). The mixed methods appraisal tool (MMAT) version 2018 for information professionals and researchers. *Education for Information, 34*(4), 285–291. https://doi.org/10.3233/efi-180221
* Hooven, C. K., Chabris, C. F., Ellison, P. T., & Kosslyn, S. M. (2004). The relationship of male testosterone to components of mental rotation. *Neuropsychologia, 42*(6), 782–790. https://doi.org/10.1016/j.neuropsychologia.2003.11.012
Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues, evidence and lessons learned. *International Journal of Selection and Assessment, 9*(1–2), 152–194. https://doi.org/10.1111/1468-2389.00171

* Houston, W. M. (1987). Race-based differential prediction in Air Force technical training programs. *Journal of Educational Measurement, 24*(4), 309–320. https://doi.org/10.1111/j.1745-3984.1987.tb00282.x
* Howard, R. C., Fenwick, P., Brown, D., & Norton, R. (1992). Relationship between CNV asymmetries and individual differences in cognitive performance, personality and gender. *International Journal of Psychophysiology, 13*(3), 191–197. https://doi.org/10.1016/0167-8760(92)90069-n
* Hugdahl, K., Thomsen, T., & Ersland, L. (2006). Sex differences in visuo-spatial processing: An *fMRI* study of mental rotation. *Neuropsychologia, 44*(9), 1575–1583. https://doi.org/10.1016/j.neuropsychologia.2006.01.026
Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*(1), 72. https://doi.org/10.1037/0033-2909.96.1.72
Hyde, J. S. (1990). Meta-analysis and the psychology of gender differences. *Signs: Journal of Women in Culture and Society, 16*(1), 55–73. https://doi.org/10.1086/494645
* Ispas, D., Iliescu, D., Ilie, A., & Johnson, R. E. (2010). Examining the criterion related validity of the general ability measure for adults: A two sample investigation. *International Journal of Selection and Assessment, 18*(2), 226–229. https://doi.org/10.1111/j.1468-2389.2010.00505.x
* Jansen, P., & Lehmann, J. (2013). Mental rotation performance in soccer players and gymnasts in an object-based mental rotation task. *Advances in Cognitive Psychology, 9*(2), 92–98. https://doi.org/10.5709/acp-0135-8
* Jaušovec, N., & Jaušovec, K. (2012). Sex differences in mental rotation and cortical activation patterns: Can training change them? *Intelligence, 40*(2), 151–162. https://doi.org/10.1016/j.intell.2012.01.005
* Jensen, A. R. (1977). An examination of culture bias in the Wonderlic Personnel Test. *Intelligence, 1*(1), 51–64. https://doi.org/10.1016/0160-2896(77)90026-5
Jensen, A. R. (1980). *Bias in Mental Testing*. Free Press.
* Jensen, A. R., & McGurk, F. C. (1987). Black-White bias in 'cultural' and 'noncultural' test items. *Personality and Individual Differences, 8*(3), 295–301. https://doi.org/10.1016/0191-8869(87)90029-8
* Johns, M., Inzlicht, M., & Schmader, T. (2008). Stereotype threat and executive resource depletion: Examining the influence of emotion regulation. *Journal of Experimental Psychology: General, 137*(4), 691. https://doi.org/10.1037/a0013834
* Johns, M., Schmader, T., & Martens, A. (2005). Knowing is half the battle: Teaching stereotype threat as a means of improving women's math performance. *Psychological Science, 16*(3), 175–179. https://doi.org/10.1111/j.0956-7976.2005.00799.x
Johnson, W., & Bouchard, T. J., Jr. (2007). Sex differences in mental abilities: G masks the dimensions on which they lie. *Intelligence, 35*(1), 23–39. https://doi.org/10.1016/j.intell.2006.03.012
* Jordan, K., Wuestenberg, T., Heinze, H. J., Peters, M., & Jaencke, L. (2002). Women and men exhibit different cortical activation patterns during mental rotation tasks. *Neuropsychologia, 40*(13), 2397–2408. https://doi.org/10.1016/s0028-3932(02)00076-3
* Josephs, R. A., Newman, M. L., Brown, R. P., & Beer, J. M. (2003). Status, testosterone, and human intellectual performance: Stereotype threat as status concern. *Psychological Science, 14*(2), 158–163. https://doi.org/10.1111/1467-9280.t01-1-01435
Kahn, S., & Ginther, D. (2017). Women and STEM. In NBER Working Papers 23525, National Bureau of Economic Research, Inc. Cambridge. https://doi.org/10.3386/w23525
* Kass, S. J., Ahlers, R. H., & Dugger, M. (1998). Eliminating gender differences through practice in an applied visual spatial task. *Human Performance, 11*(4), 337–349. https://doi.org/10.1207/s15327043hup1104_3
* Kempel, P., Gohlke, B., Klempau, J., Zinsberger, P., Reuter, M., & Hennig, J. (2005). Second-to-fourth digit length, testosterone and spatial ability. *Intelligence, 33*(3), 215–230. https://doi.org/10.1016/j.intell.2004.11.004
* Kiefer, A. K., & Sekaquaptewa, D. (2007). Implicit stereotypes, gender identification, and math-related outcomes: A prospective study of female college students. *Psychological Science, 18*(1), 13–18. https://doi.org/10.1111/j.1467-9280.2007.01841.x
Kirnan, J. P., Alfieri, J. A., Bragger, J. D., & Harris, R. S. (2009). An investigation of stereotype threat in employment tests. *Journal of Applied Social Psychology, 39*(2), 359–388. https://doi.org/10.1111/j.1559-1816.2008.00442.x
* Klein, O., Pohl, S., & Ndagijimana, C. (2007). The influence of intergroup comparisons on Africans' intelligence test performance in a job selection context. *The Journal of Psychology, 141*(5), 453–468. https://doi.org/10.3200/jrlp.141.5.453-468

