Personality and Individual Differences 229 (2024) 112776

Contents lists available at ScienceDirect

Personality and Individual Differences

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

Review

Sex differences in cognition: A meta-analysis of variance ratios in the Wechsler Intelligence Scales for Children

David Giofrè [a,*], Enrico Toffalini [b,**], Ambra Perugini [c], Lorenzo Esposito [a], Guido Amoretti [a], David C. Geary [d]

[a] Disfor, University of Genoa, Italy
[b] Department of General Psychology, University of Padua, Italy
[c] Department of Developmental Psychology and Socialization, University of Padua, Italy
[d] University of Missouri, USA

ARTICLE INFO

Keywords:
Sex differences
Variance ratio
Wechsler Intelligence Scales for Children (WISC)
Meta-analysis
Cattell-Horn-Carroll (CHC)
Children

ABSTRACT

A meta-analytic approach was used to explore sex differences in within-sex variation across different versions of the Wechsler Intelligence Scale for Children (WISC). Mean sex differences and variance ratios (*VRs*) were obtained from 75 studies that included 124 independent samples, 571 effect sizes. The outcomes included full scale IQ, individual subtests, and composite indices based on the Cattell-Horn-Carroll ability model (e.g., visual processing, *g*V, and crystallized intelligence, *g*C). Although males were more variable in several domains (e.g., *g*V, *g*C), females were more variable for processing speed (*g*S), and there were no *VR* differences for still other domains (e.g., short-term verbal memory). The cross-domain differences in *VRs* were linearly related to the magnitude of mean sex differences, suggesting that the sex with the advantage at the mean is often the more variable sex. A theoretical model that accounts for this pattern is proposed.

Darwin (1871) noted the greater vulnerability and variability of males across many species, including humans, based on higher premature mortality rates among them. Thorndike (1906) extended the idea to human mental abilities with the argument that mean sex differences were often small, but variation among males was typically larger than among females, whereas Galton (1894) argued that females were the more variable sex. In contrast, McNemar and Terman's (1936) early review suggested that the more variable sex depended on the trait in question. The issue of sex differences in within-sex variability remains unresolved. A fuller understanding of these differences is socially important, because many scientific, technical, and even artistic contributions are disproportionately made by individuals at the extremes of one or several competencies (Bernstein et al., 2019; Humphreys et al., 1993; Lubinski et al., 2023). In other words, sex differences in the numbers of individuals at the extreme right-tail of any ability distribution (or combination of distributions) could contribute to sex differences in socially visible outcomes, such as the proportion of men and women who excel in science, technology, engineering, and mathematics (STEM) fields (e.g., Kell et al., 2013).

The first goal here is to provide a conceptual start point for predicting the traits for which males, females, or neither are the more variable sex. The basic argument is that the sex with the mean advantage will (except under ideal conditions, below) be the more variable sex. The second goal was to provide a meta-analytic review of mean sex differences and sex differences in variability across the different version of the Wechsler Intelligence Scale for Children (WISC), and to examine the relation between these mean and variance differences.

1. Sex differences in trait variance

Jensen (1971) introduced the contemporary approach to the study of sex differences in trait variability, with the concept of the variance ratio (*VR*). The latter is calculated by dividing the male variance by the female variance such that *VR*s >1.0 indicate more variability among males and those <1.0 indicate more variability among females (Feingold, 1992; Hedges & Nowell, 1995). Using this approach, Jensen (1971) concluded that males tended to be more variable in general intelligence.

Subsequent studies focused on a broader range of cognitive abilities

* Correspondence to: D. Giofrè, Disfor, University of Genoa, Corso Andrea Podestà, 2, Italy.
** Correspondence to: E. Toffalini, Department of General Psychology, Via Venezia 8, Padova, Italy.
E-mail addresses: david.giofre@gmail.com (D. Giofrè), enrico.toffalini@unipd.it (E. Toffalini).

https://doi.org/10.1016/j.paid.2024.112776
Received 27 February 2024; Received in revised form 18 June 2024; Accepted 20 June 2024
Available online 26 June 2024

and confirmed that males are often more variable than females. Results on the *VR* are typically based on adults, and thus less is known about sex differences in variability in children. Feingold's (1992) assessment of *VR*s for the national norms from several standardized test batteries revealed a mixed pattern. For the Differential Aptitude Test (DAT), which is a standardized test battery for adolescents and adults, males were more variable for numerical ability, mechanical reasoning, space relations, and spelling, but variances were similar on tests of verbal reasoning, abstract reasoning, and language. On various adult Wechsler scales, males were typically more variable on the Information, Digit Span, and Block Design subtests (*VR*s ranging from 1.11 to 1.16), but females were typically more variable on a Processing Speed subtest (*VR* = 0.91). Variance ratios were close to 1.0 for other subtests including Comprehension, Similarities, Vocabulary, and Object Assembly. Irwing (2012) also found a mixed pattern for the Wechsler Adult Intelligence Scale-III (WAIS-III). Males were more variable on a composite measure of working memory and females on a composite measure of processing speed. There were no sex differences in variability for general intelligence, or summary measures of Verbal Comprehension (e.g., Vocabulary, General Knowledge) or Perceptual Organization (largely spatial abilities); for the latter males were more variable (1.05 to 1.24) for three of the four subtests, but their significance was not assessed.

Related studies have also revealed mixed patterns for general intelligence, as well as for more circumscribed domains, such as short-term memory, and for specific ability measures (Johnson et al., 2008; Keith et al., 2008; Keith et al., 2011; Lakin & Gambrell, 2014). To further complicate this issue, Feingold (1994) found that *VR*s can vary across nations and others have found differences across age, spanning from 5-year-olds to adults. In a related study, 2-to-10-year-old British boys had a greater *VR* on several different instruments including some subtests of the WISC (Arden & Plomin, 2006). These mixed patterns are also found at a basic biological level, with males being more variable in energy intake (among other things) and females for platlet count (among other things) (Halsey et al., 2023). Moreover, human males have higher metabolic rates and are more variable in these rates, but the extent of this variability peaks in early adulthood and then gradually declines (Halsey et al., 2022). At this point, it is clear that broad statements about greater male variability or for that matter greater female variability (see Galton, 1894) are not empirically supported (Halsey et al., 2023).

2. Theoretical framework

There is no principled approach to make sense of why males are more variable for some traits, females for others, and no sex differences for still others, nor is there an approach for making sense of age and contextual influences on the magnitude of any such differences. An evolutionary approach might provide a first step toward making sense of these patterns. From this perspective, sex differences are largely found for traits that facilitate competition for mates and influence mate choices (Darwin, 1871). Critically, the expression of these traits is variable and depends on the phyiscal health of individuals and the conditions underwhich they developed and are currently living (Cotton et al., 2004; Geary, 2015; Johnstone, 1995). In other words, the expression of these traits is condition-dependent and thus mean sex differences and intrasexual variation are expected to vary across time and contexts (Geary, 2021). Sex differences are predicted to be larger and intrasexual variation lower in populations with very low levels of stressor exposure, with smaller sex differences and greater intrasexual variation in populations under stress. Typically though exposure to environmental stressors is not evenly distributed across the population which could still result in mean sex differences (though not as large as they could be) along with large intrasexual variation.

Indeed, there is evidence that broad improvements in environmental conditions (across countries or generations) are associated with larger sex differences in some cognitive abilities, including females' advantage for episodic memory (Asperholm et al., 2019) and males' for spatial abilities (Lippa et al., 2010), although *VR*s have not been assessed in these studies. However, the psychometic measures used in studies of *VR*s are not a one-to-one match to evolved traits, which might limit the applicability of this approach. For instance, there's stong evidence that the brain systems that support spatial navigation and related abilities are evolutionarily old and conserved across vertibrates (Broglio et al., 2015) and that psychometic measures of visuospatial abilities are capturing these same traits (Geary, 2022). However, the psychometic tradition also includes academic measures (e.g., quantitative reasoning, reading) that do not have a deep evolutionary history. The result is that there is not a straightforward way to predict a priori which psychometric abilities will favor one sex or the other, with a few exceptions (e.g., a male advantage in visuospatial abilities).

