Psychon Bull Rev
DOI 10.3758/s13423-016-1085-7

THEORETICAL REVIEW

Sex differences in visual-spatial working memory: A meta-analysis

Daniel Voyer[1] · **Susan D. Voyer**[1] · **Jean Saint-Aubin**[2]

Abstract Visual-spatial working memory measures are widely used in clinical and experimental settings. Furthermore, it has been argued that the male advantage in spatial abilities can be explained by a sex difference in visual-spatial working memory. Therefore, sex differences in visual-spatial working memory have important implication for research, theory, and practice, but they have yet to be quantified. The present meta-analysis quantified the magnitude of sex differences in visual-spatial working memory and examined variables that might moderate them. The analysis used a set of 180 effect sizes from healthy males and females drawn from 98 samples ranging in mean age from 3 to 86 years. Multilevel meta-analysis was used on the overall data set to account for non-independent effect sizes. The data also were analyzed in separate task subgroups by means of multilevel and mixed-effects models. Results showed a small but significant male advantage (mean d = 0.155, 95 % confidence interval = 0.087-0.223). All the tasks produced a male advantage, except for memory for location, where a female advantage emerged. Age of the participants was a significant moderator, indicating that sex differences in visual-spatial working memory appeared first in the 13-17 years age group. Removing memory for location tasks from the sample affected the pattern of significant moderators. The present results indicate a male advantage in visual-spatial working memory, although age and specific task modulate the magnitude and direction of the effects. Implications for clinical applications, cognitive model building, and experimental research are discussed.

Keywords Visual-spatial working memory · Spatial abilities · Human sex differences · Meta-analysis

The existence of cognitive sex[1] differences has been under debate recently, especially in the context that they are generally small, suggesting that similarities between the sexes rather than differences might be the rule in this domain (Hyde, 2005, 2014). Nevertheless, the male advantage in mental rotation is noteworthy in producing medium to large effect sizes in meta-analyses (Cohen's d of 0.73 according to Linn & Petersen, 1985, and 0.56 according to Voyer, Voyer, & Bryden, 1995). When focusing on specific tests, the Vandenberg and Kuse Mental Rotations Test (MRT: Vandenberg & Kuse, 1978) produces the largest effect, with a magnitude of 0.94 reported by Linn and Petersen (1985) and an effect size of 0.67 reported by Voyer et al. (1995). In a more recent analysis, Maeda and Yoon (2013) examined only the Purdue Visualization of Rotations test and reported an overall effect size of 0.57 in favor of males.

Essentially, the existing data consistently point to mental rotation as producing the largest sex difference in cognitive performance. Therefore, it is not surprising that a large number of factors have been considered to account for this sex difference (for a review, see Halpern, 2013). One of the possible explanations that began to receive attention only recently relies on the notion that mental rotation requires the operation

Electronic supplementary material The online version of this article (doi:10.3758/s13423-016-1085-7) contains supplementary material, which is available to authorized users.

✉ Daniel Voyer
voyer@unb.ca

[1] Department of Psychology, University of New Brunswick, PO Box 4400, Fredericton, NB, Canada E3B 5A3

[2] School of Psychology, Université de Moncton, Moncton, NB, Canada

[1] The term "sex" is used throughout as the present study is concerned with biological sex rather than with gender, which is a social construct (Torgrimson & Minson, 2005).

Published online: 29 June 2016

of visual-spatial working memory (Christie et al., 2013; Hyun & Luck, 2007; Kaufman, 2007; Logie, 1995; Prime & Jolicoeur, 2010). In fact, this premise is accepted to the point that Vecchi, Phillips, and Cornoldi (2001) grouped measures of spatial abilities with visual-spatial working memory tasks. In terms of mechanism, Wang and Carr (2014) proposed that the ratio of visual-spatial to verbal working memory capacity determines the strategies used to solve spatial tasks and, in turn, the efficiency of the selected strategies affects performance. Essentially, males, with their better visual-spatial working memory, would select more effective holistic strategies, whereas females with their superior verbal working memory would select less effective analytic strategies, resulting in the observed male advantage (Wang & Carr, 2014). Based on such a premise, Kaufman (2007) hypothesized that sex differences in visual-spatial working memory might underlie the male advantage. He demonstrated this point by showing that visual-spatial working memory completely mediated the relation between sex and a spatial factor composed of two measures of mental rotation (DAT-Spatial Relation, MRT), although the MRT also had a large unique variance component accounted for by sex. Essentially, these data support the notion that visual-spatial working memory plays an important role both in mental rotation and in the sex difference in such tasks (see also Loring-Meier & Halpern, 1999 on this point).

If we are to accept the notion that sex difference account at least in part for the male advantage in mental rotation, we need to demonstrate a consistent male advantage in visual-spatial working memory as well. Data on this point generally support the presence of a male advantage, although contradictory data also exist. Focusing mostly on a review of the evidence based on the Corsi Blocks task, Wang and Carr (2014) concluded on the existence of a male advantage in visual-spatial working memory, which they distinguished from spatial short term memory. In contrast, Duff and Hampson (2001) reported a female advantage in a task that involved the maintenance of location in visual-spatial working memory. However, such a female advantage seems to be an exception in the literature, at least according to Wang and Carr. Nevertheless, considering that there exists no systematic review of the literature concerning sex differences in visual-spatial working memory and no meta-analysis has ever examined that question, it is impossible to draw definite conclusions at this point. Accordingly, the purpose of the present study was to conduct a meta-analysis of sex differences in visual-spatial working memory. The goals of this analysis were to quantify the overall findings as well as to examine potential moderators of these sex differences.