* Konan, P. N. D., Chatard, A., Selimbegović, L., Mugny, G., & Moraru, A. (2011). Deflecting stereotype threat through downward comparison: When comparison with immigrants boosts the performance of stigmatized native students. *Social Justice Research, 24*(2), 191–205. https://doi.org/10.1007/s11211-011-0134-7

* Koscik, T., O'Leary, D., Moser, D. J., Andreasen, N. C., & Nopoulos, P. (2009). Sex differences in parietal lobe morphology: Relationship to mental rotation performance. *Brain and Cognition, 69*(3), 451–459. https://doi.org/10.1016/j.bandc.2008.09.004

Krapohl, E., & Plomin, R. (2016). Genetic link between family socioeconomic status and children's educational achievement estimated from genome-wide SNPs. *Molecular Psychiatry, 21*(3), 437–443. https://doi.org/10.1038/mp.2015.2

* Kumari, V., & Corr, P. J. (1998). Trait anxiety, stress and the menstrual cycle: Effect on raven´s standard progressive matrices test. *Personality and Individual Differences, 24*(5), 615–623. https://doi.org/10.1016/s0191-8869(97)00233-x

* Lefkowitz, J., & Battista, M. (1995). Potential sources of criterion bias in supervisor ratings used for test validation. *Journal of Business and Psychology, 11*(3), 389–414. https://doi.org/10.1007/bf02230978

* Levy, J., & Reid, M. (1978). Variations in cerebral organization as a function of handedness, hand posture in writing, and sex. *Journal of Experimental Psychology: General, 107*(2), 119. https://doi.org/10.1037/0096-3445.107.2.119

* Maertz, C. P., Jr., Bauer, T. N., Mosley, D. C., Jr., Posthuma, R. A., & Campion, M. A. (2005). Predictors of self-efficacy for cognitive ability employment testing. *Journal of Business Research, 58*(2), 160–167. https://doi.org/10.1016/s0148-2963(03)00111-5

* Maloney, E. A., Waechter, S., Risko, E. F., & Fugelsang, J. A. (2012). Reducing the sex difference in math anxiety: The role of spatial processing ability. *Learning and Individual Differences, 22*(3), 380–384. https://doi.org/10.1016/j.lindif.2012.01.001

* Martens, A., Johns, M., Greenberg, J., & Schimel, J. (2006). Combating stereotype threat: The effect of self-affirmation on women's intellectual performance. *Journal of Experimental Social Psychology, 42*(2), 236–243. https://doi.org/10.1016/j.jesp.2005.04.010

* Marx, D. M., & Goff, P. A. (2005). Clearing the air: The effect of experimenter race on target's test performance and subjective experience. *British Journal of Social Psychology, 44*(4), 645–657. https://doi.org/10.1348/014466604x17948

* Marx, D. M., & Roman, J. S. (2002). Female role models: Protecting women's math performance. *Personality and Social Psychology Bulletin, 28*(9), 1183–1193. https://doi.org/10.1177/01461672022812004

* Mayer, D. M., & Hanges, P. J. (2003). Understanding the stereotype threat effect with" culture-free" tests: An examination of its mediators and measurement. *Human Performance, 16*(3), 207–230. https://doi.org/10.1207/s15327043hup1603_3

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*(4), 282–295.

* McCormick, C. M., & Teillon, S. M. (2001). Menstrual cycle variation in spatial ability: Relation to salivary cortisol levels. *Hormones and Behavior, 39*(1), 29–38. https://doi.org/10.1207/s15327043hup1603_3

* McFarland, L. A., Lev-Arey, D. M., & Ziegert, J. C. (2003). An examination of stereotype threat in a motivational context. *Human Performance, 16*(3), 181–205. https://doi.org/10.1207/s15327043hup1603_2

* McGee, M. G. (1978). Effects of training and practice on sex differences in mental rotation test scores. *Journal of Psychology, 100*(1), 87–90. https://doi.org/10.1080/00223980.1978.9923476

*McGlone, M. S., & Aronson, J. (2006). Stereotype threat, identity salience, and spatial reasoning. *Journal of Applied Developmental Psychology, 27*(5), 486–493. https://doi.org/10.1016/j.appdev.2006.06.003

