



# The Impasse on Gender Differences in Intelligence: a Meta-Analysis on WISC Batteries

D. Giofrè<sup>1</sup> · K. Allen<sup>2</sup> · E. Toffalini<sup>3</sup> · S. Caviola<sup>4,5</sup>

Accepted: 1 September 2022 / Published online: 22 September 2022  
© The Author(s) 2022

## Abstract

This meta-analysis reviews 79 studies ( $N=46,605$ ) that examined the existence of gender difference on intelligence in school-aged children. To do so, we limited the literature search to works that assessed the construct of intelligence through the Wechsler Intelligence Scales for Children (WISC) batteries, evaluating eventual gender differences in indices and subtests. The theoretical framework we adopted is the cross-battery approach which locates cognitive abilities into different levels, also considering the possible mediating effect of the version of the WISC being used. As for broad abilities, a notable discrepancy emerged in favour of males for visual and crystallized intelligence, while female/male differences on fluid intelligence were negligible. Conversely, females' performance on the processing speed factor was superior. Interesting results emerged at the subtest levels, albeit with less pronounced differences in performance. Results generally showed that older versions of WISC batteries displayed larger gender differences compared to the most recent ones.

**Keywords** Intelligence · Gender differences · Sex differences · WISC · Performance

Research in psychology has consistently tried to explain and understand the impact of gender<sup>1</sup> differences (e.g. Geary, 2021). Most empirical findings on this topic can be explained by different (but interrelated) theoretical approaches that capture much of the current frameworks relating to gender differences: the socio-cultural,

<sup>1</sup> Some authors prefer to use sex rather than gender when talking about female/male differences. However, we decided to use “gender” rather than “sex” because the latter tends to be used more in the biological vs. social–psychological literature.

✉ D. Giofrè  
david.giofre@gmail.com

<sup>1</sup> Disfor, University of Genoa, Genoa, Italy

<sup>2</sup> Department of Psychology, University of Durham, Durham, UK

<sup>3</sup> Department of General Psychology, University of Padua, Padua, Italy

<sup>4</sup> Department of Developmental and Social Psychology, University of Padova, Padua, Italy

<sup>5</sup> School of Psychology, University of Leeds, Leeds, UK

evolutionary, hormone and brain science approaches (e.g. Archer, 2019; Eagly & Wood, 2013). Most of these theories acknowledge the existence of both biological (*nature*) factors (e.g. physical differences, evolved traits, hormonal influences) and socio-cultural (*nurture*) factors (e.g. social and cultural role learning, and stereotyped beliefs). Looking at differences in cognitive tasks linked to gender, the socio-cultural theories propose that gender differences arise from social, cultural, psychological, and other environmental factors (Wood & Eagly, 2012).

The long-lasting debate on how gender may, or may not, impact intelligence has been addressed from several angles. While findings have consistently shown gender differences in certain cognitive domains such as verbal and spatial abilities (Halpern et al., 2007), a consensus has yet to be reached regarding the existence of gender differences in overall cognitive ability or general intelligence. As Hunt (2011) nicely summarizes, research reported evidence from three types of sources: (a) studies based on single tests (e.g. Raven Matrices, Cattell Culture Fair Intelligence Test; Colom, & García-López, 2002); (b) *g*-factor type studies, derived from a variety of test batteries (e.g. Deary, Strand, Smith, & Fernandez, 2007); and finally (c) research considering overall scores, such as IQ score derived from the Weschler Scales (see also Mackintosh, 2011 for a review). As for evidence derived from tests considered markers for intelligence, inconsistent results have been reported, often attributed to the non-representativeness of the data samples and other methodological issues (Savage-McGlynn, 2012), such as the use of different assessments (i.e. psychometric tasks vs. standardized IQ tasks, Arribas-Aguila, et al., 2019), or materials (i.e. verbal vs. spatial) (see Hunt, 2011 for a discussion). Colom and García-López (2002) tested more than 4,000 high-school students with the PMA Inductive Reasoning Test, the Advanced Progressive Matrices (APM), and the Culture-Fair Intelligence Test. They concluded that because no systematic difference was found in the measures of fluid intelligence, gender differences in general intelligence are likely non-existent (see also Waschl et al., 2020). Similar results emerged from studies considering different test batteries indicating that gender differences are very small, if not trivial (Hajovsky et al., 2018; Keith et al., 2008; Strand et al., 2006). Most studies concluded there is a male superiority (e.g. Flores-Mendoza et al., 2013; Jackson & Rushton, 2006; Lynn & Irwing, 2004), while some reported that females outperform males on the *g*-factor (e.g. Arden & Plomin, 2006; Reynolds et al., 2008).

Several explanations have been proposed to explain these scattered results. One took into consideration the effect of age: according to the developmental theory, gender differences in intelligence can vary between age groups and it is linked to differences in the progression of maturity between the two genders (Lynn, 1999). Although a few studies indicate a significant interaction between age and gender (e.g. Arden & Plomin, 2006; Lynn & Kanazawa, 2011), other researchers reported findings that were inconsistent with this developmental theory (e.g. Keith et al., 2008; Savage-McGlynn, 2012). Another explanation considers the variability of the scores (variability hypothesis). Most studies have only considered gender differences in mean scores, without considering possible differences in score variability between genders (Maccoby & Jacklin, 1974). Gender discrepancies in the spread of scores are essential to consider as they may contribute to explaining why males/females may outnumber their peers among the highest (or lowest) scoring

individuals in tests that show only a small advantage in the mean score (Feingold, 1992; Nowell & Hedges, 1998).

As mentioned above, research in this area has used the Wechsler Intelligence Scales to test for the possible presence of differences in general intellectual functioning and in other abilities (Van der Sluis, et al., 2006, 2008). Most of these studies were performed on a single standardization sample, being therefore limited to a single country, or conducted with a limited sample size thus leading to different psychometric approaches in data analyses (see Giofrè et al., 2022 for a discussion). In addition, previous studies were mainly focused on adults, while child participants were often neglected (see Chen et al., 2015; Chen et al. 2020; Dombrowski et al., 2021; Pezzuti & Orsini, 2016). We decided to address these issues by performing a meta-analysis of all evidence available in children assessed via the Wechsler Intelligence Scales for Children (WISC). This approach allowed us to have a very large sample and to obtain more robust estimates of the female/male difference in intelligence, and more broadly in factors and subtests.

The WISC is an individually administered intelligence test for children. The first version of the scale appeared in 1949, encompassing several subtests explicitly designed to assess children's intelligence in typical and atypical development (Wechsler, 1949). The first battery, however, presented with some limitations and a revised version of the scale was introduced in 1974 (Wechsler, 1974). Several refinements of the scale have been published since, including the WISC-III (Wechsler, 1991), the WISC-IV (Wechsler, 1991), and more recently the WISC-V (Wechsler, 2014). The WISC has also been adapted and standardized for use in several countries and populations and is probably the most widely used tool for assessing intelligence in children (Evers et al., 2012).

From a historical perspective, the WISC, as compared to other important intelligence batteries, presented with several advantages. One of the advantages of the WISC is its atheoretical approach. In addition, the WISC has been designed to be used easily by practitioners and to capture difficulties experienced by children in school settings (Flanagan & Kaufman, 2004), resulting in it being instrumental in driving educational practice and intervention. This approach remained fairly stable across different standardizations of the scale-up until the WISC-IV. However, with the emergence of the so-called CHC theory (named after Cattell, Horn, and Carroll), and of the so-called cross-battery approach, which postulates that the results of different batteries can be compared based on the CHC theory, important changes to the battery have been introduced (Flanagan et al., 2007).

Following the CHC approach, tasks can be distinguished based on their loadings on the respective CHC factors. A series of studies performed on WISC scales demonstrated that the CHC is a useful framework to explain the structure of the Wechsler scales (e.g. Bowden, 2013). The CHC model divides cognitive abilities into different levels: on top of the CHC hierarchy is the g-factor, several narrow abilities are at the base level, while several broad abilities sit in the middle level, e.g. fluid intelligence (gF), visual processing (gV), crystallized intelligence (gC), short-term memory (gSM), speed (gS), and quantitative knowledge (gQ), among others (e.g. Flanagan et al., 2013). In fact, several studies, using different versions of the WISC, showed that the CHC structure explained Wechsler scales better as compared

to other traditional factorial structures (e.g. Golay et al., 2013; McGill & Canivez, 2017; Scheiber, 2016; Weiss et al., 2013). This approach can be particularly useful when comparing female/male differences on cognitive abilities assessed through the Wechsler scales.

The first mention of female/male differences on the WISC appeared as early as the 1950s, shortly after the publication of the first WISC scale, with results indicating some small differences in full-scale IQ (FSIQ) in the older age groups (Seashore et al., 1950). Evidence that emerged later was mixed, with some claiming small differences favouring males, and others claiming small differences favouring females (Darley & Winitz, 1961; Mercer & Smith, 1972). The urgency of having a new and more robust version of the WISC, culminated in the standardization of the WISC-R.

The WISC was widely used particularly in Anglophone countries, but with the WISC-R the battery was adapted to several other countries around the world. The original USA standardization sample of the WISC-R was stratified for age and gender, thus making it possible to test for the presence of gender differences, however, these differences were considered trivial and not of a practical relevance (Kaufman & Doppelt, 1976). Several subsequent investigations were performed with results showing the presence of some female/male differences (e.g. males outperforming females on some tasks, such as Block Design, while females performing better on other tasks, such as Coding) in several countries, the USA, Scotland, Netherlands, China, Mauritius, New Zealand, Belgium, and Greece, among others (Al-Shahomee et al., 2016; Alexopoulos, 1979; Born & Lynn, 1994; Dai & Lynn, 1994; Jensen & Reynolds, 1983; Lynn et al., 2005; Lynn & Mulhern, 1991; van der Sluis et al., 2008). It was probably with the emergence of these investigations that the scientific debate on the possible presence of female/male differences in intelligence became harshly contested. With some authors claiming that differences were negligible, while others took the opposite view.

The introduction of the WISC-III led to several refinements in the WISC scale. The verbal/performance factor structure was replaced by a three-factor solution, also including a freedom from distractibility factor, and some newly created subtests were introduced. Using this scale, the presence of female/male differences were investigated in several different samples and countries, e.g. USA, China, Taiwan, Sudan, and Brazil among others (Bakhiet et al., 2017; Chen et al., 2016; Flores-Mendoza et al., 2007). In these studies, females consistently outperformed males on some subtests such as Coding, while males outperformed females on some others, such as Block Design. However, a males' advantage on verbal task was found in some countries (e.g. Chen et al., 2016; Flores-Mendoza et al., 2007), but not in others (e.g. Bakhiet et al., 2017).