Broad implications

So far, we have emphasized the relevance of potential sex differences in visual-spatial working memory to their relatively narrow implications for spatial abilities research. However, visual-spatial working memory tasks also are used widely for assessment and theory building in clinical (Alonso-Recio, Martín-Plasencia, Loeches-Alonso, & Serrano-Rodrigues, 2014; Barrett, Kelly, Bell, & King, 2008), developmental (Almela, van der Meij, Hidalgo, Villada, & Salvador, 2012; Teixeira, Zachi, Roque, Taub, & Ventura, 2011), and purely experimental settings (Hegarty, Montello, Richardson, Ishikawa, & Lovelace, 2006; Martin & Chaudry, 2014). Thus, visual-spatial working memory applies to many contexts and knowing whether it produces sex differences has important implications for how data are interpreted in these various contexts. Essentially, if we determine that sex differences exist in visual-spatial working memory, it might require an adjustment of norms in clinical settings, consideration of potentially different developmental trajectories for males and females, and additional components to existing theories. Yet, sex as a factor often is ignored in this area of research, as we discovered in our literature search. This means that the implications of the present paper go beyond establishing a potential link between sex differences in visual-spatial working memory and spatial ability. This paper also has implications for how we use visual-spatial working memory in clinical, developmental, and experimental settings.

With this in mind, we will now proceed to define specifically what we mean by "visual-spatial working memory." This will be followed by the identification of potential moderators that have found support in the literature.

Defining visual-spatial working memory

One issue that arises when attempting to define visual-spatial working memory is that there are as many definitions of this concept as there are theories used to explain its functioning. Wang and Carr (2014) present an excellent summary of the various theoretical perspectives and their implied definitions for visual-spatial working memory. However, our goal in defining visual-spatial working memory was to stay away from any specific theory while remaining broad in the inclusiveness of our definition. Therefore, throughout the present article, we use the term visual-spatial working memory in a theory-neutral manner to refer to the processes involved in the storage of spatial or visual information over a limited period of time. The studies that we sampled in our meta-analysis reflect this general definition.

Identifying potential moderators

The identification of potential moderators is a crucial step in a meta-analysis (Borenstein, Hedges, Higgins, & Rothstein, 2009), although it also is important to limit the number of moderators considered to reduce the risk of Type 1 error in hypothesis testing (Lipsey & Wilson, 2001). This is why

Lipsey and Wilson (2001) suggested that the identification of moderators should be based on what past researchers have considered as important in their empirical studies. Accordingly, the set of potential moderators that we have identified in what follows stems from past literature.

Task The specific task used in a study is an obvious choice as a moderator considering the contradictory findings as a function of task mentioned earlier. Specifically, as we have seen, Wang and Carr (2014) found a male advantage in reviewing the evidence available for the Corsi Blocks tasks, but they also noted the female advantage observed by Duff and Hampson (2001) in a task focusing on location.

The Corsi Blocks task is a fairly common measure of visual-spatial working memory and it involves many variations presented under different names. In the classic implementation of this task, nine blocks are placed randomly in front of the participant and the experimenter taps a number of them on each trial. Participants are asked to reproduce the order in which the blocks were tapped. The maximum number of objects one can tap correctly is the measure of span. Although some might view the Corsi task as a measure of location memory, the fact that the blocks have to be tapped in the correct order also is crucial. Therefore, the Corsi task has both a location and sequencing component. Accordingly, we categorized any task that involved such a location and sequencing component under the Corsi Blocks label, even though the authors of the specific study might have labelled it otherwise.

In contrast, some researchers have used pure location tasks as their measure of visual-spatial working memory. For example, Duff and Hampson (2001) asked their participants to remember the location of matched pairs of stimuli. In this task, only location mattered, not order of recall. In another variation of the location memory task, participants are asked to associate a location with a specific pattern and they then have to remember where a centrally presented pattern was presented (Flannery et al., 2007). This requirement to recall the pattern as well as the location adds a level of complexity to the task that distinguishes it from a pure location task. Accordingly, it was coded as a distinct task in the moderator analysis. The Cambridge Neuropsychological Test Automated Battery (CANTAB) Spatial Working Memory test is a variant of the location task in that participants have to locate one token in an array of boxes presented on a computer. However, because they are told that the token will be presented in a different box on each trial, participants have to recall where they found items on previous trials. Therefore, the location aspect is complicated by the need to remember an increasing number of distractor locations as one progresses through the task. Therefore, this task was coded as a distinct category under the label "token." The n-back task also is a well-known measure of working memory, although it might be more readily associated with verbal working memory. In an example of a visual-spatial working memory version of the n-back task, Kalmady et al. (2013) asked participants to identify the location of the dot that changed color immediately after it was presented (0 back) or two trials before (2 back). Considering that the direction of sex differences can vary as a function of task, specific task is a potentially important moderator of sex differences in visual-spatial working memory that was coded in the present analysis.

Because location seems to be a central component of so many tasks, it is important to distinguish the studies sampled here from the ones that were sampled by Voyer, Postma, Brake, and Imperato-McGinley (2007) in their meta-analysis of sex differences in object location memory. Their analysis focused mostly on tasks similar to the one proposed by Silverman and Eals (1992), in which a large array of objects is memorized and, after an intervening object identity memory task, participants are tasked to identify moved and unmoved objects. Essentially, the time interval between encoding and retrieval is typically too long and the number of objects is too numerous to fit within the limits of working memory (Cowan, 2008). Therefore, the location tasks included in the present analysis are distinct from the type of task that was discussed by Voyer et al. (2007).

Age of participants Age of participants is an obvious potential moderator in any meta-analyses of sex differences in cognitive performance as cognitive abilities are well known to change across the life span (Techentin, Voyer, & Voyer, 2014). Visual-spatial working memory is no exception considering that De Luca et al. (2003) showed clear changes in visual-spatial working memory capacity across the life span. Accordingly, age of participants was considered as a potential moderator in the present analysis.