Nonetheless, a more general principle emerges that might be applicable to psychometic measures (Geary, 2018). Relative to other traits, exaggerated traits – regardless of why they are exaggerated –require more energy expenditure and often a longer developmental period for their full expression, which in turn makes them more vulnerable to disruption (Geary, 2017). Whatever the source of the sex difference, the key idea is that traits for which males have advantages over females are predicted to be more susceptable to stressors and thus show more intrasexual variability relative to the same trait in females or other traits in males. The same applies to traits in which females have advantages. For psychometic measures, the magnitude of mean sex differences, whether they favor males or females, are thus predicted to be linearly related to the magnitude of *VR*s.

3. Current study

The main aim of this meta-analysis was to explore sex differences in variance ratios (*VR*s) for the WISC and the relation between the magnitude of mean sex differences and the magnitude of these ratios. We followed the traditional Cattell-Horn-Carroll (CHC) framework to evaluate differences in the composite indices outlined by McGrew (2009). The CHC model organizes cognitive abilities into different levels. At the apex is general ability or the *g*-factor, measured by the full-scale IQ (FSIQ), with several narrow abilities at the base level and several broad abilities situated at the intermediate level. These broad abilities include fluid intelligence (*g*F), visual processing (*g*V), crystallized intelligence (*g*C), short-term memory (*g*SM), and processing speed (*g*S), among others, as detailed by Flanagan et al. (2013). Analyses of both older and recent versions of the WISC suggest that the main subtests of the WISC battery align well with the CHC approach.

Indeed, the latest version of the WISC, the WISC-V, has been framed according to the CHC model (Wechsler, 2014). One benefit is that results from different versions of the WISC can be compared and this is generally referred to as the cross-battery approach (Bowden, 2013; Flanagan et al., 2007). In other words, WISC versions can be compared using the same subtests and using the broader cognitive domains identified in the CHC model (Keith et al., 2006; LeAdelle et al., 2005; Woodcock, 1990). In addition to investigating *VR*, we assessed whether these varied with WISC test version, age, and the magnitude of mean sex differences. The latter specifically tested the hypothesis that *VR*s will increase with increases in the magnitude of any sex differences, with greater variance for the sex with the mean advantage.

4. Method

4.1. WISC subtests

A short description of each subtest across the various versions of the WISC and included in this meta-analysis is provided below. Not all subtests are found on all WISC versions, but all versions have the subtests needed to assess the broad CHC domains.

Matrix Reasoning (MR): Children are presented with a series of patterns and asked to determine which object or picture from an array

logically completes the pattern. The test taps fluid intelligence and logical reasoning.

Picture Arrangement (PA): Children are given a series of comic-strip-like pictures in a random order. The task is to arrange them as quickly as possible to form a reasonable and meaningful story, assessing narrative reasoning and sequencing skills.

Block Design (BD): Children are asked to rearrange blocks with various color patterns on different sides to match a given pattern. This subtest assesses spatial visualization ability and motor skills.

Object Assembly (OA): Children are presented with puzzle pieces in a standard arrangement and must fit them together to form a meaningful whole. This subtest assesses visual-motor coordination and problem-solving skills.

Picture Concepts (PC): For each item, children are presented with two or three rows of pictured objects, and asked to select the single picture from each row that shares a common characteristic with the others. This subtest assessed abstract and categorical reasoning.

Mazes (MZ): Children navigate a complex maze, planning their path from start to finish without lifting the pencil. This subtest measures planning abilities and visual-motor coordination.

Vocabulary (VC): This subtest measures word knowledge and verbal concept formation by requiring children to define a list of presented words.

Similarities (SI): Children describe how two seemingly disparate objects or concepts are similar, testing abstract thinking and the ability to verbalize conceptual relationships.

Comprehension (CO): Children are asked to answer questions about social rules and expressions, assessing understanding of social conventions and practical judgment.

Information (IN): Children answer questions about a wide range of general knowledge topics, testing their breadth of acquired knowledge.

Digit Span (DS): The examiner states a sequence of number words, and the child is asked to repeat them in the same order (forward DS) or in reverse order (backward DS). This subtest assessed verbal memory and attentional capacity.

Letter-Number Sequencing (LN): Children reorder a mixed sequence of numbers and letters into their correct numerical and alphabetical sequence, assessing working memory and attentional control.

Symbol Search (SS): This test involves identifying matching symbols from a group, measuring speed of visual scanning, processing, and recognition.

Coding (CD): Children transcribe a code by pairing symbols with simple geometric shapes or numbers according to a key, testing graph-omotor speed, visual perception, and memory.

4.2. Literature search and inclusion criteria

The systematic literature search builds on previous work by Giofrè et al. (2022). Searches were conducted from 1961 through October 4, 2023, using three online databases: Scopus, PubMed, and PsychINFO. The search query included the terms: ("WISC" OR "Wechsler Intelligence Scale for Children*") AND ("Sex difference*" OR "Gender difference*"), with the * wildcard to capture any possible alterations of these terms. In the case of Scopus, the search was conducted within the Article Title, Abstract, and Keywords fields. For PubMed and PsycINFO, the search encompassed all available fields. All search results were gathered into a single spreadsheet, which listed information such as the source database, authors, title, year of publication, and publication details (e.g., volume, issue, pages, DOI, abstract) when available. Initially, 1135 records were identified, reduced to 994 after duplicates were removed.

Upon screening titles and abstracts, 146 were excluded due to the unavailability of abstracts, and 629 were deemed not relevant to our research query. Consequently, 219 full-text articles were assessed for eligibility. If access to a paper was not possible through academic library resources, additional online searches were undertaken. Authors were contacted as a last resort, particularly when relevant data was collected but not reported directly. Exclusion criteria detailed in Fig. 1 included studies that matched IQs between groups, used abbreviated WISC versions for general intelligence scoring, combined results from different batteries, lacked sex-disaggregated data, employed experimental WISC tasks, reported means without standard deviations, or had unavailable data. Articles were included if they were written in English or any other language understandable to the authors, such as Italian, Spanish, French, or Portuguese, and reported the sample size, mean, and standard deviation of standardized scores for typically developing boys and girls on at least one WISC index or subscale. If a study did not report data by sex or was missing essential information (e.g., standard deviation, mean), authors were contacted and given one month to supply the needed information. The flowchart in Fig. 1 offers a comprehensive overview of the literature search process. This meta-analysis adheres to the PRISMA guidelines for reporting systematic reviews and meta-analyses.

4.3. Coding of effects, studies, and samples

The meta-analytic dataset was coded in long form, such that multiple effect sizes could be coded per study (e.g., because multiple subtests, or even multiple samples are reported within one study) with each effect size coded on a different row. To avoid redundancies, when both subtests and indices were available, effect sizes were calculated only for subtests. If only indices were reported, they were coded. The only exception was FSIQ, which was always coded if reported, but it was examined in a separate analysis. In the case of longitudinal designs, only data from the first time point were coded.

Variances were calculated as the squares of the reported standard deviations. As the *VR* is a ratio, its logarithm was calculated for meta-analytic computations because it ensures interval scale properties. The variance of the (logarithm of the) *VR* was calculated using a Monte Carlo simulation method for simplicity. The latter was conducted for each row of the dataset. For both males and females, data were drawn from a normal distribution with the same *N*, *M*, and *SD* as in the empirical sample, and this was repeated for 10,000 iterations. For each iteration, the (logarithm of the) *VR* was computed, the variance of its sampling distribution was used as the variance of the effect size. We excluded *VR*s larger than 5.0 (or smaller than 0.2; corresponding to logarithm values beyond ± 1.61), as this value appears implausibly large. This led to the exclusion of 3 out of 574 coded effect sizes (<1 %). No studies or samples were entirely excluded due to implausible values.