More recently, the introduction of the WISC-IV led to further important changes to its internal structure. Several newly developed subtests were included while some other old subtests were removed. The WISC-IV, included four different factors, including a working memory index and a processing speed index, which replaced the freedom from distractibility factor. The scale also included some new subtests explicitly introduced to measure fluid intelligence, such as matrix reasoning. As previously mentioned, the new scale is closely aligned with the CHC model. Several studies indicated that the CHC factor solution was preferable as compared to the

traditional four-factor solution (e.g. Weiss et al., 2013). Finally, the most recent version of the WISC (WISC-V), embraced the CHC model and introduced the possibility of directly estimated broader factors, including for example a fluid intelligence (gF) and a visual processing (gV) factor. As for female/male differences, those were evaluated on several different samples and countries including, but not limited to, Italy, Germany, Romania, and Spain (Goldbeck et al., 2010; Hernández et al., 2017; Iliescu et al., 2016; Pezzuti & Orsini, 2016). Intriguingly, results showed either no or trivial female/male differences in most countries.

Based on these premises, we aimed to investigate female/male differences in cognitive abilities reflected by the WISC's indices and subtests. As for the indices, we were particularly interested in the FSIQ. Moreover, we decided to embrace the CHC theory, using subtests as indicators of the underlying broad factors, for example, tasks traditionally used for measuring gC, such as vocabulary and information, were used as indicators of this broad factor. In a similar vein, tasks measuring gF, gV, gSM, and gS were used as indicators of their underlying factor. It is worth noting, however, that we decided not to calculate a gQ factor. This decision was based on the observation that only one test was measuring this factor (arithmetic), making the introduction of a gQ factor unnecessary.

To achieve this aim, we used a meta-analytic approach. We collected all the available evidence on female/male differences on the WISC. We wanted, to evaluate whether differences on indices and on subtests exist and, if so, to quantify these differences. In addition, we wanted to evaluate the role of the version of the WISC being used. The WISC and WISC-R are very old scales, while WISC-III, WISC-IV, and WISC-V are relatively newer. We had several hypotheses. We did not expect to find any strong differences in the FSIQ, or in gF, while we expected to find differences on some broad factors: a female advantage on speed-related tasks and indices (e.g. coding and gS) and on verbal working memory tasks (e.g. digit span and gSM), and some male advantage on visuospatial tasks (e.g. block design, and gV). As for the gC factor, recent pieces of evidence indicate some male advantage on these tasks, thus we also expected to find some differences on this broad factor (Lynn, 2021). Since several recent studies claimed there has been a reduction in the female/male differences in intelligence we also expected to find that the Wechsler scale being used had an effect, with newer versions of the scale showing a decrease in the differences between groups.

## Method

### Literature Search and Inclusion Criteria

The literature search was conducted on three online databases: PsycINFO, PubMed, and Scopus. The search terms were: (“WISC” OR “Wechsler Intelligence Scale for Children\*”) AND (“Sex difference\*” OR “Gender difference\*”). The wildcard \* was used to include any possible variations of these terms. Results were exported and merged in a single spreadsheet, including source database, authors, year of publication, title, abstract, and publication details (i.e. volume, issue, pages, doi). A total

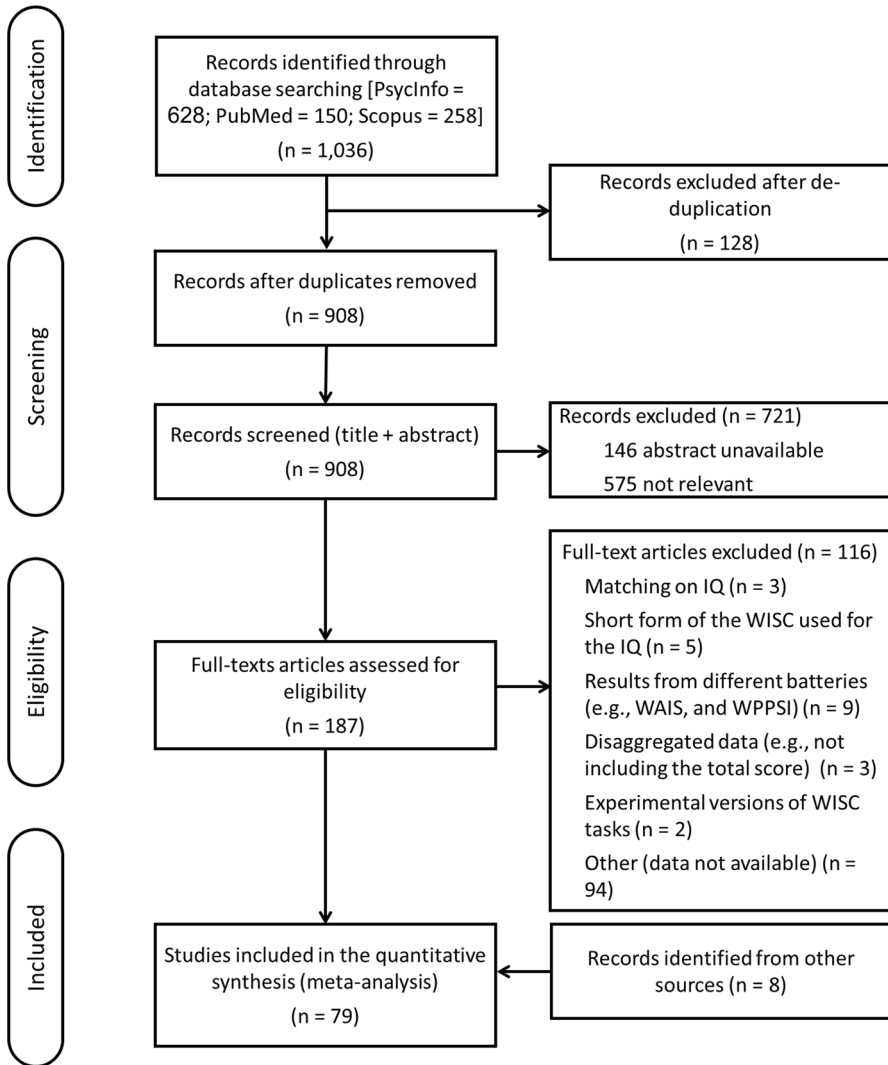


Fig. 1 Prisma flowchart

of 1,036 records were found. After removing duplicates, 908 records remained in the spreadsheet, of which 146 were excluded because the abstract could not be retrieved even after further web searches and emails to the authors. Records had publication dates between 1967 and 2021. The search was conducted on the 14th of July 2021. An additional 8 articles found via other sources were added. Further details are provided in Fig. 1.

Records were screened by title and abstract. In this phase, any record that clearly did not report primary research data (e.g. reviews), did not use any version of the WISC, or clearly collected data only on atypical/clinical samples was excluded

(Fig. 1). A total of 187 records passed to the subsequent full-text eligibility phase. Full-text papers were searched through the academic library facilities of the affiliation institutions of all authors involved in the present study. If a paper could not be found via academic facilities, an additional web search was conducted, and the authors were contacted.

Inclusion criteria were (a) papers written in English or in any other language understood by the authors of the present review (i.e. Italian, French, Spanish, or Portuguese); (b) papers reporting primary research data collected using at least one index or one subtest of the WISC; (c) participants including at least one subsample of typically developing children<sup>2</sup>; (d) papers reporting sample size and means, distinguishing between males and females; if descriptive statistics were not reported, Cohen's *d* or any other measure of the effect size with enough information to estimate its variance (e.g. sample size, standard error), was considered; (e) males and females being compared had a similar age or were in the same age range; (f) males and females being compared were not matched on any measure of intelligence (e.g. FSIQ). Before excluding a full-text paper based on criteria (d) and (e), authors were contacted via email for additional information (1 month was given to reply to the email).

The flow chart below provides more details on the literature search procedure (Fig. 1). The preferred reporting items for systematic reviews and meta-analyses (PRISMA) guidelines were followed in this meta-analysis. Two independent reviewers evaluated 20% of the papers. The agreement between the two reviewers was high: 96.77% for the screening phase, and 98.92% for the eligibility phase, and disagreements were resolved by discussion.

## Coding of the Studies, Samples, and Effects

Since our primary goal was to conduct a meta-analysis, data were coded accordingly. The spreadsheet was organized so that each row corresponded to an effect size. As several effect sizes could be calculated for each study in most cases (e.g. because a study reported data for many subtests and/or indices for one or many samples of children), there were more rows than studies (i.e. 79 studies and 640 effects). We avoided coding any redundant information: effect sizes were always calculated from subtests, except when only scores for indices were available. The only exception was the FSIQ, which was always coded if reported. Nonetheless, the latter was examined in a separate analysis. Finally, to prevent violating the independence of observations, data only from the first time point were coded in longitudinal studies. Effect sizes were calculated using the formulae for Cohen's *d* and its variance reported by Borenstein et al. (2009).

For each row in the coding spreadsheet, information about the study, the sample, and the effect size was entered. Information about the study included a label with authors and year of publication. Information about the sample included country, block/continent (with the following groups: Africa, East-Asia, East-Europe, Middle

<sup>2</sup> Children with any neurodevelopmental disorder, for example children with ADHD, autism, or specific learning disabilities were not included.

East, South-America, Western-Countries), version of the battery (WISC, WISC-R, WISC-III, WISC-IV, WISC-V), whether the sample was used for the battery standardization (of any country), number of females and males, age range and mean age of the sample (in years). If a study included data from subsamples of different ages, the latter were coded separately instead of using the combined sample. Information about the effect size also included mean and standard deviation separately by males and females, whether the score was related to an index or a subtest, the name of the index or subtest, and, if possible, the CHC/WISC-V classification. Concerning the latter, the following levels were coded: g-factor (FSIQ scores only); gF/fluid reasoning (subtests: Matrix reasoning, Picture concepts; Fluid Reasoning index from WISC-V); gV/visual processing (subtests: Block design, Mazes, Object assembly, Picture completion; VSI index from WISC-V); gC/verbal comprehension (subtests: Comprehension, Information, Similarities, Vocabulary; VCI index from WISC-IV/V); gSM/working memory (subtests: Digit span, Letter Number Sequencing; WMI index from WISC-IV/V); gS/speed of processing (subtests: Coding, Symbol search; PSI index from WISC-IV/V). More information about the analytic approach being used for multiple effects on the same sample, or for multiple samples within the same study is provided below.