Medium of presentation The medium used for stimulus presentation method (e.g., computer or physical material as in the original Corsi Blocks) could account for some variance that is not relevant to visual-spatial working memory. In particular, there is some evidence suggesting that males feel more positively about computers and that they are more comfortable in their use than females (Cooper, 2006; Kay, 2006). This sex difference still persists in more recent studies, despite the high exposure of both sexes to computers in our modern world (Scherer & Siddiq, 2015; Sieverding & Koch, 2009). If this moderator proves significant, it will suggest the presence of a third variable in the relation between sex and visual-spatial working memory. However, it is important to keep in mind that specific task and testing medium are confounded to some extent. For example, the CANTAB measures of visual-spatial working memory are always computerized, whereas the Corsi Blocks test can be presented physically or on computer.

Findings of an effect of testing medium will thus require cautious interpretation, especially if an effect of task also is found.

Stimulus type The specific stimuli used in a given task typically are a constant within a given study, but they tend to vary across tasks. For example, blocks, squares, or boxes are by far the most common stimuli (Coluccia & Martello, 2004). The use of dots or circles is quite common (Duff & Hampson, 2001; Evardone & Alexander, 2009). At this point, however, it seems that whether stimulus type has an impact in terms of performance or magnitude of sex differences remains an empirical question. The need to consider this variable arises from the possibility that verbalizable content might affect the result of individual studies (Lejbak, Vrbancic, & Crossley, 2009). For example, Lejbak et al. (2009) used not only dots and circles but also common objects as stimuli, whereas Seghete, Cservenka, Herting, & Nagel (2013) used alphanumeric characters as stimuli. Both of these studies produced, at least, a trend for a female advantage in their task. As these findings suggest, the possibility that the use of verbalizable stimuli might affect the magnitude or direction of sex differences in visual-spatial working memory, stimulus type was coded as a potential moderator.

Type of memory task: the distinction between recall and recognition The distinction between recall and recognition is crucial in memory research as these two processes sometimes can produce opposite effects. For example, in the context of verbal memory, MacLeod and Kampe (1996) reported poorer recognition of high-frequency than low-frequency words for recognition but no word frequency effect for recall. Recall and recognition also are believed to involve somewhat different cerebral networks, at least in the context of episodic memory (Cabeza et al, 1997), and they often are presented as entailing different cognitive processes (Johnson, 2013). Although there seems to be a paucity of research examining the distinction between these two processes in visual-spatial working memory, empirical and theoretical evidence suggest that it should be considered as a moderator in any meta-analysis examining a memory task. Therefore, this distinction was coded as a variable reflecting the type of memory task. However, in this case as well, specific task and instruction often are confounded. For example, the Corsi Blocks task usually involves recall, whereas pattern recognition obviously involves recognition. Therefore, a significant effect of this moderator will call for a closer look at the actual data.

Dependent variable Research on visual-spatial working memory often uses memory span or accuracy of responses as a dependent variable and these could be seen as reflecting different process. Essentially, memory span directly quantifies memory capacity, whereas accuracy has been presented as a subcomponent of memory span (Unsworth, Redick, Heitz, Broadway, & Engle, 2009). Unsworth et al. (2009) also found that processing time is correlated with memory span. Therefore, it appears that span, accuracy, and response time measure different components of memory capacity. In fact, they can produce different results in the context of sex differences in visual-spatial working memory. For example, Lejbak, Crossley, & Vrbancic (2011) reported a significant male advantage in an n-back task for accuracy but not for response time. Accordingly, as studies of visual-spatial working memory typically use at least one of these dependent variables as a measure of performance, this factor was coded as a potential moderator. However, here as well, specific task and dependent variable can be confounded. For example, the Corsi Blocks task typically produces a measure of memory span, whereas n-back tasks can involve accuracy and response time as dependent variables but not span. This means that, in this case as well, a closer look at the actual data might be required.

Current meta-analysis

Visual-spatial working memory is ubiquitous as a measure of general functioning in clinical and lifespan development settings. In addition, it often is viewed as an important component of any theory of human memory. Therefore, the present meta-analysis concerns the question of whether sex of participants should be considered as a relevant factor when investigating visual-spatial working memory in clinical and experimental settings. Furthermore, the current meta-analysis addresses whether sex differences in visual-spatial working memory warrant their hypothesized role in accounting for sex differences in spatial ability, especially mental rotation. Essentially, a comparison of the magnitude of sex differences in visual-spatial working memory with that for spatial performance will provide an indirect measure of their importance in accounting for the latter effects. Moreover, the examination of potential moderators of these sex differences should allow some speculations concerning underlying factors while also providing directions for future research. Therefore, quantifying sex differences in visual-spatial working memory and the identification of moderators of the effect sizes constituted the two primary goals of the present analysis.

To achieve these goals, a two stage process was used, similar to the approach proposed by Voyer and Voyer (2014) in the analysis of sex differences in scholastic achievement. Specifically, because many researchers used a variety of relevant measures of visual-spatial working memory in their design, the typical fixed effect or random effect meta-analysis would require collapsing across these tests or randomly selecting one effect size to avoid violation of the homogeneity of effect sizes assumption (Borenstein et al., 2009). Accordingly, the present analysis relied on the multilevel modeling approach to meta-analysis (Hox, 2008;

Raudenbush & Bryk, 2002), because it makes no assumption concerning independence of effects and it applies readily to the hierarchical structure of a meta-analysis. Therefore, as a first step, computation of the overall effect size and of the moderator analysis relied on this approach.

The second step in data analysis considered the fact that many of the moderator variables that were discussed earlier are confounded with specific task. Therefore, this second step involved the application of mixed-effects meta-analysis (when all effects sizes were independent or did not include enough samples to compute robust standard errors (Raudenbush & Bryk, 2002) or multilevel modeling (in the case of non-independent effect sizes with enough samples to compute robust standard errors) to examine the influence of moderator variables within each specific task.

This combination of approaches to meta-analysis should allow more fine grained conclusions to be reached from the data. In this way, the analytic approach proposed maximizes the contribution of the present meta-analysis.