In each row of the dataset include information on authors and year of publication, as well as country, block/continent (group by: Africa, East-Asia, East-Europe, Middle East, South-America, Western-Countries), version of the battery (WISC, WISC-R, WISC-III, WISC-IV, WISC-V), whether the study was a standardization sample, number of females and males (*N*), age range and mean age of the sample (in years). If a study included multiple subsamples of different ages, these were coded separately. The concerning row also included data on Cohen's *d*, standard deviation, and mean separately by males and females, and 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 was coded. Effect sizes were coded only if all these pieces of information were available (i.e., effect sizes were not coded if standard deviations were not explicitly reported). We included (when available) *g*-factor/FSIQ scores; *g*F/Fluid Reasoning (subtests: Matrix Reasoning, Picture Concepts; indices: Fluid Reasoning index from WISC-V); *g*V/Visual Processing (subtests: Block Design, Mazes, Object Assembly, Picture Completion; indices: VSI index from WISC-V); *g*C/Verbal Comprehension (subtests: Comprehension, Information, Similarities, Vocabulary; indices: VCI index from WISC-IV/V); *g*SM/Working Memory (subtests: Digit Span, Letter Number Sequencing; indices: WMI index from WISC-IV/V); and, *g*S/ Processing Speed (subtests: Coding, Symbol Search; PSI index from WISC-IV/V).

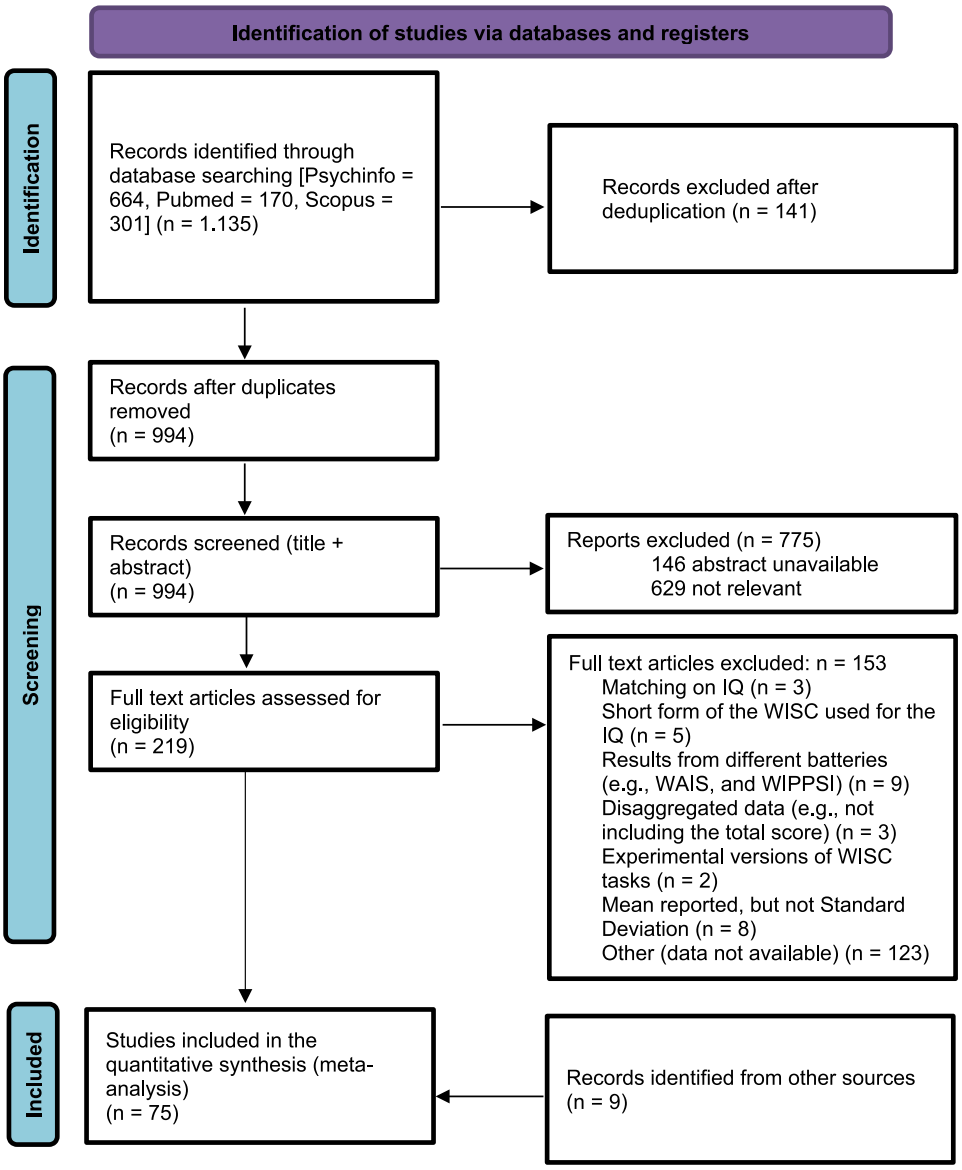

Fig. 1. Prisma Flowchart.

4.4. Analytic strategy

The dataset used in the analyses is available in the OSF repository, accessible via the following link: https://osf.io/p5ht8/.

4.4.1. Meta-analytic model fitting and heterogeneity

The R software (R Core Team, 2023) was used for data analysis. All multilevel meta-analytic models were fitted using functions implemented in the "metafor" package (Viechtbauer, 2010). Different meta-analytic models were computed for estimating the *VR* in Wechsler subtests and in CHC domains. In the latter case, we directly used Wechsler indices when available (e.g., the VCI index was considered appropriate for directly estimating *g*C). Alternatively, when specific indices were not available, relevant subtests were combined; for example, Matrix Reasoning and Picture Concepts were merged to estimate fluid intelligence (*g*F). As recommended by Borenstein et al. (2009), random-effects models were used to account for potential heterogeneity of effect sizes, as detailed below.

Since our analysis included many studies with multiple samples, and some samples contained multiple effect sizes per index or subtest, we fitted multilevel meta-analytic (MLMA) models to account for the structure of dependencies among the effects. We set random intercepts for both study and sample within study. Preliminarily, we aggregated all possible multiple effects within each sample for simplicity, following the strategy reported by Borenstein et al. (2009). With this approach, the aggregated effect size was computed as the average of the multiple effects to be combined. To estimate the aggregated variance, we used the formula for the variance of the mean of correlated variables, as indicated by Borenstein et al. (2009, pp.227-229). Since the correlation value required by this formula is generally not reported in articles, we assumed a correlation of 0.70 across effects within each sample. The rationale is that multiple effects within a sample were collected on the same participants and are likely to be highly correlated to each other. Sensitivity analyses indicated that values in the range from 0.50 to 0.90 led to virtually identical results. Subsequently, we fitted multilevel meta-analytic models with one effect size per sample, while accounting for the presence of multiple samples per study. Random intercepts were estimated for both study and sample/effect size levels. Finally, heterogeneity was estimated considering the two levels of our MLMA models, reporting tau (τ, as the estimated standard deviation across true effects across studies) and omega (ω, as the estimated standard deviation across true effects within studies).

Also, the total I^2 index was computed, as the percentage of total variance that is attributable to the variance across true effects (Higgins, 2003). High values of I^2 (i.e., $I^2 > 75$ %) indicate that most variance across the observed effect sizes can be attributed to true effect size heterogeneity, indicating a potentially large role of moderators.