## Analytic Strategy

### Meta-Analytic Model Fitting and Heterogeneity

The R free software (R Core Team, 2021) was used for all the analyses. Meta-analytic models were fitted using the multilevel random-effects models function implemented in the “metafor” (Viechtbauer, 2010) package.

We adopted the analytic strategy proposed by Borenstein et al. (2009). Random-effects models were used for all analyses to account for the presumable heterogeneity across the effect sizes. To account for the structure of dependence between the effects in the dataset, multilevel models were used. They allow us to model effects coming from a structure in which multiple samples (i.e. subgroups of different ages) are nested within studies. Studies and samples were treated as random effects. For simplicity, multiple effect sizes nested within samples were combined before modelling rather than adding an additional level in the models. Effects were combined using the formulae suggested by Borenstein et al., (2009, pp. 227–229), assuming a correlation of 0.7 among effect sizes within the same sample (alternative values between 0.5 and 0.9 had negligible impact on the results). In addition, once again for simplicity, separated meta-analytic models were fitted for different factors of intelligence or different subtests, rather than using multivariate models.

Heterogeneity across independent samples was quantified using  $\tau$ , which is a measure of estimated standard error across the true effect sizes, and by using the  $I^2$  index, which can be interpreted as the percentage of total variance that is attributable to the variance across true effects (Higgins et al., 2003). High values of  $I^2$  (i.e.  $I^2 > 75\%$ ) suggest that most variance across the observed effect sizes can be attributed to differences in the true effect sizes, indicating a substantial role of moderators.



Finally, to test the robustness of each estimated effect size, and the leverage that each single study had on it, we used the *leave-one-out* method. It consists of a sensitivity analysis in which the estimated result is re-calculated after removing one study at a time. This allows us to detect outliers that are influential cases in a meta-analysis (Viechtbauer & Cheung, 2010).

## Moderator Analyses

The version of the WISC (dichotomized into old [WISC and WISC-R] vs. new [WISC-III/IV/V]) and mean age of children (where it could be coded) were tested as the moderators of interest. Regarding mean age, we decided to test it as a moderator only for samples whose age range was not wider than 5 years (e.g. samples including children throughout most, or all, age spans covered by the Wechsler scales were not considered for this moderation analysis). Moderating analyses were conducted on multilevel models using meta-regression, and only for the analysis by CHC factors (for subtests the results are largely redundant).

Additionally, we examined whether effect sizes estimated from standardization samples differed from estimates including all data (i.e. convenience samples). We did not test this formally with a moderation analysis, however, because we did not assume that standardization samples might differ from convenience samples, but only that they might be more accurate. Thus, we re-computed the estimated average effect sizes on standardization samples alone.

## Assessment of Risk of Publication Bias

Assessing the publication bias was difficult for at least three reasons. First, even though the amount of heterogeneity was modest in absolute terms (see below), it was large relative to the small effect sizes that we investigated. In other words, the estimated standard deviation,  $\tau$ , among true effect sizes was often similar to, or even larger than, the effect size,  $d$ . Heterogeneity is generally an issue that limits the reliability of conventional meta-analytic approaches to assess publication bias (Stanley, 2017). Second, for standardized mean differences, the effect size and its variance are not independent (Borenstein et al., 2009), which creates an asymmetry in the distribution of the effect sizes in the funnel plot that increases the risk of false positives with any method based on the relationship between effect size and its variance/precision (e.g. trim-and-fill, regression tests; Zwetsloot et al., 2017). Third, we dealt with a complex multilevel data structure (samples nested within studies), which is generally an issue when adopting a conventional approach to assess publication bias (Rodgers & Pustejovsky, 2021).

To tackle the above issues, we choose to adopt the PET-PEESE meta-regression approach on adequately adapted data. The PET-PEESE method fits a regression model in which standard error (or, if the estimated bias-free effect size remains significant, variance) is entered as a moderator of the effect size, with the intercept being interpreted as the bias-free effect (Haaf, 2021). Although it risks performing poorly in presence of high heterogeneity, PET-PEESE is comparatively more robust than other conventional approaches (Stanley, 2017). In addition, regression

tests (PET-PEESE is a modification of Egger's test; Stanley & Doucouliagos, 2014) can be used within the multi-level meta-analytic framework (Rodgers & Pustejovsky, 2021). Finally, to address the issue of effect size being correlated with its variance for standardized mean differences, we adopted the suggestion by Zwetsloot et al. (2017) of using  $1/n$  as an alternate variance estimate (independent from  $d$ ) in the meta-regression test. Even so, we urge taking the results of the publication bias assessment with caution due to the expectably large heterogeneity (Stanley, 2017). Besides the issues discussed above, it should be noted that publication bias is not necessarily the only cause of asymmetry in the distribution of effect sizes in the funnel plot (e.g. Borenstein et al., 2009). For example, PET-PEESE meta-regressions that lead to estimates even larger than the original effect sizes cannot be attributed to a publication bias and are probably due to heterogeneity alone.

## Results

### Overview of the Studies

A total of 79 studies, published between 1961 and 2019, encompassing 134 distinct independent samples, and 640 effect sizes were included. The estimated total number of participants was 46,605 (23,404 males and 23,201 females). Sixty-seven studies were conducted in Western Countries, 3 in Africa, 3 in East Asia, 2 in East-Europe, 2 in the Middle East, and 2 in South America. Forty-two studies used old versions of the WISC ( $k=11$  for WISC;  $k=31$  for WISC-R), while the remaining 37 used new versions ( $k=15$  for WISC-III;  $k=21$  for WISC-IV;  $k=1$  for WISC-V). Fourteen studies used data from a standardization sample.

### Female/Male Differences on CHC broad factors

A synthesis of the main meta-analytic estimates reported in this section can be found in Table 1. Additional analyses (i.e. funnel plots, leave-one-out tests, and forest plots) were also performed and are included in the Supplemental material (Figures S1-S18).

### Full-Scale IQ

Forty-eight studies, encompassing 78 samples, included data for the FSIQ. The estimated mean effect size was significant but practically negligible, with slightly higher scores in males than in females,  $d=0.09$ , 95%CI [0.04, 0.14],  $p<0.001$ . Heterogeneity was relatively modest in absolute terms, but it appeared relevant when compared with the very small effect size,  $\tau=0.13$ ,  $I^2=55.97$  (i.e. the estimated standard error across true effect sizes is larger than the mean effect size). The leave-one-out range of variation was modest: [0.08, 0.11] (see plot in Supplemental materials). The mean effect size calculated on standardization samples alone (11 studies, 35 samples) was virtually the same  $d=0.10$  [0.03, 0.17],  $\tau=0.10$ ,  $I^2=52.96$ .

**Table 1** Summary of the main meta-analytic results of gender differences by CHC factor

Factor	<i>n</i> of studies	<i>n</i> of samples	<i>n</i> of effects	total <i>n</i> of males	total <i>n</i> of females	<i>d</i>	95% CI	Leave-one-out range	Equival. IQ scores	<i>p</i>	tau	Only from WISC-III/IV/V	
												Cohen's <i>d</i>	95% CI
g/FSIQ	48	78	78	13,709	13,696	0.093	[0.040, 0.145]	[0.081, 0.145]	1.395	.001	0.130	0.054	[− 0.015, 0.123]
gF	21	32	34	12,372	12,456	0.093	[0.038, 0.147]	[0.077, 0.107]	1.395	.001	0.103	0.052	[− 0.029, 0.133]
gV	26	39	98	11,049	11,207	0.206	[0.159, 0.254]	[0.181, 0.214]	3.090	<.001	0.092	0.183	[0.135, 0.231]
gC	41	72	142	13,489	13,740	0.137	[0.081, 0.192]	[0.127, 0.146]	2.055	<.001	0.142	0.136	[0.044, 0.229]
gSM	28	48	49	10,887	11,080	− 0.042	[− 0.091, 0.007]	[− 0.055, − 0.032]	− 0.630	.091	0.097	− 0.019	[− 0.084, 0.046]
gS	34	53	55	13,239	13,289	− 0.366	[− 0.436, − 0.296]	[− 0.380, − 0.347]	− 5.490	<.001	0.177	− 0.264	[− 0.337, − 0.190]

Positive values indicate males' advantage, while negative values indicate females' superiority

*gF*, fluid intelligence; *gV*, visual processing; *gC*, crystallized intelligence; *gSM*, short-term memory; *gS*, speed

The old/new version of the battery failed to reach significance as a moderator,  $\chi^2(1) = 1.86$ ,  $p = 0.17$ , and the effect size estimated from samples using new versions of the WISC was virtually zero,  $d = 0.05$   $[-0.02, 0.13]$ ,  $p = 0.13$ . Mean age of children could be coded for 33 studies, 60 samples, and was not a significant moderator:  $\chi^2(1) = 1.86$ ,  $p = 0.17$ ,  $B = 0.01$   $[-0.01, 0.02]$ . The PET-PEESE method did not suggest any publication bias, as the meta-regressions with both standard error and variance provided effect size estimates that were even larger than the original estimate ( $d = 0.14$  and  $0.11$ , respectively); as explained above, however, this is likely due to heterogeneity alone.

### gF/Fluid reasoning

Twenty-one studies, encompassing 32 samples, included data for this factor. The estimated mean effect size was significant, but negligible, with slightly higher scores in males than in females,  $d = 0.09$   $[0.04, 0.15]$ ,  $p < 0.001$ . Heterogeneity was modest in absolute terms,  $\tau = 0.10$ , albeit it accounted for a large portion of variance in the observed effect size,  $I^2 = 70.09$ . The leave-one-out range of variation was modest:  $[0.08, 0.11]$  (see plot in Supplemental materials). The mean effect size calculated on standardization samples alone (11 studies, 16 samples) was very similar  $d = 0.06$   $[0.03, 0.10]$ ,  $\tau = 0.06$ ,  $I^2 = 52.16$ .