*McIntyre, R. B., Paulson, R. M., & Lord, C. G. (2003). Alleviating women's mathematics stereotype threat through salience of group achievements. *Journal of Experimental Social Psychology, 39*(1), 83–90. https://doi.org/10.1016/s0022-1031(02)00513-9

*McKay, P. F., Doverspike, D., Bowen-Hilton, D., & McKay, Q. D. (2003). The effects of demographic variables and stereotype threat on Black/White differences in cognitive ability test performance. *Journal of Business and Psychology, 18*(1), 1–14. https://doi.org/10.1023/a:1025062703113

*McKeever, W. F., Rich, D. A., Deyo, R. A., & Conner, R. L. (1987). Androgens and spatial ability: Failure to find a relationship between testosterone and ability measures. *Bulletin of the Psychonomic Society, 25*(6), 438–440. https://doi.org/10.3758/bf03334734

*Moè, A. (2016). Teaching motivation and strategies to improve mental rotation abilities. *Intelligence, 59*, 16–23. https://doi.org/10.1016/j.intell.2016.10.004
*Moè, A. (2018). Effects of group gender composition on mental rotation test performance in women. *Archives of Sexual Behavior, 47*(8), 2299–2305. https://doi.org/10.1007/s10508-018-1245-0
*Moè, A., Jansen, P., & Pietsch, S. (2018). Childhood preference for spatial toys. Gender differences and relationships with mental rotation in STEM and non-STEM students. *Learning and Individual Differences, 68*, 108–115. https://doi.org/10.1016/j.lindif.2018.10.003
*Moè, A., Meneghetti, C., & Cadinu, M. (2009). Women and mental rotation: Incremental theory and spatial strategy use enhance performance. *Personality and Individual Differences, 46*(2), 187–191. https://doi.org/10.1016/j.paid.2008.09.030
*Moffat, S. D., & Hampson, E. (1996). A curvilinear relationship between testosterone and spatial cognition in humans: Possible influence of hand preference. *Psychoneuroendocrinology, 21*(3), 323–337. https://doi.org/10.1016/0306-4530(95)00051-8
Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., Stewart, L. A., & Prisma-P Group. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews*, *4*, 1-9https://doi.org/10.1186/2046-4053-4-1
*Moody, M. S. (1997). Changes in scores on the mental rotations test during the menstrual cycle. *Perceptual and Motor Skills, 84*(3), 955–961. https://doi.org/10.2466/pms.1997.84.3.955
*Moreau, D., Clerc, J., Mansy-Dannay, A., & Guerrien, A. (2012). Enhancing spatial ability through sport practice. Evidence for an effect of motor training on mental rotation performance. *Journal of Individual Differences, 33*(2), 83–88. https://doi.org/10.1027/1614-0001/a000075
Nathan, M., & Lee, N. (2013). Cultural diversity, innovation, and entrepreneurship: Firm-level evidence from London. *Economic Geography, 89*(4), 367–394. https://doi.org/10.1111/ecge.12016
Neisser, U., Boodoo, G., Bouchard, T. J., Jr., Boykin, A. W., Brody, N., Ceci, S. J., Halpern, D. F., Loehlin, J. C., Perloff, R., Sternberg, R. J., & Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51*(2), 77. https://doi.org/10.1037/0003-066x.51.2.77
*Nguyen, H. D., O'Neal, A., & Ryan, A. M. (2003). Relating test-taking attitudes and skills and stereotype threat effects to the racial gap in cognitive ability test performance. *Human Performance, 16*(3), 261–293. https://doi.org/10.1207/s15327043hup1603_5
*Nyborg, H., & Jensen, A. R. (2000). Black-white differences on various psychometric tests: Spearman's hypothesis tested on American armed services veterans. *Personality and Individual Differences, 28*(3), 593–599. https://doi.org/10.1016/s0191-8869(99)00122-1
*O'Brien, L. T., & Crandall, C. S. (2003). Stereotype threat and arousal: Effects on women's math performance. *Personality and Social Psychology Bulletin, 29*(6), 782–789. https://doi.org/10.1177/0146167203029006010
Ones, D. S., Dilchert, S., Viswesvaran, C., & Salgado, J. F. (2017). Cognitive ability: Measurement and validity for employee selection. In J. L. Farr, & N.T. Tippins (Eds.), *Handbook of Employee Selection* (pp. 251–276). Routledge. https://doi.org/10.4324/9781315690193-11
*Ortner, T. M., & Sieverding, M. (2008). Where are the gender differences? Male priming boosts spatial skills in women. *Sex Roles, 59*(3–4), 274–281. https://doi.org/10.1007/s11199-008-9448-9
Osborne, R. T (1980). Twins: Black and White. In *Foundation for Human Understanding*.
Pace, R., Pluye, P., Bartlett, G., Macaulay, A. C., Salsberg, J., Jagosh, J., & Seller, R. (2012). Testing the reliability and efficiency of the pilot mixed methods appraisal tool (MMAT) for systematic mixed studies review. *International Journal of Nursing Studies, 49*(1), 47–53. https://doi.org/10.1016/j.ijnurstu.2011.07.002
*Palumbo, M. V., & Steele-Johnson, D. (2014). Do Test Perceptions Influence Test Performance? Exploring Stereotype Threat Theory. *North American Journal of Psychology*, *16*(1), 1–12. https://doi.org/10.32469/10355/44294
*Peña, D., Contreras, M. J., Shih, P. C., & Santacreu, J. (2008). Solution strategies as possible explanations of individual and sex differences in a dynamic spatial task. *Acta Psychologica, 128*(1), 1–14. https://doi.org/10.1016/j.actpsy.2007.09.005
*Penner, A. M., & Willer, R. (2011). Stigma and glucose levels: Testing ego depletion and arousal explanations of stereotype threat effects. *Current Research in Social Psychology, 16*(3), n3.
*Pennington, C. R., & Heim, D. (2016). Creating a critical mass eliminates the effects of stereotype threat on women's mathematical performance. *British Journal of Educational Psychology, 86*(3), 353–368. https://doi.org/10.1111/bjep.12110