4.4.2. Moderator analyses

We tested the version of the Wechsler battery and the mean age of the sample as moderators via meta-regression in the meta-analytic models. Q-test of heterogeneity was used to test statistical significance of moderators. For simplicity, the version of the WISC was dichotomized into "old" [WISC and WISC-R] vs. "new" [WISC-III/IV/V]). Our aim was to determine whether more recent versions of the WISC were associated with smaller (or larger) *VR*s. Despite the inclusion of similar subtests across all WISC versions, we opted for this distinction to simplify the meta-regression analysis and avoid an overly complex model with numerous levels. It is also worth noting that the WISC and WISC-R are generally based on a dichotomous distinction between verbal and performance, which was typical of early versions of the scale. At the same time, newer versions incorporated the presence of other neuropsychological aspects, for example the presence of a WM factor (WISC-IV and WISC-IV), which was called the freedom from distractibility factor on the WISC-III, and a processing speed factor, but these were not included in the WISC-R or WISC. For these reasons, a distinction between older and newer versions of the WISC made statistical sense and had a theoretical rationale.

Mean age was in years, and if not explicitly reported, it was coded as the median point of the age range. For simplicity, clarity, and to avoid redundancy, moderator analyses were reported only for the CHC areas except where above indicated.

4.4.3. Assessment of risk of publication bias

In principle, we expected no risk of publication bias, because *VR* is rarely tested in the published literature. Therefore, there is no reason to expect that the estimated *VR* is related to the likelihood of an article being published. Nonetheless, as a routine check, we performed PET-PEESE meta-regression procedure for checking publication bias (Stanley & Doucouliagos, 2014) for each CHC area. In the PET-PEESE method, a meta-regression model is fitted including the standard error as moderator of the effect size (PET) in the first step. This is basically the same as the famous Egger's regression test of symmetry, relating effect size to its standard error. Subsequently, if the model intercept (which represents an estimated bias-free effect) remains significant, the variance is entered as a moderator instead of standard error (PEESE) for a better estimate of the bias-free effect (Haaf, 2021). Meta-regression methods such as Egger's test, and its extension in the PET (Precision Effect Test)-PEESE (Precision Effect Estimate with Standard Error) procedure, are preferred over common alternatives because they can be adapted for the use with multilevel data structures, thus accounting for possible dependencies across effect sizes. In our case, we performed the above meta-regression tests directly in our multilevel models.

4.4.4. Association between VR and Cohen's d

We aimed to examine the relationship between differences in mean scores (male vs. female Cohen's *d*) and *VR*s (male/female *VR*) across subtests and CHC areas. Preliminarily, Cohen's *d*s were calculated, for all subtests and CHC areas, using the same MLMA approach detailed above for *VR*. Subsequently, Cohen's *d*s were used as moderators of *VR*s across different subtests and across different CHC areas, using univariate meta-analytic models. It is worth noting that our goal here was not to establish whether larger *VR*s were associated with larger Cohen's *d* values at the level of study or sample/individual effect size. Instead, our focus was on examining whether entire CHC areas (or subtests) associated with higher meta-analytic *VR*s exhibited larger meta-analytic Cohen's *d*s than CHC areas (or subtests) with lower meta-analytic *VR*s.

5. Results

5.1. Overview of the studies

A total of 75 studies, 124 independent samples, 571 effect sizes, and an estimated total of 45,821 participants (22,914 males and 22,907 females) were included. The mean age was 10.53 years (SD = 2.83; Range 5.2 to 16.5 years). Sixty studies were conducted in Western Countries, three in Africa, three in East Asia, three in East-Europe, two in the Middle East, three in South America, and one included samples from both a Western and East Asian country. Thirty-four studies used old versions of the WISC ($k = 9$ for WISC; $k = 25$ for WISC-R), and 41 used new versions ($k = 16$ for WISC-III; $k = 22$ for WISC-IV; $k = 3$ for WISC-V). Additional information, including forest plots, are provided online (see the Online Appendix).

5.2. Variance ratios in CHC abilities

Table 1 summarized the meta-analytic results for CHC abilities (see also Figs. A1-A6; Appendix, for forest plots).

Meta-analytic estimates of *VR* reached statistical significance for four of the six CHC areas: *g*/FSIQ, *g*V, and *g*C, where larger variances were estimated in males than in females, and *g*S, where a larger variance was estimated in females. The statistical significance of *VR* indicates that the estimate differs from 1.00, meaning its logarithm is not 0.00. All effect sizes were small, with the largest disparity was a *VR* of 1.077 in *g*C, suggesting that variance in one sex did not exceed that in the other by more than roughly 8 %. The heterogeneity was mostly modest, with tau and omega estimates generally small. Notably, *g*SM was the only area where the heterogeneity exceeded the sampling variance, with an I^2value of 61 %.

5.3. Moderator analysis for CHC areas

The version of the battery did not significantly moderate the *VR* for any of the CHC areas, with $Q(1)$ values ranging from 0.029 to 1.941, and *p*-values ranging from 0.165 to 0.865. However, the mean age of sample significantly moderated the *VR* of *g*V, $Q(1) = 4.417$, $B = -0.042$, $p = 0.014$. Such a result indicates that sex differences in the *g*V variance were smaller in older than in younger children. For all the other CHC areas, Q(1) ranged from 0.014 to 1.472, with *p*-values ranging from 0.225 to 0.906.

5.4. Assessment of risk of publication bias for CHC areas

The PET-PEESE meta-regression did not indicate a significant relationship between variance ratios (*VR*) and their precision across most CHC areas, with the exception of *g*F. In this case, the Precision Effect Test (PET) reached significance, suggesting that *VR* was significantly moderated by its standard error in the Multilevel Meta-Analytic (MLMA) regression, $B = -1.010$, $p = 0.004$. However, since the original effect size was not statistically significant, the *VR* was not significantly different from 1 (see Table 1), the Precision Effect Estimate with Standard Error (PEESE) was not performed. For all other CHC areas, PET's *B* ranged from −0.183 to 0.247, with p-values ranging from 0.357 to 0.705, meaning that were all not statistically significant. Overall, these findings suggest no substantial risk of publication bias.

5.5. Variance ratios in subtests

Table 2 summarizes the meta-analytic results for individual WISC subtests (Figs. A7-A20; Appendix). Results were generally in line with those that emerged for CHC abilities, but there were heterogeneous *VR*s for some subtests that indexed the same CHC ability. As shown in Table 2, there was generally more variance among males than females, but these differences were only significant for the Matrix Reasoning,

Table 1
Summary of the meta-analytic results concerning CHC areas.

CHC area	Studies	Samples	Effects	N males	N females	*VR*	*Log (VR)*	*Std. Err.*	*p*	*Tau*	*Omega*	I^2	Cohen's *d*
g/FSIQ	43	67	67	12,472	12,540	1.057 [1.020, 1.095]	0.055	0.018	0.002	0	0	0 %	0.097 [0.039, 0.156]
*g*F	19	27	29	12,713	12,929	1.031 [0.997, 1.066]	0.030	0.017	0.078	0	0	0 %	0.075 [0.021, 0.130]
*g*V	24	34	80	10,688	10,868	1.064 [1.023, 1.108]	0.062	0.020	0.002	0.050	0	24 %	0.197 [0.143, 0.251]
*g*C	42	71	125	14,292	14,292	1.077 [1.040, 1.115]	0.074	0.018	< 0.001	0	0.052	15 %	0.117 [0.063, 0.171]
*g*SM	29	46	47	11,674	12,008	1.007 [0.931, 1.089]	0.007	0.040	0.860	0.135	0.078	61 %	−0.021 [−0.068, 0.027]
*g*S	32	48	50	13,584	13,771	0.947 [0.901, 0.996]	−0.054	0.026	0.036	0.030	0.094	41 %	−0.340 [−0.402, −0.279]

Note. *g*F = fluid intelligence, *g*V = visual intelligence, *g*C = crystallized intelligence, *g*SM = short term memory intelligence, *g*S = processing speed intelligence.

Table 2
Summary of the meta-analytic results concerning the Weschler subtests.