The old/new version of the battery was not a statistically significant moderator,  $\chi^2(1) = 2.22$ ,  $p = 0.13$ . However, the effect size estimated only from new versions of the WISC was virtually zero,  $d = 0.05$   $[-0.03, 0.12]$ . Mean age of children could be coded for 7 studies, 13 samples, and was a significant and positive moderator:  $\chi^2(1) = 9.06$ ,  $p = 0.003$ ,  $B = 0.06$   $[0.03, 0.10]$ . The PET-PEESE meta-regression suggested no significant moderating effect of standard error,  $B = 0.30$   $[-2.61, 3.21]$ , and the estimated bias-free effect remained close to zero, but with larger uncertainty,  $d = 0.06$   $[-0.17, 0.29]$ .

### gV/Visual Spatial

Twenty-six studies, encompassing 39 samples, included data for this factor. The estimated mean effect size was significant but small, with higher scores in males than in females,  $d = 0.21$   $[0.16, 0.25]$ ,  $p < 0.001$ . Heterogeneity was modest,  $\tau = 0.09$ ,  $I^2 = 66.28$ . The leave-one-out range of variation was modest:  $[0.18, 0.21]$  (see plot in Supplemental materials). The mean effect size calculated on standardization samples alone (11 studies, 16 samples) was very similar to the original one,  $d = 0.17$   $[0.14, 0.20]$ ,  $\tau = 0.03$ ,  $I^2 = 26.41$ .

The old/new version of the battery did not appear to be a moderator of the effect size,  $\chi^2(1) = 0.55$ ,  $p = 0.46$ ,  $\Delta d = 0.04$ . Mean age of children could be coded for 12 studies, 20 samples, and was not a significant moderator:  $\chi^2(1) = 2.46$ ,  $p = 0.06$ ,  $B = 0.03$   $[-0.00, 0.07]$ . The PET-PEESE meta-regression suggested no significant moderating effect of standard error,  $B = -0.57$   $[-2.62, 1.49]$ , and an estimated bias-free effect even larger than the original estimate,  $d = 0.31$   $[0.11, 0.50]$ ; as in the previous case, this is likely due to heterogeneity alone.

### gC/Verbal Comprehension

Forty-one studies, encompassing 72 samples, included data for this factor. The estimated mean effect size was significant but small, with higher scores in males than in females,  $d=0.14$  [0.08, 0.19],  $p<0.001$ . Heterogeneity was modest,  $\tau=0.14$ ,  $I^2=61.20$ . The leave-one-out range of variation was modest: [0.13, 0.15] (see plot in Supplemental materials). The mean effect size calculated on standardization samples alone (12 studies, 28 samples) was virtually the same,  $d=0.13$  [0.06, 0.20],  $\tau=0.11$ ,  $I^2=74.01$ .

The old/new version of the battery did not appear to be a moderator of the effect size,  $\chi^2(1)=0.05$ ,  $p=0.82$ ,  $\Delta d=0.02$ . Mean age of children could be coded for 26 studies, 52 samples, and was not a significant moderator:  $\chi^2(1)=1.54$ ,  $p=0.21$ ,  $B=0.01$  [−0.01, 0.03]. The PET-PEESE meta-regression coefficient for standard error was not significant,  $B=1.01$  [−0.51, 2.53], although the estimated bias-free effect was very close to null,  $d=0.07$  [−0.09, 0.23].

### gSM/Working Memory

Twenty-eight studies, encompassing 48 samples, included data for this factor. The estimated mean effect size was near zero,  $d=-0.04$  [−0.09, 0.01],  $p=0.09$ . Heterogeneity was modest,  $\tau=0.10$ ,  $I^2=54.76$ . The leave-one-out range of variation was modest: [−0.06, −0.03] (see plot in Supplemental materials). The mean effect size calculated on standardization samples alone (10 studies, 24 samples) was the same,  $d=-0.04$  (−0.10, 0.03),  $\tau=0.10$ ,  $I^2=65.05$ .

The old/new version of the battery did not appear to be a moderator of the effect size,  $\chi^2(1)=1.61$ ,  $p=0.20$ ,  $\Delta d=0.06$ . Mean age of children could be coded for 16 studies, 31 samples, and was not a significant moderator:  $\chi^2(1)=1.51$ ,  $p=0.22$ ,  $B=0.01$  [−0.01, 0.04]. Since the effect was already virtually zero, the publication bias assessment was not conducted.

### gS/Speed of Processing

Thirty-four studies, encompassing 53 samples, included data for this factor. The estimated mean effect size was significant and medium, with higher scores in females than in males,  $d=-0.37$  [−0.44, −0.30],  $p<0.001$ . Heterogeneity was substantial,  $\tau=0.18$ ,  $I^2=79.42$ . The leave-one-out range of variation was modest: [−0.38, −0.35] (see plot in Supplemental materials). The mean effect size calculated on standardization samples alone (13 studies, 29 samples) was very similar,  $d=-0.30$  [−0.39, −0.21],  $\tau=0.16$ ,  $I^2=79.64$ .

The old/new version of the battery was a significant moderator of the effect size,  $\chi^2(1)=9.32$ ,  $p=0.002$ ; the effect was larger when estimated with older versions of the battery,  $d=-0.46$  [−0.56, −0.36] than with newer versions,  $d=-0.26$  [−0.34, −0.19]. Mean age of children could be coded for 20 studies, 34 samples, and was not a significant moderator:  $\chi^2(1)=0.41$ ,  $p=0.52$ ,  $B=-0.01$  [−0.03, 0.02]. The PET-PEESE meta-regression suggested no significant moderating effect of standard error,  $B=-0.19$  [−2.89, 2.51], and

the estimated bias-free effect remained very similar to the original estimate,  $d = -0.40 [-0.58, -0.21]$ .

### Female/Male Differences on Subtests

A summary of the main results by subtest is reported in Table 2. As the results of the analyses by subtest largely reflect the results by CHC factors reported in the previous section, here we note only the main discrepancies. No publication bias assessment was conducted for subtests since the analysis already conducted for the CHC factors (previous section) suggested that this was not an issue.

Concerning the gF factor subtests, all effects were very modest to negligible. The largest concerned Picture arrangement, with an estimated mean effect size of  $d = 0.13 [0.08, 0.12]$ . As far as Matrix reasoning is concerned, a subtest tapping heavily on core aspects of fluid reasoning, and very similar to Raven's matrices, the effect, favouring females, is virtually zero:  $d = -0.04 [-0.10, 0.02]$ .

Concerning the gV factor subtests, there were no notable discrepancies across subtests, with estimated mean effect sizes close to  $d \approx 0.20$  for all of them, always in favour of males over females (Table 2).

Concerning the gC factor subtests, a notable discrepancy emerged between Information, which presented the largest estimated difference in favour of males,  $d = 0.26 [0.17, 0.36]$ , and the other subtests, which presented very small to negligible estimated effect sizes (see Table 2).

Concerning the gSM factor subtests, both effect sizes were negligible. Nonetheless, the digit span reached the conventional level of significance,  $d = -0.06 [-0.11, -0.01]$ ,  $p = 0.02$ , with higher scores in females than in males.

Concerning the gS factor subtests, both effect sizes were significantly in favour of females over males, but the estimated mean effect size was substantially larger for Coding,  $d = -0.42 [-0.49, -0.34]$ , than for Symbol search,  $d = -0.15 [-0.24, -0.06]$ .

As for the Arithmetic reasoning subtest, which was a single indicator tapping on gQ. The estimated mean effect size was significant but modest, with higher scores in males than in females,  $d = 0.13 [0.06, 0.20]$ , which remained practically the same when only the newer versions of the WISC battery were considered. As for most of the other subtests, the estimated heterogeneity was modest,  $\tau = 0.11$ .

It should be noted that for most subtests, the estimated mean effect sizes calculated using only the scores from the newer versions of the WISC, were generally smaller or they remained the same, but they were never larger, vis-à-vis the estimated effects using all scores. This is consistent with the effect sizes aggregated by the CHC factor (Table 1), all of which were closer to zero when calculated only from the newer versions of the WISC as compared to the overall estimates.

**Table 2** Summary of the main meta-analytic results of gender differences by subtest

Subtest	CHC	<i>n</i> of studies	<i>n</i> of samples/effects	total <i>n</i> of males	total <i>n</i> of females	<i>d</i>	95% CI	Leave-one-out range	<i>p</i>	tau	Only from WISC-III/IV/V	
											<i>d</i>	95%CI
Matrix Reasoning	gF	5	6	4,744	4,656	-0.041	[-0.102, 0.021]	[-0.062, -0.009]	0.194	0.000	-0.041	[-0.102, 0.021]
Picture Concepts	gF	2	2	1,925	1,925	0.018	[-0.088, 0.124]	[-0.033, 0.075]	0.740	0.043	0.018	[-0.088, 0.124]
Picture Arrang	gF	16	26	7,628	7,800	0.130	[0.079, 0.181]	[0.116, 0.142]	<.001	0.026	0.139	[0.047, 0.231]
Block Design	gV	26	39	11,049	11,207	0.206	[0.162, 0.249]	[0.186, 0.213]	<.001	0.071	0.196	[0.150, 0.241]
Object Assembly	gV	17	25	8,769	8,940	0.212	[0.132, 0.292]	[0.186, 0.226]	<.001	0.132	0.189	[0.069, 0.309]
Picture Completion	gV	17	25	8,604	8,754	0.177	[0.111, 0.243]	[0.150, 0.197]	<.001	0.11	0.124	[0.030, 0.218]
Mazes	gV	6	9	4,945	5,144	0.239	[0.151, 0.327]	[0.199, 0.261]	<.001	0.000	0.213	[0.105, 0.322]
Vocabulary	gC	28	46	10,656	10,863	0.138	[0.073, 0.203]	[0.127, 0.149]	<.001	0.122	0.124	[0.023, 0.225]
Similarities	gC	20	27	10,484	10,650	0.033	[-0.010, 0.076]	[0.025, 0.056]	0.136	0.000	0.023	[-0.080, 0.126]
Comprehension	gC	19	28	9,344	9,509	0.096	[0.033, 0.159]	[0.063, 0.114]	0.003	0.107	0.045	[0.006, 0.085]
Information	gC	18	29	7,803	7,962	0.263	[0.168, 0.358]	[0.239, 0.288]	<.001	0.175	0.155	[-0.005, 0.315]
DigitSpan	gSM	19	28	9,532	9,762	-0.059	[-0.107, -0.011]	[-0.071, -0.049]	0.017	0.069	-0.032	[-0.103, 0.039]
Letts-Num Seq	gSM	2	2	1,925	1,925	0.025	[-0.038, 0.088]	[0.000, 0.059]	0.434	0.000	0.025	[-0.038, 0.088]
Symbol Search	gS	5	6	4,604	4,778	-0.149	[-0.235, -0.064]	[-0.193, -0.123]	0.001	0.078	-0.149	[-0.235, -0.064]
Coding	gS	24	32	11,376	11,471	-0.419	[-0.492, -0.345]	[-0.433, -0.398]	<.001	0.117	-0.333	[-0.425, -0.241]
Arithmetic Reas	gQ	17	24	7,922	8,105	0.129	[0.060, 0.198]	[0.104, 0.143]	<.001	0.113	0.138	[0.087, 0.189]