Phillips, K., & Silverman, I. (1997). Differences in the relationship of menstrual cycle phase to spatial performance on two- and three-dimensional tasks. *Hormones and Behavior, 32*(3), 167–175. https://doi.org/10.1006/hbeh.1997.1418

*Picucci, L., Caffò, A. O., & Bosco, A. (2011). Besides navigation accuracy: Gender differences in strategy selection and level of spatial confidence. *Journal of Environmental Psychology, 31*(4), 430–438. https://doi.org/10.1016/j.jenvp.2011.01.005

*Pletzer, B., Steinbeisser, J., Van Laak, L., & Harris, T. (2019). Beyond biological sex: Interactive effects of gender role and sex hormones on spatial abilities. *Frontiers in Neuroscience, 13*, 675. https://doi.org/10.3389/fnins.2019.00675

*Ployhart, R. E., & Ehrhart, M. G. (2002). Modeling the practical effects of applicant reactions: Subgroup differences in test-taking motivation, test performance, and selection rates. *International Journal of Selection and Assessment, 10*(4), 258–270. https://doi.org/10.1111/1468-2389.00216

Ployhart, R. E., & Holtz, B. C. (2008). The diversity–validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology, 61*(1), 153–172. https://doi.org/10.1111/j.1744-6570.2008.00109.x

Ployhart, R. E., Ziegert, J. C., & McFarland, L. A. (2003). Understanding racial differences on cognitive ability tests in selection contexts: An integration of stereotype threat and applicant reactions research. *Human Performance, 16*(3), 231–259. https://doi.org/10.1207/s15327043hup1603_4

Popay, J., Roberts, H., Sowden, A., Petticrew, M., Arai, L., Rodgers, M., Britten, N., Roen, K., & Duffy, S. (2006). Guidance on the conduct of narrative synthesis in systematic reviews. *A Product from the ESRC Methods Programme Version, 1*(1), b92.

*Puts, D. A., Cárdenas, R. A., Bailey, D. H., Burriss, R. P., Jordan, C. L., & Breedlove, S. M. (2010). Salivary testosterone does not predict mental rotation performance in men or women. *Hormones and Behavior, 58*(2), 282–289. https://doi.org/10.1016/j.yhbeh.2010.03.005

*Quinn, D. M., & Spencer, S. J. (2001). The interference of stereotype threat with women's generation of mathematical problem-solving strategies. *Journal of Social Issues, 57*(1), 55–71. https://doi.org/10.1111/0022-4537.00201

*Rahe, M., & Jansen, P. (2022). Sex differences in mental rotation: The role of stereotyped material, perceived performance and extrinsic spatial ability. *Journal of Cognitive Psychology, 34*(3), 400–409. https://doi.org/10.1080/20445911.2021.2011896

*Rahe, M., Ruthsatz, V., & Quaiser-Pohl, C. (2020). Influence of the stimulus material on gender differences in a mental-rotation test. *Psychological Research Psychologische Forschung, 85*(8), 2892–2899. https://doi.org/10.1007/s00426-020-01450-w

*Razani, J., Murcia, G., Tabares, J., & Wong, J. (2007). The effects of culture on WASI test performance in ethnically diverse individuals. *The Clinical Neuropsychologist, 21*(5), 776–788. https://doi.org/10.1080/13854040701437481

*Reeb, M. (1976). Differential test validity for ethnic groups in the Israel army and the effects of educational level. *Journal of Applied Psychology, 61*(3), 253–261. https://doi.org/10.1037/0021-9010.61.3.253

*Reeve, C. L., Bonaccio, S., & Charles, J. E. (2008). A policy-capturing study of the contextual antecedents of test anxiety. *Personality and Individual Differences, 45*(3), 243–248. https://doi.org/10.1016/j.paid.2008.04.006

*Rilea, S. L. (2008). Sex and hemisphere differences when mentally rotating meaningful and meaningless stimuli. *Laterality, 13*(3), 217–233. https://doi.org/10.1080/13576500701809846