Subtest	CHC Area	Studies	Samples	Effects	N males	N females	*VR*	*Log (VR)*	*Std. Err.*	*p*	*Tau*	*Omega*	I^2	*Cohen's d*
MR	*g*F	7	8	8	5836	5878	1.066 [1.013, 1.122]	0.064	0.026	0.0145	0	0	0 %	−0.041 [−0.088, 0.006]
PCn	*g*F	2	2	2	1925	1925	1.031[0.943, 1.127]	0.031	0.045	0.498	0	–	0	0.018 [−0.088, 0.124]
PA	*g*F	12	19	19	6877	7051	1.010 [0.963, 1.059]	0.010	0.024	0.6748	0	0	0 %	0.134 [0.082, 0.186]
BD	*g*V	24	34	34	10,688	10,868	1.127 [1.027, 1.237]	0.119	0.048	0.012	0.177	0.033	71 %	0.194 [0.149, 0.239]
OA	*g*V	13	18	18	8018	8191	0.987 [0.930, 1.048]	−0.013	0.031	0.6764	0.058	0.036	33 %	0.240 [0.158, 0.322]
PC	*g*V	14	19	19	7885	8036	1.095 [1.045, 1.147]	0.091	0.024	< 0.001	0.022	0	5 %	0.154 [0.077, 0.231]
MZ	*g*V	6	9	9	4945	5144	1.051 [0.905, 1.219]	0.049	0.076	0.516	0.156	0.077	80 %	0.239 [0.151, 0.327]
VC	*g*C	27	43	43	11,413	11,752	1.088 [0.989, 1.196]	0.084	0.049	0.0832	0.201	0	72 %	0.103 [0.043, 0.163]
SI	*g*C	19	24	24	10,539	10,727	1.089 [1.039, 1.142]	0.085	0.024	< 0.001	0	0.056	25 %	0.042 [−0.002, 0.083]
CO	*g*C	16	23	23	9009	9176	0.968 [0.920, 1.018]	−0.033	0.026	0.2005	0	0.053	21 %	0.062 [0.033, 0.091]
IN	*g*C	15	23	23	7459	7618	1.233 [1.150, 1.323]	0.210	0.036	< 0.001	0.032	0.091	42 %	0.251 [0.143, 0.358]
DS	*g*SM	18	24	24	9571	9828	1.062 [0.920, 1.227]	0.061	0.073	0.4093	0.162	0.23	88 %	−0.039 [−0.091, 0.014]
LN	*g*SM	3	3	3	2659	2768	1.019 [0.945, 1.099]	0.019	0.039	0.6289	0	–	0 %	0.046 [−0.011, 0.102]
SS	*g*S	6	7	7	4962	5157	1.005 [0.946, 1.068]	0.005	0.031	0.8679	0	0.033	16 %	−0.156 [−0.229, −0.084]
CD	*g*S	21	26	26	11,707	11,934	0.936 [0.879, 0.996]	−0.067	0.032	0.0377	0.107	0	56 %	−0.393 [−0.452, −0.333]

Note. BD = Block Design; CD = Coding; CO = Comprehension; DS = Digit Span; IN = Information; LN = Letter-Number sequencing; MR = Matrix Reasoning; MZ = Mazes; OA = Object Assembly; PA = Picture Arrangement; PC = Picture Concepts; PCn = Picture Concepts; SI = Similarities; SS = Symbol Search; VC = Vocabulary.

Block Design, Picture Completion, Similarities, and Information subtests. In keeping with the results for *g*S, there was more variation among females for Coding, but not Symbol Search.

5.6. Relationship between VR and Cohen's d across CHC abilities and subtests

Meta-regression indicated that CHC areas associated with a larger *VR* also displayed larger Cohen's *d*. The test for this moderator was significant, $Q(1) = 16.38$, $B = 0.237$, $p < 0.001$. Scatter plots and meta-regression estimates for CHC areas are shown in left panel of Fig. 2. Notably, the model's intercept was significantly different from zero, $logVR = 0.025$, $p = 0.015$, corresponding to a *VR* of 1.025. This suggests that for a hypothetical CHC area with perfectly identical mean scores in males and females, males would still exhibit slightly higher variance.

Similar results were observed at the subtests' level, $Q(1) = 5.35$, $B = 0.23$, $p = 0.021$. Scatter plots and meta-regression estimates are shown in the right panel of Fig. 2. In this case, the intercept was not statistically significant, but the estimate is very close to the results obtained with the CHC areas, $logVR = 0.033$, $p = 0.051$, corresponding to a VR of 1.034.

6. Discussion

Our meta-analytic results confirmed McNemar and Terman's (1936) early conclusion that sex differences in intrasexual variability depend on the trait being assessed, and is consistent with recent studies of more basic biological traits showing the same mixed pattern (Halsey et al., 2023). That said and across different versions of the WISC and across age, males were generally more variable than females, with significantly more variation among males for FSIQ, visual processing, and crystallized intelligence within the CHC framework, as well as more male variation on several of the associated subtests (e.g., Block Design, Information). At

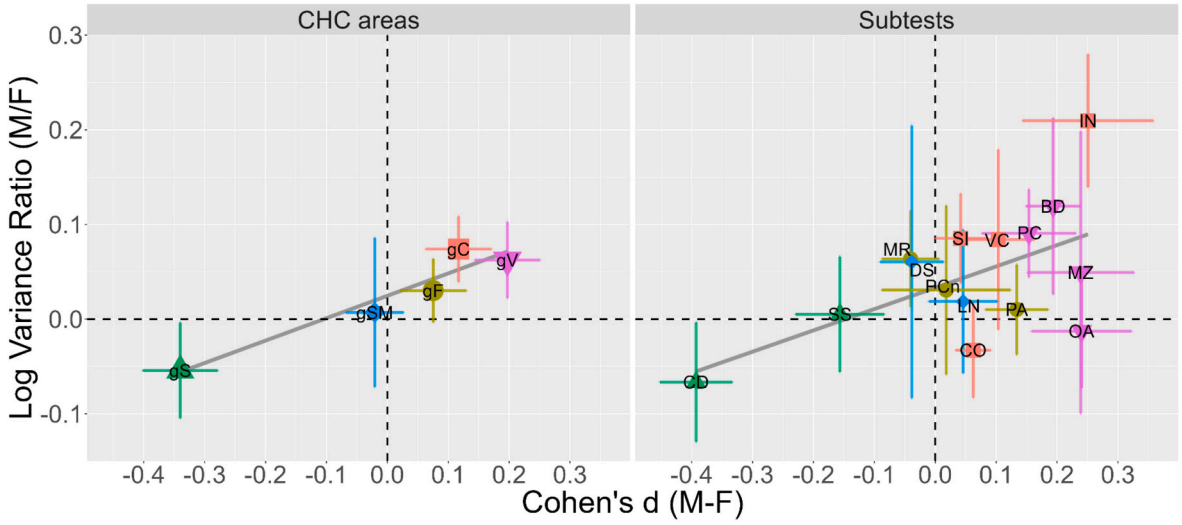

Fig. 2. Scatter plots of meta-analytic estimates of Variance ratio (logarithm) and Cohen's d for CHC areas and for subtests. Solid grey lines represent meta-regression estimate of Cohen's d as a moderator of VR. Error bars (both horizontal and vertical) represent 95 % CIs of meta-analytic estimates.
Note. CHC areas: gF = fluid intelligence/reasoning; gV = visual processing; gC = crystallized intelligence; gSM = working memory; gS = processing speed. Subtests: BD = Block Design; CD = Coding; CO = Comprehension; DS = Digit Span; IN = Information; LN = Letter-Number sequencing; MR = Matrix Reasoning; MZ = Mazes; OA = Object Assembly; PA = Picture Arrangement; PC = Picture Completion; PCn = Picture Concepts; SI = Similarities; SS = Symbol Search; VC = Vocabulary.

the same time, females were consistently more variable on the processing speed factor in the CHC framework. However, this result was largely driven by a single subtest (Coding), in which females consistently outperformed males, while at the same time being more variable. There were no consistent sex differences in variability for short-term memory in the CHC framework or for the associated subtests (Digit Span, Letter-Number Sequence).