Positive values indicate males' advantage, while negative values indicate females' superiority  
*gF*, fluid intelligence; *gV*, visual processing; *gC*, crystallized intelligence; *gSM*, short-term memory; *gS*, speed

## Discussion

Over the past years, differences between females and males in terms of neuropsychological and cognitive functioning have been extensively documented (Halpern & Wai, 2019; Miller & Halpern, 2014). Meta-analyses have shown gender differences favouring males on some specific cognitive abilities, such as spatial abilities (Moffat et al., 1998; Nazareth, et al., 2019), mental rotation (Maeda & Yoon, 2013), mathematics, and science achievement (Reilly, et al., 2015). Conversely, differences have also been found in favour of females who often outperform males in verbal tasks, such as in reading and writing achievements (Petersen, 2018; Reilly, et al., 2019). In addition, other aspects of human cognition, such as memory, processing speed, and intelligence, have been investigated with a particular focus on gender-related differences. Among the latter, intelligence is the construct that provoked, and still does, a considerable volume of research trying to answer the questions about whether, and in which abilities, females and males differ (Johnson, et al., 2008).

The main aim of the present meta-analysis was to investigate the presence of gender differences in intelligence on the WISC. We estimated standardized differences on the FSIQ, and on tasks and indices tapping on some broader factors derived from the CHC model. We also evaluated the moderating effect of the WISC version, in which results from the WISC and WISC-R, were compared with the results obtained from the WISC-III, WISC-IV, and WISC-V.

Concerning the first aim, we investigated the effect on the FSIQ. We found that there is a male advantage which, albeit statistically significant, is negligible in terms of magnitude (equivalent to a difference of 1.395 IQ points) on the FSIQ. Intriguingly, when only the newer version of the WISC was included, the difference dropped further and became statistically non-significant; equivalent to a difference of 0.81 IQ points. In fact, differences in IQ seem to be attributable, at least in part, to the battery being used, becoming smaller with new batteries as compared to the older ones. To achieve a more precise understanding of this phenomenon is useful to investigate differences at the level of the broad CHC factors.

Following a cross battery approach, tasks were distinguished based on their belonging to specific CHC broader factors. To understand differences in the FSIQ, it is useful to start with differences in tasks tapping on fluid intelligence, gF, which is the most closely related factor to general intelligence and therefore to the FSIQ (e.g. Giofrè & Cornoldi, 2015; Keith et al., 2006; Weiss et al., 2013). Results for the gF factor are very similar to the FSIQ, with a very small and not statically significant difference favouring males with the newer version of the WISC. This finding seems to indicate that it is unlikely that there are any female/male differences in fluid intelligence tasks. Further, investigation of the results at the subtests level, indicated that the larger difference was on the Picture Arrangement subtest. However, newer tasks tapping on fluid intelligence showed either no differences whatsoever (on picture concepts) or differences favouring females (on matrix reasoning, which is the subtest with the highest loading on fluid intelligence). This finding is extremely interesting because it seems to indicate that



the FSIQ, which is where the debate originated in the first place (with the early versions of the WISC), probably does not reflect female/male differences in fluid intelligence. In fact, newly devised tasks, explicitly developed to load on fluid intelligence showed either no differences or differences favouring females.

As for other specific abilities, as expected, we found that males outperformed females in the gV factor. We identified a male advantage equivalent to a difference of 3.09 IQ points, which was somewhat lower using new batteries (2.75 IQ points). Looking at each individual subtest we noticed that differences were somewhat larger on some tasks, particularly those included in the older version of the WISC scales, such as mazes and object assembly, while somewhat lower on picture completion. Block design probably deserves a separate discussion. This task is one of the few tasks that have been maintained in the newer version of the WISC, making it possible to evaluate changes over time. Interestingly, differences in this task favouring males seem to be large and relatively unchanged between various versions of the WISC. The block design task is widely used in literature to assess visuospatial and visuo-constructive abilities (e.g. Cardillo et al., 2017, 2018, 2020). Visuo-constructive abilities are closely related to the maturation of other cognitive visuospatial skills, such as visuomotor coordination, perceptual abilities, mental rotation skills, and working memory (WM) (Lingo VanGilder et al., 2021; Morra & Panesi, 2017). These results align with evidence indicating a male advantage on tasks requiring some visuospatial transformation of the stimuli. The male advantage on tasks requiring generating and mentally manipulating images in memory (i.e. tasks requiring visuospatial working memory or mental imagery to a larger extent) has been shown repeatedly (Halpern & Wai, 2019; Hedges & Nowell, 1995; Loring-Meier & Halpern, 1999; Maeda & Yoon, 2013).

As for crystallized intelligence (gC), we found some statistically significant differences, albeit small in terms of the effect size, favouring males (equivalent to roughly 2 IQ points, independent of the version of the WISC being used). Looking at each specific subtest, we found that differences were larger on the information subtest. It is worth noting, however, that this task, which was a primary task (i.e. adding to the calculation of the FSIQ) in the older version of the WISC, is now a supplementary task, hardly being used, and not essential for the calculation of the FSIQ. As for the other subtests, some small, but statistically significant, differences were found in vocabulary and comprehension. Vice versa, in the similarities subtest, differences were not statistically significant and were small in terms of magnitude. These results are explored further below.

There is a host of research indicating a female advantage in verbal tasks. The female advantage is almost ubiquitous in reading, writing, and comprehension tasks (Hedges & Nowell, 1995; Reynolds et al., 2015), and this holds true across several different countries and realities (see Giofrè et al., 2020 for a discussion). However, the female advantage seems to be more pronounced in reading and writing tasks or in tasks that require rapid access to, and use of, phonological and semantic information from long-term memory (see Halpern & Wai, 2019 for a discussion). Tasks such as vocabulary or similarities do not necessarily require rapid access to the information in long-term memory, and this can, at least in part, explain why females do not necessarily outperform males on these tasks. The male advantage in these tasks, however,

seems to be particularly puzzling. It has been suggested that such a result could be due to a male advantage on general knowledge (crystallised intelligence) tasks (Lynn, 2021). Being that as it may, further research is needed on this issue. Notably, recent research using the WISC-IV found that nearly all WISC-IV scores conveyed significant and nontrivial amounts of variation and this was particularly true for the verbal comprehension index (McDermott et al., 2013). It is also true that verbal tasks are very hard to evaluate and some specific factors such as anxiety and other personality factors can play a role in assessing those tasks (e.g. vocabulary) (Babad et al., 1975; McDermott et al., 2013). On the other hand, it can also be argued that tasks such as vocabulary do not require an extensive verbal manipulation of the stimuli, and this might explain, at least in part, why females do not outperform males on this task. For these reasons, differences in verbal tasks are not necessarily related to differences in performance, but this hypothesis should be further investigated.

Concerning gSM, a female advantage on this factor was hypothesized. Females tend to outperform males in rapid access to information, and this might constitute an advantage on some working memory tasks (Halpern & Wai, 2019). We found that on average females were superior as compared to males on gSM tasks, however, this difference was trivial and not statistically significant (roughly  $-0.63$ , and  $-0.29$  IQ points, when considering the newer version of the WISC). It is worth noting, however, that tasks such as letter-number sequencing require the manipulation of numbers and letters to a certain extent, thus involving rapid access to the information of both sequences in long-term memory, which could in part explain this result (Halpern & LaMay, 2000).

Finally, for gS, the results are in line with several pieces of evidence indicating a female superiority on processing speed (Delaney et al., 1981; Roivainen, 2011). In this case, differences were larger and statistically significant (equivalent to a difference in IQ points of roughly  $-5.49$ , and  $-3.96$ , when considering the newer version of the WISC). Looking at the individual subtests, standardized differences on symbol search were modest, while differences on coding were much larger. This outcome is remarkably stable and robust over time: Camarata and Woodcock (2006) showed that males scored significantly lower than females across different school-cohorts in processing speed tasks. One possible explanation can be linked to the processing speed definition itself, for which males perform worse than females because of their limits in maintaining attention and concentration while performing a simple and repetitive task for an extended period (Flanagan et al., 2000; see also Giofrè et al., 2022 for a discussion). This latter finding is quite interesting since we noticed a female superiority in virtually all studies included in this meta-analysis with very few exceptions. This finding is not trivial since this task has been maintained in newer versions of the WISC-III, WISC-IV, and WISC-V. In addition, the WISC-V has reduced the number of subtests needed to calculate the IQ (only 7 subtests), and the presence of these subtests could potentially produce some biases in the estimation of the FSIQ in male participants. We think that this aspect is worth further investigation.

Arithmetic represents a special case. In this case, differences favour males and are statistically significant, albeit modest in terms of magnitude. It is worth mentioning that this task is probably very spurious and has loadings on several factors, including

working memory (in the WISC-IV for example arithmetic is included in the working memory factor). In any case, arithmetic is not included as a principal subtest in the newer version of the scale and is not needed for the calculation of the FSIQ.

Overall, our findings are in line with those using other types of well-known intelligence tests such as the Woodcock-Johnson Tests of Cognitive Abilities (e.g. Keith et al., 2008) or the Kaufman Assessment Battery for Children (e.g. Hajovsky et al., 2018; Reynolds et al., 2008). Most studies have reported that females outperformed males in the speed of processing (gS). Using the earlier forms of the Woodcock-Johnson battery, Camarata and Woodcock (2006) found that no gender difference was observed in overall estimates of general intellectual ability, but specific patterns could be observed in the context of broad cognitive abilities. They found that females scored significantly higher on the tests of gS while males performed better on estimates of verbal comprehension knowledge (gC). Similar findings of female superiority in gS have been reported by others (e.g. Burns & Nettelbeck, 2005; Keith et al., 2008; van der Sluis et al., 2006). Moreover, other researchers concluded in favour of a male advantage on the broad cognitive abilities of working memory (gSM), visual-spatial ability and visual processing (gV), and crystallized intelligence (gC) (Dolan et al., 2006; Keith et al., 2008; Reynolds et al., 2008; Strand, et al., 2006).