*Roberts, H. E., & Skinner, J. (1996). Gender and racial equity of the Air Force Officer Qualifying Test in officer training school selection decisions. *Military Psychology, 8*(2), 95–113. https://doi.org/10.1207/s15327876mp0802_4

*Roberts, J., & Bell, M. (2000). Sex differences on a computerized mental rotation task disappear with computer familiarization. *Perceptual and Motor Skills, 91*(3), 1027–1034. https://doi.org/10.2466/pms.2000.91.3f.1027

*Rosenberg, L., & Park, S. (2002). Verbal and spatial functions across the menstrual cycle in healthy young women. *Psychoneuroendocrinology, 27*(7), 835–841. https://doi.org/10.1016/s0306-4530(01)00083-x

Roth, P. L., Bevier, C. A., Bobko, P., Switzer, F. S., III., & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology, 54*(2), 297–330. https://doi.org/10.1111/j.1744-6570.2001.tb00094.x

*Roth, P. L., Le, H., Oh, I. S., Van Iddekinge, C. H., Buster, M. A., Robbins, S. B., & Campion, M. A. (2014). Differential validity for cognitive ability tests in employment and educational settings: Not much more than range restriction? *Journal of Applied Psychology, 99*(1), 1–20. https://doi.org/10.1037/a0034377

*Rotundo, M., & Sackett, P. R. (1999). Effect of rater race on conclusions regarding differential prediction in cognitive ability tests. *Journal of Applied Psychology, 84*(5), 815–822. https://doi.org/10.1037/0021-9010.84.5.815

*Rushton, J. P., & Skuy, M. (2000). Performance on Raven's matrices by African and White university students in South Africa. *Intelligence, 28*(4), 251–265. https://doi.org/10.1016/s0160-2896(00)00035-0

*Rushton, J. P., Bons, T. A., Vernon, P. A., & Cvorovic, J. (2007a). Genetic and environmental contributions to population group differences on the Raven's Progressive Matrices estimated from twins reared together and apart. *Royal Society b: Biological Sciences, 274*(1619), 1773–1777. https://doi.org/10.1098/rspb.2007.0461

*Rushton, J. P., Cvorovic, J., & Bons, T. A. (2007b). General mental ability in South Asians: Data from three Roma (Gypsy) communities in Serbia. *Intelligence, 35*(1), 1–12. https://doi.org/10.1016/j.intell.2006.09.002

*Rushton, J. P., Skuy, M., & Bons, T. A. (2004). Construct validity of Raven's advanced progressive matrices for African and non-African engineering students in South Africa. *International Journal of Selection and Assessment, 12*(3), 220–229. https://doi.org/10.1111/j.0965-075x.2004.00276.x

* Ryan, L. R., Brownlow, S., & Patterson, B. (2015). Women's Mental Rotation Abilities as a Function of Priming. *Psychology*, *6*(03), 217. https://doi.org/10.4236/psych.2015.63021

*Saccuzzo, D. P., Craig, A. S., Johnson, N. E., & Larson, G. E. (1996). Gender differences in dynamic spatial abilities. *Personality and Individual Differences, 21*(4), 599–607. https://doi.org/10.1016/0191-8869(96)00090-6

*Sackett, P. R., Zhang, C., & Berry, C. M. (2023). Challenging conclusions about predictive bias against Hispanic test takers in personnel selection. *Journal of Applied Psychology, 108*(2), 341–349. https://doi.org/10.1037/apl0000978

Sackett, P. R., Zhang, C., Berry, C. M., & Lievens, F. (2022). Revisiting meta-analytic estimates of validity in personnel selection: Addressing systematic overcorrection for restriction of range. *Journal of Applied Psychology, 107*(11), 2040. https://doi.org/10.1037/apl0000994

*Sanchis-Segura, C., Aguirre, N., Cruz-Gómez, Á. J., Solozano, N., & Forn, C. (2018). Do gender-related stereotypes affect spatial performance? Exploring when, how and to whom using a chronometric two-choice mental rotation task. *Frontiers in Psychology, 9*, 1261–1261. https://doi.org/10.3389/fpsyg.2018.01261

*Scherbaum, C. A., Blanshetyn, V., Marshall-Wolp, E., McCue, E., & Strauss, R. (2011). Examining the effects of stereotype threat on test-taking behaviors. *Social Psychology of Education, 14*(3), 361–375. https://doi.org/10.1007/s11218-011-9154-2

*Schmader, T., & Johns, M. (2003). Converging evidence that stereotype threat reduces working memory capacity. *Journal of Personality and Social Psychology, 85*(3), 440–452. https://doi.org/10.1037/0022-3514.85.3.440

*Schmader, T. (2002). Gender identification moderates stereotype threat effects on women's math performance. *Journal of Experimental Social Psychology, 38*(2), 194–201. https://doi.org/10.1006/jesp.2001.1500

*Schmader, T., Forbes, C. E., Zhang, S., & Mendes, W. B. (2009). A metacognitive perspective on the cognitive deficits experienced in intellectually threatening environments. *Personality and Social Psychology Bulletin, 35*(5), 584–596. https://doi.org/10.1177/0146167208330450