6.1. Sex differences in intrasexual variability

In keeping with many prior studies, there were small sex differences for mean FSIQ (d = 0.10) scores, but males were more variable than females (Arden & Plomin, 2006; Deary et al., 2003; Deary et al., 2007; Jensen, 1971; Johnson et al., 2008; Schroeder, 2021). Although debated, mean sex differences for general ability may favor females during the preschool and early elementary school years, with males eventually catching up and potentially surpassing females by early adulthood (Arden & Plomin, 2006; Lynn, 1994). For compositive measures, such as FSIQ, or latent *g* measures, the differences are generally small, however. For 10-year-olds, Ardin and Plomin found a small male advantage of about the same magnitude as we found, Deary and colleagues (Deary et al., 2003; Deary et al., 2007) found no difference for 11-year-olds or across brothers and sisters, whereas Irwing (2012) found a small male advantage for the WAIS-III.

In contrast to these mixed results for mean sex differences in *g*, evidence for greater male variation is more consistent. For the WISC, males were about 6 % more variable in FSIQ than females (*VR* = 1.06), which is the same *g*-factor *VR* found by Arden and Plomin for 9-year-olds (*VR* = 1.06) but considerably smaller than the *VR* for 10-year-olds (*VR* = 1.2) and the *VR* (1.12) for 11-year-olds reported by Deary et al. (2003). In an analysis of the WAIS-III standardization data, Irwing (2012) found an overall *VR* = 1.04 for 16- to 89-year-olds, with no significant age trends, and Deary et al. (2007) found *VR*s from 1.09 to 1.11 (depending on how *g* was measured) comparing brothers and sisters who were adolescents or young adults. Given these patterns, our results are at the lower end of the *VR* estimates for *g*. In any case, one proposal is that *g* is indexing, at least in part, top-down attentional control (Kane & Engle, 2002), and there is evidence that males have advantages in attentional focus (e.g., Stoet, 2017) but many of them also have attentional deficits (Willcutt, 2012). The combination would result in greater male variability on *g* measures to the extent *g* is capturing top-down attentional abilities.

As mentioned, the two other areas with greater male variation were visual processing and crystallized intelligence. The former captures visuospatial abilities and sex differences are commonly found for species in which males have larger travel ranges than females (Gaulin, 1992), which is the case for humans (MacDonald & Hewlett, 1999). At the same time, males' spatial abilities are more susceptible to disruption by stressor exposure, such as man-made toxins, that in turn results in a male advantage but also greater male vulnerability and variability (Geary, 2019). The male advantage and greater variability on crystallized measures is not as readily explained. Males do not appear to have an advantage in the semantic memory system that would facilitate the acquisition of general knowledge (Maitland et al., 2004). When examining individual subtests, the information subtest shows significant sex differences for means and *VR* (see also Liu & Lynn, 2015). Differences also appeared in similarities and vocabulary but were more modest in terms of magnitude. The comprehension subtest showed minimal differences, likely due to lower saturation on crystallized intelligence (Keith et al., 2006). Attention difficulties, particularly in males, can affect performance on tasks requiring prolonged concentration, potentially explaining the higher *VR* in some subtests. At the same time, acquiring crystallized knowledge typically requires extensive study, and reading many books or other materials, which many males may find tedious or difficult: for example, males also tend to have a higher rates of reading problems (Quinn & Wagner, 2015). Consequently, a considerable proportion of males may avoid reading activities, crucial for acquiring the crystallized knowledge assessed in some WISC tasks, which ultimately might contribute to higher *VR* in certain tasks, and this is particularly true for the information subtest. At the same time, some males may be motivated to acquire broad knowledge or extensive knowledge in one or several domains as a way to display their expertise or intelligence as related to their status striving.

Notably, we found significant the mean difference and the *VR* difference in Processing Speed (*g*S), favoring females (see also Feingold, 1992; Irwing, 2012), but this effect was largely driven by one subtest, that is Coding. These results are in keeping with other findings for typically developing children where in virtually all studies females outperformed males on the Coding subtest (Giofrè et al., 2022). Similarly, for children with various neurodevelopmental disabilities, such as specific learning disabilities, there is a striking females advantage on this

subtest (Giofrè et al., 2023; Giofrè et al., 2024; Giofrè & Cornoldi, 2015). These measures are considered to be indexes of fine motor abilities and Processing Speed for simple information (e.g., Lyle & Johnson, 1974), but females' advantage in episodic memory might also be a contributing factor (see Asperholm et al., 2019). One alternative explanation may be tied to the definition of processing speed itself, where females may have advantages because many males have difficulty maintaining attention while engaging in a simple and repetitive task over an extended period (see Giofrè et al., 2022; Flanagan et al., 2000 for a discussion). Indeed, this particular subtest appears to depend heavily on attentional mechanisms, and this is also consistent with recent findings indicating that males generally struggle to maintain attention for prolonged periods (Geary et al., 2023). Finally, the Coding subtest is likely the one in which children with ADHD demonstrate the most significant impairments, confirming that this subtest is probably reliant on attentional mechanisms to a large extent (Giofrè et al., 2023).

In keeping with prior studies, males had a small advantage for fluid abilities (*g*F, $d = 0.08$) and tended to be more variable, but the effects were not significant (e.g., Rodríguez-Cancino & Concha-Salgado, 2023). Sex differences are sometimes found for short-term memory, with females often outperforming males on verbal measures and males outperforming females on visuospatial measures (Pauls et al., 2013). The WISC short-term memory measures are verbal but generally not as difficult as the measures that typically show a female advantage, which might explain why we found no mean sex differences or differences in variation. Whatever the reason, the key point is that the lack of sex differences in variability for fluid abilities and short-term memory confirms the *VR* varies across traits (Johnson et al., 2008; Keith et al., 2008; Keith et al., 2011; Lakin & Gambrell, 2014) and thus broad statements about male or female variability are not supported (Halsey et al., 2023).

In terms of subtests, differences were often statistically significant, but small in terms of the magnitude. It is also worth noting that due to the large power of meta-analytical designs, even smaller effects might be statistically significant. The two highest *VR*s were on the Information subtest, measuring crystallized intelligence (Feingold, 1992; Hedges & Nowell, 1995), and the Block Design subtest, measuring visuospatial abilities (Feingold, 1992; Schroeder, 2021). These subtests also exhibit the highest standardized mean sex differences (Giofrè et al., 2022). It is worth noting that the Comprehension subtest was the only crystallized intelligence task in which differences were not statistically significant. However, the comprehension subtest is more likely to be related to daily life skills than the other measures of crystallized intelligence. In fact, the loading of this task on both general intelligence and on the verbal factor appears to be lower in both children with typical and atypical development (Giofrè et al., 2019; Giofrè & Cornoldi, 2015).

Effects of moderators were generally not statistically significant. The effects reported here did not vary with the use of older or more recent WISC tests, suggesting the mean sex differences and the *VR*s are stable. The only statistically significant effect we found was related to age in the visual processing CHC area. The results indicated a reduction of *VR* in older children. Traditional meta-regression analyses, which used Cohen's *d* as a regressor for each CHC area, did not produce significant results. However, meta-regression indicated that CHC areas associated with a larger *VR* also displayed larger Cohen's *d*, this finding was not observed at the level of the subtest, despite the effect being very similar in terms of magnitude.

6.2. Theoretical implications

In this study we found evidence, at least for some cognitive domains, that *VR* tends to vary with the magnitude of mean sex differences. To further assess this pattern, we correlated *d* and *VR* values for the DAT and the WAIS reported in Feingold (1992). Consistent with our findings for the WISC, larger mean sex differences on DAT and WAIS subtests were associated with larger within-sex variation (rs = 0.71, 0.90, respectively); these are particularly striking correlations. The results are in keeping with our proposed framework, whereby advantages at trait means are associated with increased vulnerabilities to stressors that in turn would contribute to increases in within-sex variation. The basic idea comes from studies of traits that facilitate competition for mates and influence mate choices (Cotton et al., 2004; Geary, 2015; Johnstone, 1995). These traits are typically more elaborated or complex in one sex or the other and their condition dependence (vulnerability) makes them honest signals of the underlying health or specific abilities (e.g., foraging) of the individual (Zahavi, 1975).