Despite the importance in terms of both theoretical and clinical implications, these results shall be considered in light of a few limitations that should be addressed in future research. The number of studies reporting disaggregated data for males and females was not particularly large, and in fact, some standardizations completely neglect to report those data. We believe that this is an important problem and that publishers should be hard-pressed to report and make available those data for the scientific community. Also, the WISC-V has only been recently standardized, and it is not available in several countries yet. It would therefore be interesting to repeat the meta-analysis in future years when this battery will eventually be accessible to more countries. Finally, most papers only report and discuss results on principal indexes, while it would be interesting to also evaluate differences on the other subtests as well.

The moderating effect of age was tested, and we found that, in general, the results were trivial and not statistically significant, with one exception. In one case (gF) the effect of age was statistically significant (differences increased with age). It is worth noting that the reliability of intelligence is related to age. That is, at early ages batteries are somewhat less reliable and this might have important implications, particularly when investigating age trends in the female/male context (McCall, 1977). For this reason, we decided to repeat the analysis of age as a moderator after removing a study with very young children (5.2 years old). Results showed that the effect of age as a moderator is not very stable as the removal of one further study (using the leave one out method) makes the effect negligible and not statistically significant. For these reasons, the results of age as a moderator can hardly be brought as supporting evidence for increasing differences in gF and we believe that they should be interpreted with extreme caution because a very limited number of studies provided data disaggregated for age and gender. Further, results were provided using a large age range (we used a 5-year age range interval criterion, which could also have impacted on the results). Finally, most of the data are based on older versions of the WISC, and we noticed that the female/male gap decreased when newer versions of the WISC were used. Before

making any conclusions, these results should be corroborated using a larger number of studies and samples and should only be taken as indicative in the meantime.

In this paper, we decided to follow the classical CHC model, because it is probably the most widely used and WISC tasks usually fit nicely with this model. However, it would also be interesting to test gender differences using different intelligence formulations. For example, the g-VPR (verbal-perceptual-rotation) model (introduced by Johnson & Bouchard, 2005) postulates the presence of differences not at the g level but only at the specific abilities level (Johnson & Bouchard, 2007). Our results closely align with this hypothesis, with males performing generally better on tasks requiring some spatial manipulation of the materials, and females outperforming males on speeded tasks. It should also be noted that the WISC is probably the most widely used intelligence scale, but other scales are also available. For this reason, we believe that it would be important that future meta-analyses focus on other intelligence scales, such as the Woodcock-Johnson or the Stanford-Binet, and on other age groups (for example, focusing on the adult population; see also Collaer & Hines, 1995). Finally, it is also important to note that, in the standardization process of newer batteries, items (or tasks) that are not invariant across groups are usually removed and this might have some impact on the estimation of female/male differences, for example, the information subtest. However, the same issue applies to other subtests including arithmetic. It has been argued that the intercept of some subtests (e.g. information) is not invariant across males and females, meaning that differences in these tasks probably do not reflect genuine group differences in this ability (Wicherts & Dolan, 2010). This alongside other observations (see also Wicherts, 2016 for a discussion), might also explain why some tasks are no longer included as principal subtests in newer versions of the WISC.

To sum up, we found that female/male differences on FSIQ and gF were negligible, and this is particularly true when considering the newer version of the WISC. Some differences favouring males were found on some tasks, on the gV and gC factors and subtests. On the other hand, females, tend to outperform males on gSM, but to a very small extent and only on digit span, and on gS, in this case to a large degree and in particular on the coding subtest, which is in absolute terms the task with the largest standardized differences, as compared to all WISC subtests. The presence of some tasks favouring males and others favouring females instead can probably explain why differences on FSIQ are trivial in terms of magnitude (Jensen, 1998; see pp. 531–532). In fact, despite performing similarly on the FSIQ and gF tasks, males and females do not perform similarly on other tasks, and this makes the investigation of female/male differences particularly interesting from a scientific perspective.

To conclude, our results are very close to those obtained by Arthur Jensen (1998; see pp. 531–532). When addressing the question of female/male differences in intelligence by analysing tests that load heavily on the g-factor, he concluded that there is no evidence was found for gender differences in the mean level of g or in the variability of g and that males, on average, excel on some factors; while females excel on others.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s10648-022-09705-1>.

**Funding** Open access funding provided by Università degli Studi di Genova within the CRUI-CARE Agreement.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Al-Shahomee, A. A., Abdalla, S.E.-G., & Lynn, R. (2016). Sex differences on the WISC-R in Libya. *Mankind Quarterly*, 57(1), 91–94. <https://doi.org/10.46469/mq.2016.57.1.12>
- Alexopoulos, D. (1979). *Revision and standardization of the Wechsler intelligence scale for children - revised (WISC-R) for the age range 13 - 15 years in Greece*.
- Arden, R., & Plomin, R. (2006). Sex differences in variance of intelligence across childhood. *Personality and Individual Differences*, 41(1), 39–48. <https://doi.org/10.1016/j.paid.2005.11.027>
- Arribas-Aguila, D., Abad, F. J., & Colom, R. (2019). Testing the developmental theory of sex differences in intelligence using latent modeling: Evidence from the TEA Ability Battery (BAT-7). *Personality and Individual Differences*, 138, 212–218. <https://doi.org/10.1016/j.paid.2018.09.043>
- Babad, E. Y., Mann, M., & Mar-Hayim, M. (1975). Bias in scoring the WISC subtests. *Journal of Consulting and Clinical Psychology*, 43(2), 268–268. <https://doi.org/10.1037/h0076368>
- Bakhiet, S. F., Alburasan, I. S., Al Qudah, M. F., Abduljabbar, A. S., Aljomaa, S. S., Toto, H. S. A., & Lynn, R. (2017). Sex differences on the wis-III among children in Sudan and the United States. *Journal of Biosocial Science*, 49(6), 792–797. <https://doi.org/10.1017/S0021932016000432>
- Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2009). *Introduction to meta-analysis*. John Wiley & Sons, Ltd.
- Born, M. P. H., & Lynn, R. (1994). Sex Differences on the Dutch WISC-R: A comparison with the USA and Scotland. *Educational Psychology*, 14(2), 249–254. <https://doi.org/10.1080/0144341940140208>
- Bowden, S. C. (2013). Theoretical convergence in assessment of cognition. *Journal of Psychoeducational Assessment*, 31(2), 148–156. <https://doi.org/10.1177/0734282913478035>
- Burns, N. R., & Nettelbeck, T. (2005). Inspection time and speed of processing: Sex differences on perceptual speed but not IT. *Personality and Individual Differences*, 39(2), 439–446. <https://doi.org/10.1016/j.paid.2005.01.022>
- Camarata, S., & Woodcock, R. (2006). Sex differences in processing speed: Developmental effects in males and females. *Intelligence*, 34(3), 231–252. <https://doi.org/10.1016/j.intell.2005.12.001>
- Cardillo, R., Mammarella, I. C., Garcia, R. B., & Cornoldi, C. (2017). Local and global processing in block design tasks in children with dyslexia or nonverbal learning disability. *Research in Developmental Disabilities*, 64, 96–107. <https://doi.org/10.1016/j.ridd.2017.03.011>
- Cardillo, R., Menazza, C., & Mammarella, I. C. (2018). Visuoconstructive abilities and visuospatial memory in autism spectrum disorder without intellectual disability: Is the role of local bias specific to the cognitive domain tested? *Neuropsychology*, 32(7), 822–834. <https://doi.org/10.1037/neu0000472>
- Cardillo, R., Lanfranchi, S., & Mammarella, I. C. (2020). A cross-task comparison on visuospatial processing in autism spectrum disorders. *Autism*, 24(3), 765–779. <https://doi.org/10.1177/1362361319888341>
- Collaer, M. L., & Hines, M. (1995). Human behavioral sex differences: A role for gonadal hormones during early development?. *Psychological Bulletin*, 118(1), 55–107. <https://doi.org/10.1037/0033-2909.118.1.55>
- Chen, H.-Y., Lynn, R., & Cheng, H. (2016). Sex Differences on the WISC-III in Taiwan and the United States. *Mankind Quarterly*, 57(1), 66–71. <https://doi.org/10.46469/mq.2016.57.1.7>