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*(2), 262–274. https://doi.org/10.1037/0033-2909.124.2.262

*Schöning, S., Engelien, A., Kugel, H., Schäfer, S., Schiffbauer, H., Zwitserlood, P., Pletziger, E., Beizai, P., Kersting, A., Ohrmann, P., Greb, R. R., Lehmann, W., Heindel, W., Arolt, V., & Konrad, C. (2007). Functional anatomy of visuo-spatial working memory during mental rotation is influenced by sex, menstrual cycle, and sex steroid hormones. *Neuropsychologia, 45*(14), 3203–3214. https://doi.org/10.1016/j.neuropsychologia.2007.06.011

*Schuster, C., Martiny, S. E., & Schmader, T. (2015). Distracted by the unthought-suppression and reappraisal of mind wandering under stereotype threat. *PLoS One, 10*(3), e0122207. https://doi.org/10.1371/journal.pone.0122207

*Shapiro, J. R., Williams, A. M., & Hambarchyan, M. (2013). Are all interventions created equal? A multi-threat approach to tailoring stereotype threat interventions. *Journal of Personality and Social Psychology, 104*(2), 277–288. https://doi.org/10.1037/a0030461

*Sharps, M. J., Price, J. L., & Williams, J. K. (1994). Spatial cognition and gender: Instructional and stimulus influences on mental image rotation performance. *Psychology of Women Quarterly, 18*(3), 413–425. https://doi.org/10.1111/j.1471-6402.1994.tb00464.x

*Sharps, M. J., Welton, A. L., & Price, J. L. (1993). Gender and task in the determination of spatial cognitive performance. *Psychology of Women Quarterly, 17*(1), 71–83. https://doi.org/10.1111/j.1471-6402.1993.tb00677.x

Shen, W., Kiger, T. B., Davies, S. E., Rasch, R. L., Simon, K. M., & Ones, D. S. (2011). Samples in applied psychology: Over a decade of research in review. *Journal of Applied Psychology, 96*(5), 1055–1064. https://doi.org/10.1037/a0023322

Shewach, O. R., Sackett, P. R., & Quint, S. (2019). Stereotype threat effects in settings with features likely versus unlikely in operational test settings: A meta-analysis. *Journal of Applied Psychology, 104*(12), 1514.

*Shih, M., Ambady, N., Richeson, J. A., Fujita, K., & Gray, H. M. (2002). Stereotype performance boosts: The impact of self-relevance and the manner of stereotype activation. *Journal of Personality and Social Psychology, 83*(3), 638–647. https://doi.org/10.1037/0022-3514.83.3.638

*Shih, M., Pittinsky, T. L., & Ambady, N. (1999). Stereotype susceptibility: Identity salience and shifts in quantitative performance. *Psychological Science, 10*(1), 80–83. https://doi.org/10.1111/1467-9280.00111

*Shute, V. J., Pellegrino, J. W., Hubert, L., & Reynolds, R. W. (1983). The relationship between androgen levels and human spatial abilities. *Bulletin of the Psychonomic Society, 21*(6), 465–468. https://doi.org/10.3758/bf03330010

Silverman, I., Choi, J., & Peters, M. (2007). The hunter-gatherer theory of sex differences in spatial abilities: Data from 40 countries. *Archives of Sexual Behavior, 36*(2), 261–268. https://doi.org/10.1007/s10508-006-9168-6

Silverman, I., & Phillips, K. (1993). Effects of estrogen changes during the menstrual cycle on spatial performance. *Ethology and Sociobiology, 14*(4), 257–269. https://doi.org/10.1016/0162-3095(93)90021-9

Silverman, I., Kastuk, D., Choi, J., & Phillips, K. (1999). Testosterone levels and spatial ability in men. *Psychoneuroendocrinology, 24*(8), 813–822. https://doi.org/10.1016/s0306-4530(99)00031-1

Šimić, N., & Santini, M. (2012). Verbal and spatial functions during different phases of the menstrual cycle. *Psychiatria Danubina, 24*(1), 73–79. https://doi.org/10.2478/10004-1254-61-2010-2055

Skuy, M., Gewer, A., Osrin, Y., Khunou, D., Fridjhon, P., & Rushton, J. P. (2002). Effects of mediated learning experience on Raven's matrices scores of African and non-African university students in South Africa. *Intelligence, 30*(3), 221–232. https://doi.org/10.1016/s0160-2896(01)00085-x

Smith, J. L., & White, P. H. (2002). An examination of implicitly activated, explicitly activated, and nullified stereotypes on mathematical performance: It's not just a woman's issue. *Sex Roles, 47*(3/4), 179–191. https://doi.org/10.1023/a:1021051223441

Sokolowski, H. M., Hawes, Z., & Lyons, I. M. (2019). What explains sex differences in math anxiety? A closer look at the role of spatial processing. *Cognition, 182*, 193–212. https://doi.org/10.1016/j.cognition.2018.10.005

Spearman, C. E. (1904). "General intelligence", objectively determined and measured. *The American Journal of Psychology, 15*(2), 201–293. https://doi.org/10.2307/1412107

Spearman, C. E. (1927). *The Abilities of Man* (6th ed.). Macmillan.