Method

Study selection

Studies retrieval was initially performed through a search for research in the PsycINFO, ERIC, and Dissertation & Theses databases. The option "include related terms" was turned on for all database searches. The searches also included foreign language articles as the databases provided an English abstract for all of them. The first search was conducted in November 2014 and subsequent searches were limited to that date. The initial search used the terms *spatial* and *immediate memory* or *working memory* or *short term memory*. The total number of hits was 8,848 for this original search. After removing non-human research, literature reviews, and research summaries, the number of relevant hits was reduced to 2,072 and coding started. Although this search was as inclusive as possible, a major obstacle quickly became obvious. Specifically, in the vast majority of this research, sex was completely ignored as a possible influence on performance. In particular, the number of men and women tested was rarely presented and performance was not divided by sex. In fact, it was impossible to tell for most of these papers whether they had actually tested participants of both sexes. After attempting to code half of the relevant hits, nearly 1,000 authors had to be contacted to obtain information relevant to our research question. This state of affairs suggests that adherence to the Journal Articles Reporting Standard elaborated by the American Psychological Association (APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008) is very poor in this literature—at least as far as report of the sex composition of the sample is concerned (Voyer & Voyer, 2015). In addition, this suggests that many researchers involved in visual-spatial working memory research do not view sex as a relevant factor. In fact, it is noteworthy that none of the articles appearing in a recent special issue of Attention, Perception, & Psychophysics on visual-spatial memory (October 2014, Volume 76, Issue 7) reported a preliminary analysis investigating sex differences.

This generalized lack of report for information crucial to our research questions led us to a change of search strategy. Specifically, we conducted a second search with *sex* or *gender* and *spatial* and *immediate memory* or *working memory* or *short term memory* as search terms. This search resulted in 330 hits after removal of duplicates and this was reduced to 194 after removal of non-human research, reviews, and summaries. Of course, the inclusion of *sex* or *gender* in the search terms might have limited the hits to those where sex differences were central to the research question. However, even then, information pertaining to sex was still missing from some of the articles as this factor was actually not the focus for much of the retrieved research. Furthermore, we reasoned that if this literature shows significant sex differences after taking publication bias into account (see below), it will indicate that researchers who completely disregard sex in their visual-spatial working memory research might be overlooking a meaningful factor.

Considering our concerns that the inclusion of *sex* or *gender* in the search terms might result in a sample biased toward studies showing sex differences, we made additional efforts to obtain unpublished research. Specifically, theses and dissertations were considered as a possible source of unpublished material. Furthermore, a posting requesting unpublished research was sent to the mailing list of the Spatial Learning Network (SILC) and of the Canadian Society for Brain, Behavior, and Cognitive Science (CSBBCS). As a result of these efforts, 18 effect sizes from unpublished research (16 theses, 2 unpublished papers) were coded into the data sample (from a total of 182 effect sizes). Therefore, the final data set included a small number of effect sizes drawn from unpublished work.

Selection criteria

A number of selection criteria allowed us to determine whether a study could be included in the meta-analysis. Accordingly, as a first step, the second author carefully read the abstract for each study to determine if the inclusion criteria were met. When fit with the inclusion criteria remained unclear after consultation of the abstract, the actual paper was consulted by the three authors.

The specific criteria used in making inclusion decisions required studies to have both male and female participants. A study had to report on at least one task that reflected a pure measure of visual-spatial working memory. Determining whether a task measured visual-spatial working memory was

based on the theory-neutral definition we presented earlier: Visual-spatial working memory refers to the processes involved in the storage of spatial or visual information over a limited period of time.

Of course, the studies had to include relevant data for calculating the effect size (see the "measure of effect size" section) to be included in the meta-analysis. When information was missing and the study was published in 2004 or more recently, the first author was contacted. We expected that a publication year of 2004 would be recent enough so that the authors might still have access to the data. Twenty authors had to be contacted in this manner, and five of them provided data that allowed us to include their study in the meta-analysis.

A number of exclusion criteria were also defined. Specifically, data based on special populations were not included. For example, studies that examined individuals with brain damage were not included, although when such studies reported on a control group that met the other criteria, the control group data were included. However, control groups that were sex-matched to a clinical group were excluded as they did not form random samples. Finally, to avoid duplication, when data from the same sample were reported in multiple articles, only the first report on these data was included.

Reference lists from papers retrieved in the database search also were used to identify additional relevant studies. The data collection window ended in November 2014 and resulted in a sample of 182 effects sizes from 98 samples from 69 different articles after the application of the inclusion/exclusion criteria. A summary of the studies included in the final sample are presented in Table 1. In addition, forest plots are presented separately for each task in Figs. 1, 2, 3, 4, 5, and 6.

Coding of variables

A number of variables were coded as factors that might moderate sex differences in visual-spatial working memory. Specifically, characteristics relevant to the samples themselves (sample level variables) and factors inherent to the tasks used in each study (measure level variables) were considered.

Sample level variables Mean age of the participants in a sample was included both as a continuous and a categorical variable. When considered categorically, five groups were defined: aged younger than 13 years (29 samples), between 13 and 17 years (inclusive; 10 samples), between 18 and 29 years (inclusive; 36 samples), between 30 and 49 years (inclusive; 10 samples), and above 49 years (13 samples). These categories were arbitrary, but we hoped that they would capture various periods of development reasonably well. It is important to note that mean age of the sample was not always reported in the retrieved studies. However, when the school grade was given, the age variable was coded using the approach proposed by Voyer et al. (1995). For example, in North America, children in grade 1 are typically 6 years old, whereas first-year undergraduate students are usually 19 years old.

Because year of publication was coded routinely, it was included as a potential moderator. This factor often is interpreted as an indirect way to assess how social changes might promote fluctuations in sex differences (Feingold, 1988).

Measure level variables The specific task used is likely the most obvious measure level characteristic. Accordingly, tasks were classified as belonging in the following categories: Corsi Blocks task or equivalent (k = 70 effect sizes), n-back task (k = 19), memory for patterns (k = 37), memory for location (k = 26), and memory for a token (k = 21). In addition, other tasks that clearly involved visual-spatial working memory but did not fit in any of the already defined categories were classified as "other tasks" (k = 9).