The model is directly applicable to some human cognitive sex differences, including the visuospatial abilities assessed by psychometric measures (see Geary, 2017; Geary, 2018), but in theory should be applicable to any complex trait that shows mean sex differences. Our results and the results for the DAT and WAIS, which are based on adults, seem to support this argument but it should only be considered suggestive at this point. It is unclear if the same pattern will emerge for less complex, basic biological traits that also show sex differences. Moreover, the correlations between *d* and *VR* values might vary across time and context. The samples used in our studies and for the DAT and WAIS standardization samples analyzed by Feingold (1992) are largely from Western contexts with comparatively low levels of stressors and it is not clear whether the magnitude of our effects (and those for the DAT and WAIS) will be the same in other contexts. In theory, populations living in sheltered contexts could show increases in mean sex differences and decreases in *VR*, although it is not clear if these idealized conditions exit or if all traits will respond in the same way (Geary, 2021). In populations under stress (e.g., poor health care, poor nutrition), the magnitude of sex differences could decrease along with an increase in *VR*. These are empirical questions that remain to be answered.

6.3. Limitations and conclusions

Some limitations should be considered for future studies. First, our analyses were limited to the WISC. Although the Wechsler scales are widely used, it would be particularly interesting to expand these analyses to other scales, such as the Woodcock-Johnson (Keith et al., 2008). Additionally, the focus on the WISC necessarily limited our study to children and adolescents. The linear relation between *d* and *VR* emerged in our analysis of the WAIS (from Feingold, 1992), but a more thorough analysis of adults is needed. While we found evidence for statistically significant *VR*, effect sizes were generally small. Finally, our moderator analysis is based on a limited number of samples; therefore, the results of these analyses should be interpreted with caution. Future research could address these limitations to provide a more comprehensive understanding of sex differences in intelligence.

Funding

This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Ethics

This research is based on secondary data and does not include any human participants.

CRediT authorship contribution statement

David Giofrè: Writing – review & editing, Writing – original draft, Project administration, Methodology, Formal analysis, Data curation, Conceptualization. **Enrico Toffalini:** Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Data curation, Conceptualization. **Ambra Perugini:** Writing – original draft, Methodology, Data curation. **Lorenzo Esposito:** Writing – original draft, Data curation. **Guido Amoretti:** Writing – original draft. **David Geary:**

Writing – review & editing, Writing – original draft, Methodology, Conceptualization.

Declaration of competing interest

None.

Data availability

Data and materials for the current report are available under reasonable request emailing to the corresponding authors.

Acknowledgements

This paper was finalized while David Giofrè was spending a period of research at the University of Missouri with David Geary. These authors are grateful to the Fulbright commission for having supported David Giofrè's stay in the USA. We would like to express our sincere appreciation to the founders who generously supported this research endeavor. The views expressed in this paper are those of the authors alone and do not necessarily reflect the views of the funders, who bear no responsibility for the content presented herein.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.paid.2024.112776.

References

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

Asperholm, M., Nagar, S., Dekhtyar, S., & Herlitz, A. (2019). The magnitude of sex differences in verbal episodic memory increases with social progress: Data from 54 countries across 40 years. *PLoS One, 14*(4), Article e0214945. https://doi.org/10.1371/journal.pone.0214945

Bernstein, B. O., Lubinski, D., & Benbow, C. P. (2019). Psychological constellations assessed at age 13 predict distinct forms of eminence 35 years later. *Psychological Science, 30*(3), 444–454. https://doi.org/10.1177/0956797618822524

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2009). *Introduction to meta-analysis.* John Wiley & Sons, Ltd.

Bowden, S. C. (2013). Theoretical convergence in assessment of cognition. *Journal of Psychoeducational Assessment, 31*(2), 148–156. https://doi.org/10.1177/0734282913478035

Broglio, C., Martín-Monzón, I., Ocaña, F. M., Gómez, A., Durán, E., Salas, C., & Rodríguez, F. (2015). Hippocampal pallium and map-like memories through vertebrate evolution. *Journal of Behavioral and Brain Science, 05*(03), 109–120. https://doi.org/10.4236/jbbs.2015.53011

Cotton, S., Fowler, K., & Pomiankowski, A. (2004). Do sexual ornaments demonstrate heightened condition-dependent expression as predicted by the handicap hypothesis? *Proceedings of the Royal Society of London. Series B: Biological Sciences, 271*(1541), 771–783. https://doi.org/10.1098/rspb.2004.2688

Darwin, C. (1871). *The descent of man and selection in relation to sex. 2 vols.* London: J. Murray.

Deary, I. J., Irwing, P., Der, G., & Bates, T. C. (2007). Brother–sister differences in the g factor in intelligence: Analysis of full, opposite-sex siblings from the NLSY1979. *Intelligence, 35*(5), 451–456. https://doi.org/10.1016/j.intell.2006.09.003

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

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

Feingold, A. (1994). Gender differences in variability in intellectual abilities: A cross-cultural perspective. *Sex Roles, 30*(1–2), 81–92. https://doi.org/10.1007/BF01420741

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., 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.

Galton, F. (1894). The relative sensitivity of men and women at the nape of the neck (by Weber's test). *Nature, 50*(1280), 40–42. https://doi.org/10.1038/050040a0

Gaulin, S. J. C. (1992). Evolution of sex differences in spatial ability. *Yearbook of Physical Anthropology, 35*, 125–151. https://doi.org/10.1002/ajpa.1330350606

Geary, D. C. (2015). Evolution of vulnerability: Implications for sex differences in health and development. In *Evolution of Vulnerability: Implications for Sex Differences in Health and Development.* https://doi.org/10.1016/C2014-0-00387-5

Geary, D. C. (2017). Evolutionary framework for identifying sex- and species-specific vulnerabilities in brain development and functions. *Journal of Neuroscience Research, 95*(1-2), 355–361. https://doi.org/10.1002/jnr.23794

Geary, D. C. (2018). Efficiency of mitochondrial functioning as the fundamental biological mechanism of general intelligence (g). *Psychological Review, 125*(6), 1028–1050. https://doi.org/10.1037/rev0000124

Geary, D. C. (2019). Evolutionary perspective on sex differences in the expression of neurological diseases. *Progress in Neurobiology, 176*, 33–53. https://doi.org/10.1016/j.pneurobio.2018.06.001

Geary, D. C. (2021). Now you see them, and now you don't: An evolutionarily informed model of environmental influences on human sex differences. *Neuroscience & Biobehavioral Reviews, 125*, 26–32. https://doi.org/10.1016/j.neubiorev.2021.02.020

Geary, D. C. (2022). Spatial ability as a distinct domain of human cognition: An evolutionary perspective. *Intelligence, 90*, Article 101616. https://doi.org/10.1016/j.intell.2021.101616

Geary, D. C., Hoard, M. K., Nugent, L., & Ünal, Z. E. (2023). Sex differences in developmental pathways to mathematical competence. *Journal of Educational Psychology, 115*(2), 212–228. https://doi.org/10.1037/edu0000763

Giofrè, D., Allen, K., Toffalini, E., & Caviola, S. (2022). The impasse on gender differences in intelligence: A meta-analysis on WISC batteries. *Educational Psychology Review, 34*(4), 2543–2568. https://doi.org/10.1007/s10648-022-09705-1

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., Lievore, R., Allen, K., Tonizzi, I., Mammarella, I. C., & Toffalini, E. (2024). *Understanding Sex/Gender Differences in Intelligence Profiles of Children with Autism: A Comprehensive WISC Meta-Analysis.* Manuscript Submitted for Publication.