*Spencer, B., & Castano, E. (2007). Social class is dead. Long live social class! Stereotype threat among low socioeconomic status individuals. *Social Justice Research, 20*(4), 418–432. https://doi.org/10.1007/s11211-007-0047-7

Spencer, S. J., Logel, C., & Davies, P. G. (2016). Stereotype threat. *Annual Review of Psychology, 67*, 415–437. https://doi.org/10.1146/annurev-psych-073115-103235

*Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology, 35*(1), 4–28. https://doi.org/10.1006/jesp.1998.1373

*Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69*(5), 797–811. https://doi.org/10.1037/0022-3514.69.5.797

*Stenlund, T., Eklöf, H., & Lyrén, P. E. (2017). Group differences in test-taking behaviour: An example from a high-stakes testing program. *Assessment in Education: Principles, Policy & Practice, 24*(1), 4–20. https://doi.org/10.1080/0969594x.2016.1142935

Suzuki, L. A., & Valencia, R. R. (1997). Race–ethnicity and measured intelligence: Educational implications. *American Psychologist, 52*(10), 1103. https://doi.org/10.1037/0003-066X.52.10.1103

*Tanzer, N. K., Gittler, G., & Ellis, B. B. (1995). Cross-cultural validation of item complexity in a LLTM-calibrated spatial ability test. *European Journal of Psychological Assessment, 11*(3), 170–183. https://doi.org/10.1027/1015-5759.11.3.170

*te Nijenhuis, J., & van der Flier, H. (1997). Comparability of GATB scores for immigrants and majority group members: Some Dutch findings. *Journal of Applied Psychology, 82*(5), 675–687. https://doi.org/10.1037/0021-9010.82.5.675

*Te Nijenhuis, J., & Van der Flier, H. (2000). Differential prediction of immigrant versus majority group training performance using cognitive ability and personality measures. *International Journal of Selection and Assessment, 8*(2), 54–60. https://doi.org/10.1111/1468-2389.00133

*Terlecki, M. S., & Newcombe, N. S. (2005). How important is the digital divide? The relation of computer and videogame usage to gender differences in mental rotation ability. *Sex Roles, 53*, 433–441. https://doi.org/10.1007/s11199-005-6765-0

*Terlecki, M. S., Newcombe, N. S., & Little, M. (2008). Durable and generalized effects of spatial experience on mental rotation: Gender differences in growth patterns. *Applied Cognitive Psychology, 22*(7), 996–1013. https://doi.org/10.1002/acp.1420

*Thames, A. D., Panos, S. E., Arentoft, A., Byrd, D. A., Hinkin, C. H., & Arbid, N. (2015). Mild test anxiety influences neurocognitive performance among African Americans and European Americans: Identifying interfering and facilitating sources. *Cultural Diversity and Ethnic Minority Psychology, 21*(1), 105–113. https://doi.org/10.1037/a0037530

*Thilers, P. P., Macdonald, S. W., & Herlitz, A. (2006). The association between endogenous free testosterone and cognitive performance: A population-based study in 35 to 90 year-old men and women. *Psychoneuroendocrinology, 31*(5), 565–576. https://doi.org/10.1016/j.psyneuen.2005.12.005

*Thomsen, T., Hugdahl, K., Ersland, L., Barndon, R., Lundervold, A., & Smievoll, A. I. (2000). Functional magnetic resonance imaging (fMRI) study of sex differences in a mental rotation task. *Medical Science Monitor, 6*(6), 1186–1196. https://doi.org/10.1016/s1053-8119(00)91315-6

Thurstone, L. L. (1938). *Primary mental abilities*. University of Chicago Press.

*Tine, M., & Gotlieb, R. (2013). Gender-, race-, and income-based stereotype threat: The effects of multiple stigmatized aspects of identity on math performance and working memory function. *Social Psychology of Education, 16*(3), 353–376. https://doi.org/10.1007/s11218-013-9224-8

*Toivainen, T., Pannini, G., Papageorgiou, K. A., Malanchini, M., Rimfeld, K., Shakeshaft, N., & Kovas, Y. (2018). Prenatal testosterone does not explain sex differences in spatial ability. *Scientific Reports, 8*(1), 13653–13661. https://doi.org/10.1038/s41598-018-31704-y

*Unterrainer, J., Wranek, U., Staffen, W., Gruber, T., & Ladurner, G. (2000). Lateralized cognitive visuospatial processing: Is it primarily gender-related or due to quality of performance? A HMPAO-SPECT study. *Neuropsychobiology, 41*(2), 95–101. https://doi.org/10.1159/000026639

*Vanderpool, M., & Catano, V. M. (2008). Comparing the performance of Native North Americans and predominantly White military recruits on verbal and nonverbal measures of cognitive ability. *International Journal of Selection and Assessment, 16*(3), 239–248. https://doi.org/10.1111/j.1468-2389.2008.00430.x