Type of memory task was coded as a measure level variable. Therefore, whether the task involved recall (k = 126) or recognition (k = 25) was noted. In some cases, this distinction was not relevant, resulting in 31 effect sizes coded as "not relevant." This last category reflects tasks in which, for example, participants have to keep track of where they already looked as they searched for a specific token in an array of virtual boxes.

Testing medium was coded. This moderator consisted of three categories reflecting computer (k = 117), physical medium, including paper, cards, and blocks (k = 55), and medium not reported (k = 10).

Stimulus type also was coded as a measure level variable. This moderator was categorized as blocks/squares/boxes (k = 91), dots/circles (k = 32), geometric patterns (k = 19), lines/arrows (k = 9), verbalizable shapes and objects (k = 13), alphanumeric characters (k = 6), and stimuli not reported (k = 12).

Finally, whether the dependent variable was a measure of memory span (k = 58), number of trials to criterion (k = 13), search errors (k = 35), accuracy (k = 41), response time (k = 30), and other or not reported (k = 5) was coded.

A number of steps were taken to warrant the validity of coding. As a starting point, we prepared a coding sheet that included an entry for all coded variables. A subset of 17 studies (accounting for 51 effect sizes) was coded independently by the first and second authors, two experienced meta-analysts. This coding involved 19 variables, although they were not all used in the moderator analysis. Specifically, the coded variables were: sample ID (required for multilevel analysis), authors, year of publication, publication status (published or not), mean age of sample, national origin, number of males, number of females, task, instructions, stimuli, target feature, response medium, delay, interference task, testing medium, dependent variable, type of memory, and effect size. Therefore, a total of 969 entries (19 variables x 51 effect sizes) produced only 8 disagreements, resulting in an inter-rater

Table 1 Studies included in the present analysis

Authors	Year	*Nm*	*Nf*	Age	Task	Type	Stimuli	Medium	Dependent variable	*d*	*SE*
Aarnoudse-Moens et al.	2012	87	103	8.3	Corsi	recall	blocks	computer	span	0.00	0.15
Almela et al.	2012	44	44	63.0	Token	NA	blocks	computer	errors	0.04	0.21
Alonso-Recio et al.	2014	24	25	64.9	N-back	recognition	blocks	computer	accuracy	-0.03	0.29
Barrett et al.	2008	12	14	44.5	Token	NA	blocks	computer	errors	0.62	0.40
Barrett et al.	2008	12	14	44.5	Token	NA	blocks	computer	errors	0.72	0.41
Bosco et al.	2004	53	54	22.5	Corsi	recall	blocks	physical	span	0.44	0.20
Bosco et al.	2004	53	54	22.5	Pattern	recall	blocks	physical	span	0.66	0.20
Breitberg et al.	2013	26	18	29.0	Corsi	recall	blocks	computer	span	0.35	0.31
Bücker et al.	2014	39	59	22.5	Token	NA	blocks	computer	errors	0.41	0.21
Caldwell et al.	2005	12	9	16.4	Pattern	recognition	lines	computer	accuracy	0.36	0.44
Caldwell et al.	2005	12	9	16.4	Pattern	recognition	lines	computer	RT	0.42	0.45
Cansino et al.	2013	750	750	50.0	N-back	recall	circles	computer	accuracy	0.44	0.05
Cansino et al.	2013	750	750	50.0	N-back	recall	circles	computer	RT	0.52	0.05
Capitani et al.	1991	229	266	53.3	Corsi	recall	blocks	physical	span	0.25	0.09
Casey et al.	2011	64	60	9.0	Corsi	recall	blocks	computer	trials	0.21	0.18
Colom et al.	2013	45	59	19.9	Pattern	recall	circles	computer	accuracy	0.00	0.20
Coluccia & Martello	2004	53	57	22.5	Corsi	recall	blocks	physical	span	0.43	0.19
De Luca et al.	2003	13	16	9.7	Corsi	recall	blocks	computer	span	0.14	0.37
De Luca et al.	2003	13	16	12.9	Corsi	recall	blocks	computer	span	0.33	0.38
De Luca et al.	2003	21	18	17.7	Corsi	recall	blocks	computer	span	0.78	0.33
De Luca et al.	2003	19	20	24.4	Corsi	recall	blocks	computer	span	0.00	0.32
De Luca et al.	2003	21	18	38.8	Corsi	recall	blocks	computer	span	0.61	0.33
De Luca et al.	2003	6	13	55.9	Corsi	recall	blocks	computer	span	0.46	0.50
De Luca et al.	2003	13	16	9.7	Token	NA	blocks	computer	errors	-0.04	0.37
De Luca et al.	2003	13	16	12.9	Token	NA	blocks	computer	errors	0.73	0.39
De Luca et al.	2003	21	18	17.7	Token	NA	blocks	computer	errors	0.81	0.33
De Luca et al.	2003	19	20	24.4	Token	NA	blocks	computer	errors	0.97	0.34
De Luca et al.	2003	21	18	38.8	Token	NA	blocks	computer	errors	0.39	0.32
De Luca et al.	2003	6	13	55.9	Token	NA	blocks	computer	errors	0.28	0.50
De Luca et al.	2003	13	16	9.7	Token	NA	blocks	computer	errors	-0.46	0.38
De Luca et al.	2003	13	16	12.9	Token	NA	blocks	computer	errors	-0.19	0.37
De Luca et al.	2003	21	18	17.7	Token	NA	blocks	computer	errors	0.23	0.32
De Luca et al.	2003	19	20	24.4	Token	NA	blocks	computer	errors	0.20	0.32
De Luca et al.	2003	21	18	38.8	Token	NA	blocks	computer	errors	0.51	0.33
De Luca et al.	2003	6	13	55.9	Token	NA	blocks	computer	errors	0.05	0.49
Duff & Hampson	2001	46	44	20.0	Location	NA	circles	physical	errors	-0.64	0.22
Duff & Hampson	2001	44	44	20.8	Location	NA	circles	physical	errors	-0.75	0.22
Duff & Hampson	2001	46	46	21.0	Location	NA	geometric	physical	errors	-0.75	0.22
Duff & Hampson	2001	44	44	20.8	Location	NA	circles	physical	RT	-0.36	0.21
Duff & Hampson	2001	46	44	20.0	Location	NA	circles	physical	RT	-0.46	0.21
Duff & Hampson	2001	46	46	21.0	Location	NA	geometric	physical	RT	-0.73	0.22
Evardone & Alexander	2009	55	50	20.0	Location	recall	circles	physical	errors	-0.26	0.20
Fikke et al.	2011	10	25	14.7	Token	NA	blocks	computer	errors	0.53	0.38
Flannery et al.	2007	48	20	32.9	Pattern	recall	geometric	computer	errors	-0.30	0.27
Flannery et al.	2007	48	20	32.9	Pattern	recall	geometric	computer	accuracy	-0.05	0.27
Flannery et al.	2007	48	20	32.9	Pattern	recall	geometric	computer	RT	-0.59	0.27
Fournet et al.	2012	89	125	61.0	Corsi	recall	blocks	computer	trials	0.23	0.14
Fournet et al.	2012	89	125	61.0	Corsi	recall	blocks	computer	trials	0.15	0.14
Fournet et al.	2012	65	73	70.7	Corsi	recall	blocks	computer	trials	0.38	0.17