- Chen, H., Zhang, O., Raiford, S. E., Zhu, J., & Weiss, L. G. (2015). Factor invariance between genders on the Wechsler intelligence scale for children—fifth edition. *Personality and Individual Differences*, 86, 1–5. <https://doi.org/10.1016/j.paid.2015.05.020>
- Colom, R., & García-López, O. (2002). Sex differences in fluid intelligence among high school graduates. *Personality and Individual Differences*, 32(3), 445–451. [https://doi.org/10.1016/S0191-8869\(01\)00040-X](https://doi.org/10.1016/S0191-8869(01)00040-X)
- Dai, X. Y., & Lynn, R. (1994). Gender differences in intelligence among chinese children. *Journal of Social Psychology*, 134(1), 123–125. <https://doi.org/10.1080/00224545.1994.9710892>
- Darley, F. L., & Winitz, H. (1961). Comparison of male and female kindergarten children on the wisc. *Journal of Genetic Psychology*, 99(1), 41–49. <https://doi.org/10.1080/00221325.1961.10534389>
- Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence*, 35(1), 13–21. <https://doi.org/10.1016/j.intell.2006.02.001>
- Delaney, H. D., Norman, R. D., & Miller, D. A. (1981). An exploration of the verbal encodability hypothesis for sex differences in the digit-symbol (symbol-digit) test. *Intelligence*, 5(2), 199–208. [https://doi.org/10.1016/0160-2896\(81\)90009-X](https://doi.org/10.1016/0160-2896(81)90009-X)
- 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 modeling 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>
- Dombrowski, S. C., Watkins, M. W., McGill, R. J., Canivez, G. L., Holingue, C., Pritchard, A. E., & Jacobson, L. A. (2021). Measurement Invariance of the Wechsler Intelligence Scale for Children, 10-Subtest Primary Battery: Can Index Scores be Compared across Age, Sex, and Diagnostic Groups? *Journal of Psychoeducational Assessment*, 39(1), 89–99. <https://doi.org/10.1177/0734282920954583>
- Eagly, A. H., & Wood, W. (2013). The nature–nurture debates: 25 years of challenges in understanding the psychology of gender. *Perspectives on Psychological Science*, 8(3), 340–357. <https://doi.org/10.1177/1745691613484767>
- Evers, A., Muñoz, J., Bartram, D., Boben, D., Egeland, J., Fernández-Hermida, J. R., Frans, Ö., Gintiliéné, G., Hagemeister, C., Halama, P., Iliescu, D., Jaworowska, A., Jiménez, P., Manthouli, M., Matesic, K., Schittekatte, M., Sümer, H. C., & Urbánek, T. (2012). Testing practices in the 21st century. *European Psychologist*, 17(4), 300–319. <https://doi.org/10.1027/1016-9040/a000102>
- Feingold, A. (1992). Sex differences in variability in intellectual abilities: A new look at an old controversy. *Review of Educational Research*, 62(1), 61–84. <https://doi.org/10.3102/00346543062001061>
- Flanagan, D. P., Alfonso, V. C., & Reynolds, M. R. (2013). Broad and narrow CHC abilities measured and not measured by the Wechsler Scales: Moving beyond within-battery factor analysis. *Journal of Psychoeducational Assessment*, 31(2), 202–223. <https://doi.org/10.1177/0734282913478047>
- Flanagan, D. P., & Kaufman, A. S. (2004). *Essentials of assessment with WISC-IV*. Wiley.
- Flanagan, D. P., McGrew, K. S., & Ortiz, S. (2000). *The Wechsler Intelligence Scales and Gf-Gc theory: A contemporary approach to interpretation*. Allyn & Bacon.
- Flanagan, D. P., Ortiz, S., & Alfonso, V. C. (2007). Use of the cross-battery approach in the assessment of diverse individual. In A. S. Kaufman & N. L. Kaufman (Eds.), *Essentials of cross-battery assessment second edition* (3rd ed., pp. 146–205). Wiley.
- Flores-Mendoza, C. E., Mansur-Alves, M., Lelê, Á. J., & Bandeira, D. R. (2007). Inexistência de diferenças de sexo no fator g (inteligência geral) e nas habilidades específicas em crianças de duas capitais brasileiras. *Psicologia Reflexão e Crítica*, 20(3), 499–506. <https://doi.org/10.1590/s0102-7972007000300018>
- Flores-Mendoza, C., Widaman, K. F., Rindermann, H., Primi, R., Mansur-Alves, M., & Pena, C. C. (2013). Cognitive sex differences in reasoning tasks: Evidence from Brazilian samples of educational settings. *Intelligence*, 41(1), 70–84. <https://doi.org/10.1016/j.intell.2012.11.002>
- Geary, D. C. (2021). *Male, female: The evolution of human sex differences* (3rd ed.). American Psychological Association. <https://doi.org/10.1037/12072-000>
- Giofrè, D., Allen, K., Toffalini, E., Mammarella, I. C., & Caviola, S. (2022). Decoding gender differences: Intellectual profiles of children with specific learning disabilities. *Intelligence*, 90, 101615. <https://doi.org/10.1016/j.intell.2015.07.002>
- Giofrè, D., & Cornoldi, C. (2015). The structure of intelligence in children with specific learning disabilities is different as compared to typically development children. *Intelligence*, 52, 36–43. <https://doi.org/10.1016/j.intell.2015.07.002>

- Giofrè, D., Cornoldi, C., Martini, A., & Toffalini, E. (2020). A population level analysis of the gender gap in mathematics: Results on over 13 million children using the INVALSI dataset. *Intelligence*, *81*, 101467. <https://doi.org/10.1016/j.intell.2020.101467>
- Golay, P., Reverte, I., Rossier, J., Favez, N., & Lecerf, T. (2013). Further insights on the french WISC-IV factor structure through bayesian structural equation modeling. *Psychological Assessment*, *25*(2), 496–508. <https://doi.org/10.1037/a0030676>
- Goldbeck, L., Daseking, M., Hellwig-Brida, S., Waldmann, H. C., & Petermann, F. (2010). Sex Differences on the German Wechsler Intelligence Test for Children (WISC-IV). *Journal of Individual Differences*, *31*(1), 22–28. <https://doi.org/10.1027/1614-0001/a000003>
- Grigorenko, E. (2000). Intelligence and Educational Practices: Deja Vu all Over Again. *Educational Psychology Review*, *12*(3), 361–364. <https://doi.org/10.1023/A:1009074107281>
- Haaf, J. M. (2021). Conventional publication bias correction methods. *PsyArXiv*. <https://doi.org/10.31234/osf.io/gv4tw>
- Hajovsky, D. B., Villeneuve, E. F., Reynolds, M. R., Niileksela, C. R., Mason, B. A., & Shudak, N. J. (2018). Cognitive ability influences on written expression: Evidence for developmental and sex-based differences in school-age children. *Journal of School Psychology*, *67*, 104–118. <https://doi.org/10.1016/j.jsp.2017.09.001>
- Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, R. C., Hyde, J. S., & Gernsbacher, M. A. (2007). The science of sex differences in science and mathematics. *Psychological Science in the Public Interest*, *8*(1), 1–51. <https://doi.org/10.1111/j.1529-1006.2007.00032.x>
- Halpern, D. F., & LaMay, M. L. (2000). Smarter Sex: A Critical Review of Sex Differences in Intelligence. *Educational Psychology Review*, *12*(2), 229. <https://doi.org/10.1023/A:1009027516424>
- Halpern, D. F., & Wai, J. (2019). Sex differences in intelligence. *The Cambridge Handbook of Intelligence*, 317–345. <https://doi.org/10.1017/9781108770422.015>
- Hedges, L., & Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science*, *269*(5220), 41–45. <https://doi.org/10.1126/science.7604277>
- Hernández, A., Aguilar, C., Paradell, È., Muñoz, M. R., Vannier, L. C., & Vallar, F. (2017). El efecto de las variables demográficas en la evaluación de la aptitud cognitiva. *Psicothema*, *29*(4), 469–474. <https://doi.org/10.7334/psicothema2017.33>
- Hunt, E. (2011). *Human Intelligence*. Cambridge University Press.
- Iliescu, D., Ilie, A., Ispas, D., Dobrean, A., & Clinciu, A. I. (2016). Sex differences in intelligence: A multi-measure approach using nationally representative samples from Romania. *Intelligence*, *58*, 54–61. <https://doi.org/10.1016/j.intell.2016.06.007>
- Irwing, P., & Lynn, R. (2005). Sex differences in means and variability on the progressive matrices in university students: A meta-analysis. *British Journal of Psychology*, *96*(4), 505–524. <https://doi.org/10.1348/000712605X53542>
- Jackson, D. N., & Rushton, J. P. (2006). Males have greater g: Sex differences in general mental ability from 100,000 17- to 18-year-olds on the Scholastic Assessment Test. *Intelligence*, *34*(5), 479–486. <https://doi.org/10.1016/j.intell.2006.03.005>
- Jensen, A. R. (1998). *The g factor: The science of mental ability*. Praeger.
- Jensen, A. R., & Reynolds, C. R. (1983). Sex differences on the WISC-R. *Personality and Individual Differences*, *4*(2), 223–226. [https://doi.org/10.1016/0191-8869\(83\)90029-6](https://doi.org/10.1016/0191-8869(83)90029-6)
- Johnson, W., & Bouchard, T. J. (2005). The structure of human intelligence: It is verbal, perceptual, and image rotation (VPR), not fluid and crystallized. *Intelligence*, *33*(4), 393–416. <https://doi.org/10.1016/j.intell.2004.12.002>
- Johnson, W., & Bouchard, T. J. (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>
- Johnson, W., Carothers, A., & Deary, I. J. (2008). Sex differences in variability in general intelligence: A new look at the old question. *Perspectives on Psychological Science*, *3*, 518–531. <https://doi.org/10.1111/j.1745-6924.2008.00096.x>
- Kaufman, A. S., & Doppelt, J. E. (1976). Analysis of WISC-R Standardization Data in Terms of the Stratification Variables. *Child Development*, *47*(1), 165. <https://doi.org/10.2307/1128295>
- Keith, T. Z. T., Fine, J. G. J., Taub, G. G. E., Reynolds, M. R., & Kranzler, J. H. (2006). Higher-order, multisample, confirmatory factor analysis of the Wechsler Intelligence Scale for Children - fourth edition: What does it measure? *School Psychology Review*, *35*(1), 108–127.
- Keith, T. Z., Reynolds, M. R., Patel, P. G., & Ridley, K. P. (2008). Sex differences in latent cognitive abilities ages 6 to 59: Evidence from the Woodcock-Johnson III tests of cognitive abilities. *Intelligence*, *36*(6), 502–525. <https://doi.org/10.1016/j.intell.2007.11.001>