*Vasta, R., Knott, J. A., & Gaze, C. E. (1996). Can spatial train- ing erase the gender differences on the water-level task? *Psychology of Women Quarterly, 20*(4), 549–567. https://doi.org/10.1111/j.1471-6402.1996.tb00321.x

*Vernon, P. A., & Jensen, A. R. (1984). Individual and group differences in intelligence and speed of information processing. *Personality and Individual Differences, 5*(4), 411–423. https://doi.org/10.1016/0191-8869(84)90006-0

*Voyer, D., & Jansen, P. (2016). Sex differences in chronometric mental rotation with human bodies. *Psychological Research Psychologische Forschung, 80*(6), 974–984. https://doi.org/10.1007/s00426-015-0701-x

*Vuoksimaa, E., Kaprio, J., Eriksson, C. J., & Rose, R. J. (2012). Pubertal testosterone predicts mental rotation performance of young adult males. *Psychoneuroendocrinology, 37*(11), 1791–1800. https://doi.org/10.1016/j.psyneuen.2012.03.013

*Vuoksimaa, E., Kaprio, J., Kremen, W. S., Hokkanen, L., Viken, R. J., Tuulio-Henriksson, A., & Rose, R. J. (2010). Having a male co-twin masculinizes mental rotation performance in females. *Psychological Science, 21*(8), 1069–1071. https://doi.org/10.1177/0956797610376075

*Walsh, M., Hickey, C., & Duffy, J. (1999). Influence of item content and stereotype situation on gender differences in mathematical problem solving. *Sex Roles, 41*(3–4), 219–240. https://doi.org/10.1023/a:1018854212358

Warne, R. T. (2021). Between-group mean differences in intelligence in the United States are> 0% genetically caused: Five converging lines of evidence. *The American Journal of Psychology, 134*(4), 480–501. https://doi.org/10.5406/amerjpsyc.134.4.0479

*Wee, S., Newman, D. A., & Joseph, D. L. (2014). More than g: Selection quality and adverse impact implications of considering second-stratum cognitive abilities. *Journal of Applied Psychology, 99*(4), 547–563. https://doi.org/10.1037/a0035183

*Weger, U. W., Hooper, N., Meier, B. P., & Hopthrow, T. (2012). Mindful maths: Reducing the impact of stereo- type threat through a mindfulness exercise. *Consciousness and Cognition, 21*(1), 471–475. https://doi.org/10.1016/j.concog.2011.10.011

*Weiss, E., Siedentopf, C. M., Hofer, A., Diesenhammer, E. A., Hoptman, M. J., Kremser, C., et al. (2003). Sex differences in brain activation pattern during a visuospatial cognitive task: A functional magnetic resonance imaging study in healthy volunteers. *Neuroscience Letters, 344*(3), 169–172. https://doi.org/10.1016/s0304-3940(03)00406-3

*Wicherts, J. M. (2017). Psychometric problems with the method of correlated vectors applied to item scores (including some nonsensical results). *Intelligence, 60*, 26–38. https://doi.org/10.1016/j.intell.2016.11.002

*Wicherts, J. M., Dolan, C. V., & Hessen, D. J. (2005). Stereotype threat and group differences in test performance: A question of measurement invariance. *Journal of Personality and Social Psychology, 89*(5), 696–716. https://doi.org/10.1037/0022-3514.89.5.696

*Wister, J. A., Stubbs, M. L., & Shipman, C. (2013). Mentioning menstruation: A stereotype threat that diminishes cognition? *Sex Roles, 68*(1–2), 19–31. https://doi.org/10.1007/s11199-012-0156-0

Woods, S. A., & Patterson, F. (2024). A critical review of the use of cognitive ability testing for selection into graduate and higher professional occupations. *Journal of Occupational and Organizational Psychology, 97*(1), 253–272. https://doi.org/10.1111/joop.12470

*Wraga, M., Duncan, L., Jacobs, E. C., Helt, M., & Church, J. (2006). Stereotype susceptibility narrows the gender gap in imagined self-rotation performance. *Psychonomic Bulletin & Review, 13*(5), 813–819. https://doi.org/10.3758/bf03194002

*Wraga, M., Helt, M., Jacobs, E., & Sullivan, K. (2007). Neural basis of stereotype-induced shifts in women's mental rotation performance. *Social Cognitive and Affective Neuroscience, 2*(1), 12–19. https://doi.org/10.1093/scan/nsl041

*Yang, C. F. J., Hooven, C. K., Boynes, M., Gray, P. B., & Pope, H. G., Jr. (2007). Testosterone levels and mental rotation performance in Chinese men. *Hormones and Behavior, 51*(3), 373–378. https://doi.org/10.1016/j.yhbeh.2006.12.005

*Zoref, L., & Williams, P. (1980). A look at content bias in IQ tests. *Journal of Educational Measurement, 17*(4), 313–322. https://doi.org/10.1111/j.1745-3984.1980.tb00834.x