Table 1 (continued)

Authors	Year	*Nm*	*Nf*	Age	Task	Type	Stimuli	Medium	Dependent variable	*d*	*SE*
Fournet et al.	2012	65	73	70.7	Corsi	recall	blocks	computer	trials	0.11	0.17
Fournet et al.	2012	43	54	79.0	Corsi	recall	blocks	computer	trials	0.40	0.21
Fournet et al.	2012	43	54	79.0	Corsi	recall	blocks	computer	trials	0.26	0.21
Fournet et al.	2012	89	125	61.0	Pattern	recall	blocks	computer	trials	0.33	0.14
Fournet et al.	2012	65	73	70.7	Pattern	recall	blocks	computer	trials	0.67	0.18
Fournet et al.	2012	43	54	79.0	Pattern	recall	blocks	computer	trials	-0.04	0.20
Geiger & Litwiller	2005	9	28	19.0	Other	recall	alpha	physical	span	0.73	0.39
Girard	2014	6	8	34.6	Corsi	recall	blocks	computer	accuracy	0.50	0.55
Girard et al	2010	6	15	34.3	Corsi	recall	blocks	physical	span	-0.14	0.48
Girard et al	2010	6	15	34.3	Corsi	recall	blocks	physical	span	-0.24	0.48
Hampson & Morley	2013	31	39	21.6	Location	NA	circles	physical	errors	-0.10	0.24
Hampson & Morley	2013	31	39	21.6	Location	NA	circles	physical	RT	-0.32	0.24
Hartley et al.	2004	6	7	19.0	Token	NA	blocks	computer	errors	0.60	0.57
Hartley et al.*	2004	6	7	19.0	Location	recognition	blocks	computer	accuracy	2.60*	0.76
Hartley et al.	2004	6	7	19.0	Pattern	recognition	geometric	computer	accuracy	0.63	0.57
Hartley et al.	2004	6	7	19.0	Location	recognition	blocks	computer	RT	0.86	0.58
Hartley et al.*	2004	6	7	19.0	Pattern	recognition	geometric	computer	RT	1.93*	0.67
Hayward	2014	17	21	20.2	Other	recall	verbal	computer	accuracy	0.89	0.34
Hegarty et al.	2006	83	135	22.0	Other	recall	lines	computer	accuracy	0.40	0.14
Hernández-Balderas et al.	2012	14	14	9.6	Pattern	recognition	circles	computer	accuracy	0.16	0.38
Hoesing	1998	14	26	19.0	Pattern	recognition	geometric	computer	errors	-0.12	0.33
Hoesing	1998	14	26	19.0	Pattern	recognition	circles	computer	RT	-0.29	0.33
Hoesing	1998	14	26	19.0	Pattern	recognition	geometric	computer	RT	0.49	0.34
Kalmady et al	2013	14	11	25.5	N-back	recall	circles	computer	accuracy	0.23	0.40
Kalmady et al	2013	14	11	25.5	N-back	recall	circles	computer	accuracy	-0.35	0.41
Kalmady et al	2013	14	11	25.5	N-back	recall	circles	computer	RT	0.48	0.41
Kalmady et al	2013	14	11	25.5	N-back	recall	circles	computer	RT	0.38	0.41
Kaufman	2007	50	50	17.0	Corsi	recall	blocks	computer	accuracy	0.67	0.21
Kaufman	2007	50	50	17.0	Corsi	recall	blocks	computer	accuracy	0.63	0.20
Kaufman	2007	50	50	17.0	Corsi	recall	blocks	computer	accuracy	0.22	0.20
Kokubo et al.	2012	44	50	17.0	Other	NA	alpha	computer	RT	-0.16	0.21
Kokubo et al.	2012	44	50	17.0	Other	NA	alpha	computer	RT	0.08	0.21
Krikorian et al.	1996	119	160	18.0	Pattern	recall	circles	physical	trials	0.31	0.12
Krinzinger et al.	2012	60	80	7.5	Corsi	recall	blocks	physical	accuracy	-0.05	0.17
Krinzinger et al.	2012	60	80	7.5	Corsi	recall	blocks	physical	accuracy	0.00	0.17
Kuhn & Holling	2014	27	32	9.0	Pattern	recall	geometric	computer	accuracy	0.22	0.26
Lawton & Hatcher	2005	72	209	22.9	Other	recognition	verbal	computer	accuracy	0.42	0.14
Lawton & Hatcher	2005	72	209	22.9	Other	recognition	verbal	computer	RT	0.41	0.14
Lejbak et al.	2009	20	20	19.0	Location	recall	verbal	computer	errors	-0.58	0.32
Lejbak et al.	2009	20	20	19.0	Location	recall	verbal	physical	errors	-0.62	0.32
Lejbak et al.	2009	20	20	19.0	Location	recall	verbal	computer	errors	-0.28	0.32
Lejbak et al.	2009	20	20	19.0	Location	recall	verbal	physical	errors	-0.77	0.33
Lejbak et al.	2009	20	20	19.0	Location	recall	geometric	computer	errors	-0.29	0.32
Lejbak et al.	2009	20	20	19.0	Location	recall	geometric	physical	errors	-0.60	0.32
Lejbak et al.	2009	20	20	19.0	Location	recall	verbal	computer	RT	-0.10	0.32
Lejbak et al.	2009	20	20	19.0	Location	recall	verbal	physical	RT	-0.46	0.32
Lejbak et al.	2009	20	20	19.0	Location	recall	verbal	computer	RT	0.08	0.32
Lejbak et al.	2009	20	20	19.0	Location	recall	verbal	physical	RT	-0.17	0.32
Lejbak et al.	2009	20	20	19.0	Location	recall	geometric	computer	RT	0.45	0.32