Giofrè, D., Pastore, M., Cornoldi, C., & Toffalini, E. (2019). Lumpers vs. splitters: Intelligence in children with specific learning disorders. *Intelligence, 76*, Article 101380. https://doi.org/10.1016/j.intell.2019.101380

Giofrè, D., Toffalini, E., Esposito, L., & Cornoldi, C. (2023). Gender differences in the Wechsler intelligence scale for children in a large group of Italian children with attention deficit hyperactivity disorder. *Journal of Intelligence, 11*(9), 178. https://doi.org/10.3390/jintelligence11090178

Haaf, J. M. (2021). Conventional publication bias correction methods. *PsyArXiv*. https://doi.org/10.31234/osf.io/gv4tw

Halsey, L. G., Careau, V., Pontzer, H., Ainslie, P. N., Andersen, L. F., Anderson, L. J., … Speakman, J. R. (2022). Variability in energy expenditure is much greater in males than females. *Journal of Human Evolution, 171*, Article 103229. https://doi.org/10.1016/j.jhevol.2022.103229

Halsey, L. G., Esteves, G. P., & Dolan, E. (2023). Variability in variability: Does variation in morphological and physiological traits differ between men and women? *Royal Society Open Science, 10*(9). https://doi.org/10.1098/rsos.230713

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

Higgins, J. P. T. (2003). Measuring inconsistency in meta-analyses. *BMJ, 327*(7414), 557–560. https://doi.org/10.1136/bmj.327.7414.557

Humphreys, L. G., Lubinski, D., & Yao, G. (1993). Utility of predicting group membership and the role of spatial visualization in becoming an engineer, physical scientist, or artist. *Journal of Applied Psychology, 78*, 250–261. https://doi.org/10.1037/0021-9010.78.2.250

Irwing, P. (2012). Sex differences in g: An analysis of the US standardization sample of the WAIS-III. *Personality and Individual Differences, 53*(2), 126–131. https://doi.org/10.1016/j.paid.2011.05.001

Jensen, A. R. (1971). The race x sex x ability interaction. In R. Cancro (Ed.), *Contributions to intelligence* (pp. 107–161). Grune and Stratton. Johnson.

Johnson, W., Carothers, A., Deary, I. J., & Carothers, A. (2008). Sex differences in variability in general intelligence: A new look at the old question. *Perspectives on Psychological Science, 3*(6), 518–531. https://doi.org/10.1111/j.1745-6924.2008.00096.x

Johnstone, R. A. (1995). Sexual selection, honest advertisement and the handicap principle: Reviewing the evidence. *Biological Reviews, 70*(1), 1–65. https://doi.org/10.1111/j.1469-185X.1995.tb01439.x

Kane, M. J., & Engle, R. W. (2002). The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. *Psychonomic Bulletin & Review, 9*, 637–671. https://doi.org/10.3758/BF03196323

Keith, T. Z., Fine, J. G., Taub, 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

Keith, T. Z., Reynolds, M. R., Roberts, L. G., Winter, A. L., & Austin, C. A. (2011). Sex differences in latent cognitive abilities ages 5 to 17: Evidence from the differential

ability scales—Second edition. *Intelligence, 39*(5), 389–404. https://doi.org/10.1016/j.intell.2011.06.008
Kell, H. J., Lubinski, D., Benbow, C. P., & Steiger, J. H. (2013). Creativity and technical innovation: Spatial ability's unique role. *Psychological Science, 24*(9), 1831–1836. https://doi.org/10.1177/0956797613478615
Lakin, J. M., & Gambrell, J. L. (2014). Sex differences in fluid reasoning: Manifest and latent estimates from the cognitive abilities test. *Journal of Intelligence, 2*(2), 36–55. https://doi.org/10.3390/jintelligence2020036
LeAdelle, P., McGrew, K. S., Knopik, S. N., & Ford, L. (2005). The general (g), broad, and narrow CHC stratum characteristics of the WJ III and WISC-III tests: A confirmatory cross-battery investigation. *School Psychology Quarterly, 20*(1), 66–88. https://doi.org/10.1521/scpq.20.1.66.64191
Lippa, R. A., Collaer, M. L., & Peters, M. (2010). Sex differences in mental rotation and line angle judgments are positively associated with gender equality and economic development across 53 nations. *Archives of Sexual Behavior, 39*(4), 990–997. https://doi.org/10.1007/s10508-008-9460-8
Liu, J., & Lynn, R. (2015). Chinese sex differences in intelligence: Some new evidence. *Personality and Individual Differences, 75*, 90–93. https://doi.org/10.1016/j.paid.2014.11.002
Lubinski, D., Benbow, C. P., McCabe, K. O., & Bernstein, B. O. (2023). Composing meaningful lives: Exceptional women and men at age 50. *Gifted Child Quarterly, 67*(4), 278–305. https://doi.org/10.1177/00169862231175831
Lyle, J. G., & Johnson, E. G. (1974). Analysis of WISC coding: 5. Prediction of coding performance. *Perceptual and Motor Skills, 39*(1), 111–114. https://doi.org/10.2466/pms.1974.39.1.111
Lynn, R. (1994). Sex differences in brain size and intelligence: A paradox resolved. *Personality and Individual Differences, 17*, 257–271. https://doi.org/10.1016/0191-8869(94)90030-2
MacDonald, D. H., & Hewlett, B. S. (1999). Reproductive interests and forager mobility. *Current Anthropology, 40*, 501–523. https://doi.org/10.1086/200047
Maitland, S. B., Herlitz, A., Nyberg, L., Bäckman, L., & Nilsson, L. G. (2004). Selective sex differences in declarative memory. *Memory & Cognition, 32*(7), 1160–1169. https://doi.org/10.3758/bf03196889
McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence, 37*(1), 1–10. https://doi.org/10.1016/j.intell.2008.08.004
McNemar, Q., & Terman, L. M. (1936). Sex differences in variational tendency. *Genetic Psychology Monographs, 18*, 1–65. https://link.ezproxy.neu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=1936-03117-001&site=ehost-live&scope=site.
Pauls, F., Petermann, F., & Lepach, A. C. (2013). Gender differences in episodic memory and visual working memory including the effects of age. *Memory, 21*(7), 857–874. https://doi.org/10.1080/09658211.2013.765892
Quinn, J. M., & Wagner, R. K. (2015). Gender differences in Reading impairment and in the identification of impaired readers. *Journal of Learning Disabilities, 48*(4), 433–445. https://doi.org/10.1177/0022219413508323
R Core Team. (2023). R: A language and environment for statistical computing (3.1.2). R Foundation for Statistical Computing. http://www.r-project.org/.
Rodríguez-Cancino, M., & Concha-Salgado, A. (2023). WISC-V Measurement Invariance According to Sex and Age: Advancing the Understanding of Intergroup Differences in Cognitive Performance. *Journal of Intelligence, 11*(9). https://doi.org/10.3390/jintelligence11090180
Schroeder, D. H. (2021). *Sex differences in variability. Encyclopedia of Evolutionary Psychological Science, December*. https://doi.org/10.1007/978-3-319-19650-3_304590
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
Stoet, G. (2017). Sex differences in the Simon task help to interpret sex differences in selective attention. *Psychological Research, 81*(3), 571–581. https://doi.org/10.1007/s00426-016-0763-4
Thorndike, E. L. (1906). Sex in education. *Bookman, 23*, 211–214.
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
Wechsler, D. (2014). *Wechsler intelligence scale for children-fifth edition*. Pearson.
Willcutt, E. G. (2012). The prevalence of DSM-IV attention-deficit/hyperactivity disorder: A meta-analytic review. *Neurotherapeutics, 9*, 490–499. https://doi.org/10.1007/s13311-012-0135-8
Woodcock, R. W. (1990). Theoretical foundations of the Wj-R measures of cognitive ability. *Journal of Psychoeducational Assessment, 8*(3), 231–258. https://doi.org/10.1177/073428299000800303
Zahavi, A. (1975). Mate selection–a selection for a handicap. *Journal of Theoretical Biology, 53*, 205–214. https://doi.org/10.1016/0022-5193(75)90111-3