- Lingo VanGilder, J., Lohse, K. R., Duff, K., Wang, P., & Schaefer, S. Y. (2021). Evidence for associations between Rey-Osterrieth Complex Figure test and motor skill learning in older adults. *Acta Psychologica*, 214, 103261. <https://doi.org/10.1016/j.actpsy.2021.103261>
- Loring-meier, S., & Halpern, D. F. (1999). Sex differences in visuospatial working memory: Components of cognitive processing. *Psychonomic Bulletin & Review*, 6(3), 464–471. <https://doi.org/10.3758/BF03210836>
- Lynn, R. (1999). Sex differences in intelligence and brain size: A developmental theory. *Intelligence*, 27(1), 1–1.
- Lynn, R. (2021). Sex Differences in verbal abilities in the Wechsler tests: A review. *Mankind Quarterly*, 61(3), 688–706. <https://doi.org/10.46469/mq.2021.61.3.18>
- Lynn, R., Fergusson, D. M., & Horwood, L. J. (2005). Sex differences on the WISC-R in New Zealand. *Personality and Individual Differences*, 39(1), 103–114. <https://doi.org/10.1016/j.paid.2004.12.009>
- Lynn, R., & Irwing, P. (2004). Sex differences on the progressive matrices: A meta-analysis. *Intelligence*, 32(5), 481–498. <https://doi.org/10.1016/j.intell.2004.06.008>
- Lynn, R., & Kanazawa, S. (2011). A longitudinal study of sex differences in intelligence at ages 7, 11 and 16 years. *Personality and Individual Differences*, 51(3), 321–324. <https://doi.org/10.1016/j.paid.2011.02.028>
- Lynn, R., & Mulhern, G. (1991). A comparison of sex differences on the Scottish and American standardisation samples of the WISC-R. *Personality and Individual Differences*, 12(11), 1179–1182. [https://doi.org/10.1016/0191-8869\(91\)90082-M](https://doi.org/10.1016/0191-8869(91)90082-M)
- Maccoby, E. E., & Jacklin, C. N. (1974). *The psychology of sex differences*. Stanford University Press.
- Maeda, Y., & Yoon, S. Y. (2013). A meta-analysis on gender differences in mental rotation ability measured by the Purdue spatial visualization tests: Visualization of rotations (PSVT:R). *Educational Psychology Review*, 25(1), 69–94. <https://doi.org/10.1007/s10648-012-9215-x>
- McDermott, P. A., Watkins, M. W., & Rhoad, A. M. (2013). Whose IQ Is It?—Assessor Bias Variance in High-Stakes Psychological Assessment. *Psychological Assessment*, 26(1), No Pagination Specified. <https://doi.org/10.1037/a0034832>
- McCall, R. B. (1977). Childhood IQ's as Predictors of Adult Educational and Occupational Status. *Science*, 197(4302), 482–483. <https://doi.org/10.1126/science.197.4302.482>
- McGill, R. J., & Canivez, G. L. (2017). Confirmatory factor analyses of the WISC-IV Spanish core and supplemental subtests: Validation evidence of the Wechsler and CHC models. *International Journal of School & Educational Psychology*, 1–13. <https://doi.org/10.1080/21683603.2017.1327831>
- Mercer, J. R., & Smith, J. M. (1972). Subtest estimates of the WISC full scale IQ's for children. *Vital and Health Statistics Ser. 1: Programs and Collection Procedures*, 2(47), 1–42.
- Miller, D. I., & Halpern, D. F. (2014). The new science of cognitive sex differences. *In Trends in Cognitive Sciences*, 18(1), 37–45. <https://doi.org/10.1016/j.tics.2013.10.011>
- Moffat, S. D., Hampson, E., & Hatzipantelis, M. (1998). Navigation in a “Virtual” maze: Sex differences and correlation with psychometric measures of spatial ability in humans. *Evolution and Human Behavior*, 19(2), 73–87. [https://doi.org/10.1016/S1090-5138\(97\)00104-9](https://doi.org/10.1016/S1090-5138(97)00104-9)
- Morra, S., & Panesi, S. (2017). From scribbling to drawing: The role of working memory. *Cognitive Development*, 43, 142–158. <https://doi.org/10.1016/j.cogdev.2017.03.001>
- Nazareth, A., Huang, X., Voyer, D., & Newcombe, N. (2019). A meta-analysis of sex differences in human navigation skills. *Psychonomic Bulletin & Review*, 26(5), 1503–1528. <https://doi.org/10.3758/s13423-019-01633-6>
- Nowell, A., & Hedges, L. V. (1998). Trends in gender differences in academic achievement from 1960 to 1994: An analysis of differences in mean, variance and extreme scores. *Sex Roles*, 39, 21–43. <https://doi.org/10.1023/A:1018873615316>
- Petersen, J. (2018). Gender Difference in Verbal Performance: A Meta-analysis of United States State Performance Assessments. *Educational Psychological Review*, 30, 1269–1281. <https://doi.org/10.1007/s10648-018-9450-x>
- Pezzuti, L., & Orsini, A. (2016). Are there sex differences in the Wechsler Intelligence Scale for Children - Forth Edition? *Learning and Individual Differences*, 45, 307–312. <https://doi.org/10.1016/j.lindif.2015.12.024>
- R Core Team. (2021). *R: A language and environment for statistical computing* (3.1.2). R Foundation for Statistical Computing. <http://www.r-project.org/>
- Reilly, D., Neumann, D. L., & Andrews, G. (2015). Sex differences in mathematics and science achievement: A meta-analysis of National Assessment of Educational Progress assessments. *Journal of Educational Psychology*, 107(3), 645–662. <https://doi.org/10.1037/edu0000012>



- Reilly, D., Neumann, D. L., & Andrews, G. (2019). Gender differences in reading and writing achievement: Evidence from the National Assessment of Educational Progress (NAEP). *American Psychologist*, 74(4), 445–458. <https://doi.org/10.1037/amp0000356>
- Reynolds, M. R., Keith, T. Z., Ridley, K. P., & Patel, P. G. (2008). Sex differences in latent general and broad cognitive abilities for children and youth: Evidence from higher-order MG-MACS and MIMIC models. *Intelligence*, 36(3), 236–260. <https://doi.org/10.1016/j.intell.2007.06.003>
- Reynolds, M. R., Scheiber, C., Hajovsky, D. B., Schwartz, B., & Kaufman, A. S. (2015). Gender differences in academic achievement: Is writing an exception to the gender similarities hypothesis? *The Journal of Genetic Psychology*, 176(4), 211–234. <https://doi.org/10.1080/00221325.2015.1036833>
- Rodgers, M. A., & Pustejovsky, J. E. (2021). Evaluating meta-analytic methods to detect selective reporting in the presence of dependent effect sizes. *Psychological Methods*, 26(2), 141–160. <https://doi.org/10.1037/met0000300>
- Roivainen, E. (2011). Gender differences in processing speed: A review of recent research. *Learning and Individual Differences*, 21(2), 145–149. <https://doi.org/10.1016/j.lindif.2010.11.021>
- Savage-McGlynn, E. (2012). Sex differences in intelligence in younger and older participants of the Raven's Standard Progressive Matrices Plus. *Personality and Individual Differences*, 53(2), 137–141. <https://doi.org/10.1016/j.paid.2011.06.013>
- Scheiber, C. (2016). Is the Cattell–Horn–Carroll-Based Factor Structure of the Wechsler Intelligence Scale for Children—Fifth Edition (WISC-V) Construct Invariant for a Representative Sample of African-American, Hispanic, and Caucasian Male and Female Students Ages 6 to 16 Ye. *Journal of Pediatric Neuropsychology*, 2(3–4), 79–88. <https://doi.org/10.1007/s40817-016-0019-7>
- Seashore, H., Wesman, A., & Doppelt, J. (1950). The standardization of the Wechsler Intelligence Scale for Children. *Journal of Consulting Psychology*, 14(2), 99–110. <https://doi.org/10.1037/h0056307>
- Stanley, T. D. (2017). Limitations of PET-PEESE and other meta-analysis methods. *Social Psychological and Personality Science*, 8(5), 581–591. <https://doi.org/10.1177/2F1948550617693062>
- Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, 5(1), 60–78. <https://doi.org/10.1002/jrsm.1095>
- Strand, S., Deary, I. J., & Smith, P. (2006). Sex differences in cognitive abilities test scores: A UK national picture. *British Journal of Educational Psychology*, 76(3), 463–480. <https://doi.org/10.1348/000709905X50906>
- van der Sluis, S., Derom, C., Thiery, E., Bartels, M., Polderman, T. J. C., Verhulst, F. C., Jacobs, N., van Gestel, S., de Geus, E. J. C., Dolan, C. V., Boomsma, D. I., & Posthuma, D. (2008). Sex differences on the WISC-R in Belgium and The Netherlands. *Intelligence*, 36(1), 48–67. <https://doi.org/10.1016/j.intell.2007.01.003>
- van der Sluis, S., Posthuma, D., Dolan, C. V., de Geus, E. J. C., Colom, R., & Boomsma, D. I. (2006). Sex differences on the Dutch WAIS-III. *Intelligence*, 34, 273–289. <https://doi.org/10.1016/j.intell.2005.08.002>
- Viechtbauer, W. (2010). Conducting Meta-Analyses in R with the metafor Package. *Journal of Statistical Software*, 36(3). <https://doi.org/10.18637/jss.v036.i03>
- Viechtbauer, W., & Cheung, M.W.-L. (2010). Outlier and influence diagnostic for meta-analysis. *Research Synthesis Methods*, 1, 112–125. <https://doi.org/10.1002/jrsm.11>
- Waschl, N., & Burns, N. R. (2020). Sex differences in inductive reasoning: A research synthesis using meta-analytic techniques. *Personality and Individual Differences*, 164, 109959. <https://doi.org/10.1016/j.paid.2020.109959>
- Wechsler, D. (1949). *Wechsler intelligence scale for children*. The Psychological Corporation.
- Wechsler, D. (1974). *Wechsler intelligence scale for children—revised*. Psychological Corporation.
- Wechsler, D. (1991). *The Wechsler intelligence scale for children—third edition*. The Psychological Corporation.
- Wechsler, D. (2014). *Wechsler intelligence scale for children—Fifth Edition*. NCS Pearson.
- Weiss, L. G., Keith, T. Z., Zhu, J., & Chen, H. (2013). Technical and practical issues in the structure and clinical invariance of the Wechsler Scales: A rejoinder to commentaries. *Journal of Psychoeducational Assessment*, 31(2), 235–243. <https://doi.org/10.1177/0734282913478050>
- Wicherts, J. M. (2016). The importance of measurement invariance in neurocognitive ability testing. *Clinical Neuropsychologist*, 30(7), 1006–1016. <https://doi.org/10.1080/13854046.2016.1205136>
- Wicherts, J. M., & Dolan, C. V. (2010). Measurement invariance in confirmatory factor analysis: An illustration using IQ test performance of minorities. *Educational Measurement: Issues and Practice*, 29(3), 39–47. <https://doi.org/10.1111/j.1745-3992.2010.00182.x>

- Wood, W., & Eagly, A. H. (2012). Biosocial construction of sex differences and similarities in behavior. In M. P. Zanna, J. M. Olson (Eds.), *Advances in Experimental Social Psychology* (Vol. 46, pp. 55–123). Academic Press.
- Zwetsloot, P.-P., Van Der Naald, M., Sena, E. S., Howells, D. W., IntHout, J., De Groot, J. A., Chamuleau, S. A., MacLeod, M. R., & Wever, K. E. (2017). Standardized mean differences cause funnel plot distortion in publication bias assessments. *ELife*, 6<https://doi.org/10.7554/eLife.24260>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.