Table 1 (continued)

Authors	Year	*Nm*	*Nf*	Age	Task	Type	Stimuli	Medium	Dependent variable	*d*	*SE*
Lejbak et al.	2009	20	20	19.0	Location	recall	geometric	physical	RT	-0.32	0.32
Lejbak et al.	2009	18	18	18.6	N-back	recognition	circles	computer	accuracy	1.06	0.36
Lejbak et al.	2011	18	18	18.6	N-back	recognition	circles	computer	RT	0.00	0.33
León et al.	2014	10	10	4.0	Corsi	recall	nr	NR	span	-0.50	0.45
León et al.	2014	10	10	5.0	Corsi	recall	nr	NR	span	-0.85	0.47
León et al.	2014	10	10	6.0	Corsi	recall	nr	NR	span	0.00	0.45
León et al.	2014	10	10	7.5	Corsi	recall	nr	NR	span	0.50	0.45
León et al.	2014	10	10	9.5	Corsi	recall	nr	NR	span	-0.49	0.45
León et al.	2014	10	10	4.0	Corsi	recall	nr	NR	span	0.74	0.46
León et al.	2014	10	10	5.0	Corsi	recall	nr	NR	span	0.24	0.45
León et al.	2014	10	10	6.0	Corsi	recall	nr	NR	span	0.84	0.47
León et al.	2014	10	10	7.5	Corsi	recall	nr	NR	span	-0.19	0.45
León et al.	2014	10	10	9.5	Corsi	recall	nr	NR	span	-0.11	0.45
Levin et al.	2005	35	32	20.7	Location	recall	geometric	computer	accuracy	0.00	0.24
Levin et al.	2005	6	5	20.7	Location	recall	geometric	computer	accuracy	0.00	0.61
Levin et al.	2005	35	32	20.7	Location	recall	geometric	computer	RT	0.00	0.24
Levin et al.	2005	6	5	20.7	Location	recall	geometric	computer	RT	0.00	0.61
Mammarella et al.	2010	11	10	9.0	Pattern	recognition	lines	computer	other	0.28	0.44
Mammarella et al.	2010	11	10	9.0	Pattern	recognition	circles	computer	other	0.20	0.44
Mammarella et al.	2010	11	10	9.0	Pattern	recognition	circles	computer	other	-0.11	0.44
Martin & Chaudry	2014	41	45	20.0	Corsi	recall	blocks	physical	trials	-0.59	0.22
Miller	2003	40	40	21.8	Pattern	recognition	blocks	computer	span	-0.25	0.22
Miller	2003	40	40	21.8	Other	recall	alpha	computer	span	-0.27	0.22
Miller & Halpern	2013	49	28	18.2	Other	recall	lines	computer	accuracy	0.12	0.24
Minor & Park	1999	107	106	27.3	Pattern	recall	circles	computer	accuracy	0.05	0.14
Nalçaci et al.	1997	60	61	20.0	Pattern	recall	blocks	computer	accuracy	0.36	0.18
Nalçaci et al.	1997	60	61	20.0	Pattern	recall	blocks	computer	RT	0.66	0.19
Nalçaci et al.	2000	66	32	20.0	Pattern	recall	blocks	computer	accuracy	0.55	0.22
Orsini et al.	1986	127	111	24.5	Corsi	recall	blocks	physical	span	0.48	0.13
Orsini et al.	1986	118	124	34.5	Corsi	recall	blocks	physical	span	0.35	0.13
Orsini et al.	1986	64	102	44.5	Corsi	recall	blocks	physical	span	0.47	0.16
Orsini et al.	1986	70	99	54.5	Corsi	recall	blocks	physical	span	0.33	0.16
Orsini et al.	1986	72	74	64.5	Corsi	recall	blocks	physical	span	0.15	0.17
Orsini et al.	1986	128	140	74.5	Corsi	recall	blocks	physical	span	0.21	0.12
Orsini et al.	1986	53	72	85.7	Corsi	recall	blocks	physical	span	0.31	0.18
Pagulayan et al.	2006	148	192	18.0	Corsi	recall	blocks	physical	span	0.06	0.11
Pangelinan et al.	2011	68	104	9.1	Token	NA	blocks	computer	errors	0.00	0.16
Piccardi et al.	2014	16	21	4.6	Corsi	recall	blocks	physical	span	0.01	0.33
Piccardi et al.	2014	29	21	5.6	Corsi	recall	blocks	physical	span	1.29	0.31
Piccardi et al.	2014	22	28	6.4	Corsi	recall	blocks	physical	span	0.26	0.29
Piccardi et al.	2014	25	26	7.4	Corsi	recall	blocks	physical	span	-0.60	0.29
Piccardi et al.	2014	23	19	8.6	Corsi	recall	blocks	physical	span	0.05	0.31
Piccardi et al.	2014	17	21	10.2	Corsi	recall	blocks	physical	span	-0.05	0.33
Piccardi et al.	2014	16	21	4.6	Corsi	recall	blocks	physical	span	0.04	0.33
Piccardi et al.	2014	29	21	5.6	Corsi	recall	blocks	physical	span	0.37	0.29
Piccardi et al.	2014	22	28	6.4	Corsi	recall	blocks	physical	span	-0.31	0.29
Piccardi et al.	2014	25	26	7.4	Corsi	recall	blocks	physical	span	-0.54	0.29
Piccardi et al.	2014	23	19	8.6	Corsi	recall	blocks	physical	span	-1.08	0.33
Piccardi et al.	2014	17	21	10.2	Corsi	recall	blocks	physical	span	-0.03	0.33

Table 1 (continued)

Authors	Year	*N*m	*N*f	Age	Task	Type	Stimuli	Medium	Dependent variable	*d*	*SE*
Postma et al.	2004	32	32	21.4	Corsi	recall	blocks	physical	span	0.45	0.25
Postma et al.	1999	23	34	23.0	Pattern	recall	verbal	computer	other	0.71	0.28
Price	2009	20	16	2.9	Token	NA	blocks	physical	errors	0.18	0.34
Roesch-Ely et al.	2009	20	20	33.5	Pattern	recall	circles	computer	other	-0.34	0.32
Rubin	2009	27	30	29.0	Pattern	recall	nr	computer	accuracy	0.02	0.27
Rubin	2009	27	30	29.0	Pattern	recall	nr	computer	accuracy	-0.26	0.27
Ruggiero et al.	2008	30	30	23.5	Corsi	recall	blocks	physical	span	0.35	0.26
Ruggiero et al.	2008	16	16	23.8	Corsi	recall	blocks	physical	span	0.35	0.36
Ruggiero et al.	2008	16	16	64.7	Corsi	recall	blocks	physical	span	0.29	0.36
Ruggiero et al.	2008	30	30	23.5	Corsi	recall	blocks	physical	span	0.18	0.26
Ruggiero et al.	2008	16	16	23.8	Corsi	recall	blocks	physical	span	0.56	0.36
Ruggiero et al.	2008	16	16	64.7	Corsi	recall	blocks	physical	span	0.70	0.36
Savage	2013	25	56	46.9	Pattern	recall	circles	computer	accuracy	0.37	0.24
Savage	2013	25	56	46.9	Pattern	recall	circles	computer	accuracy	0.60	0.25
Schweinsburg et al.	2005	24	25	14.8	N-back	recognition	lines	computer	accuracy	0.05	0.29
Schweinsburg et al.	2005	24	25	14.8	N-back	recognition	lines	computer	RT	0.29	0.29
Seghete et al.	2013	19	15	14.0	N-back	recall	alpha	computer	accuracy	-0.47	0.35
Seghete et al.	2013	19	15	14.0	N-back	recall	alpha	computer	RT	-0.05	0.35
Shikhman	2007	174	229	30.0	N-back	recall	circles	computer	accuracy	0.22	0.10
Shikhman	2007	174	229	30.0	N-back	recall	circles	computer	accuracy	0.13	0.10
Shikhman	2007	174	229	30.0	N-back	recall	circles	computer	accuracy	0.13	0.10
Shikhman	2007	174	229	30.0	N-back	recall	circles	computer	accuracy	-0.05	0.10
Squeglia et al.	2011	31	24	17.8	N-back	recognition	lines	computer	accuracy	0.06	0.27
Squeglia et al.	2011	31	24	17.8	N-back	recognition	lines	computer	RT	0.23	0.27
Szabo et al.	2011	53	105	66.5	Pattern	recognition	circles	computer	accuracy	0.28	0.17
Szabo et al.	2011	53	105	66.5	Pattern	recognition	circles	computer	RT	0.69	0.17
Teixeira et al.	2011	8	10	6.5	Corsi	recall	blocks	computer	span	0.45	0.48
Teixeira et al.	2011	9	19	9.0	Corsi	recall	blocks	computer	span	-0.41	0.41
Teixeira et al.	2011	7	4	12.5	Corsi	recall	blocks	computer	span	-0.52	0.64
Verkade et al.	2011	83	72	5.6	Corsi	recall	circles	computer	trials	-0.10	0.16
Visu-Petra et al.	2008	111	112	6.3	Corsi	recall	blocks	physical	span	0.00	0.13
Vock & Holling	2008	206	168	11.4	Pattern	recall	geometric	computer	accuracy	0.16	0.10
Weisberg & Newcombe	2014	29	45	19.0	Pattern	recall	blocks	computer	accuracy	0.31	0.24
Wong et al.	2014	27	22	10.1	Pattern	recall	verbal	computer	errors	0.25	0.29
Yerys et al.	2009	13	7	10.3	Token	NA	blocks	computer	errors	0.21	0.47

Year = year of publication; *N*m = number of males; *N*f = number of females; Age = mean age of the sample; alpha= alphanumeric shapes; verbal= verbalizable shapes; NA = not applicable; NR = not reported; RT = reaction time; *d* = biased effect size

Asterisk (*) after the authors names denotes an outlier

reliability of 99.2 % (961 agreements/969 total entries). This high inter-rater reliability suggests that coding was quite straightforward. Accordingly, the remaining material was coded by the second author.

Measure of effect size The standardized mean difference between the performance of males and females (Cohen's *d*; Cohen, 1988) was the measure of effect size. In particular, the effect size was calculated in such a way that positive values reflected a male advantage and negative values reflected a female advantage. When means and standard deviations were available, we calculated the effect sizes with the formula presented by Cohen (1988). This was the case for 155 of the 182 effect sizes (85.2 %). As an inferential statistic (typically *t* test, *p*, *r*, or *F*) was available in the remaining cases, the formulae presented by Lipsey and Wilson (2001) were used as appropriate. Regardless of the underlying statistic, effect sizes were computed using the calculator provided

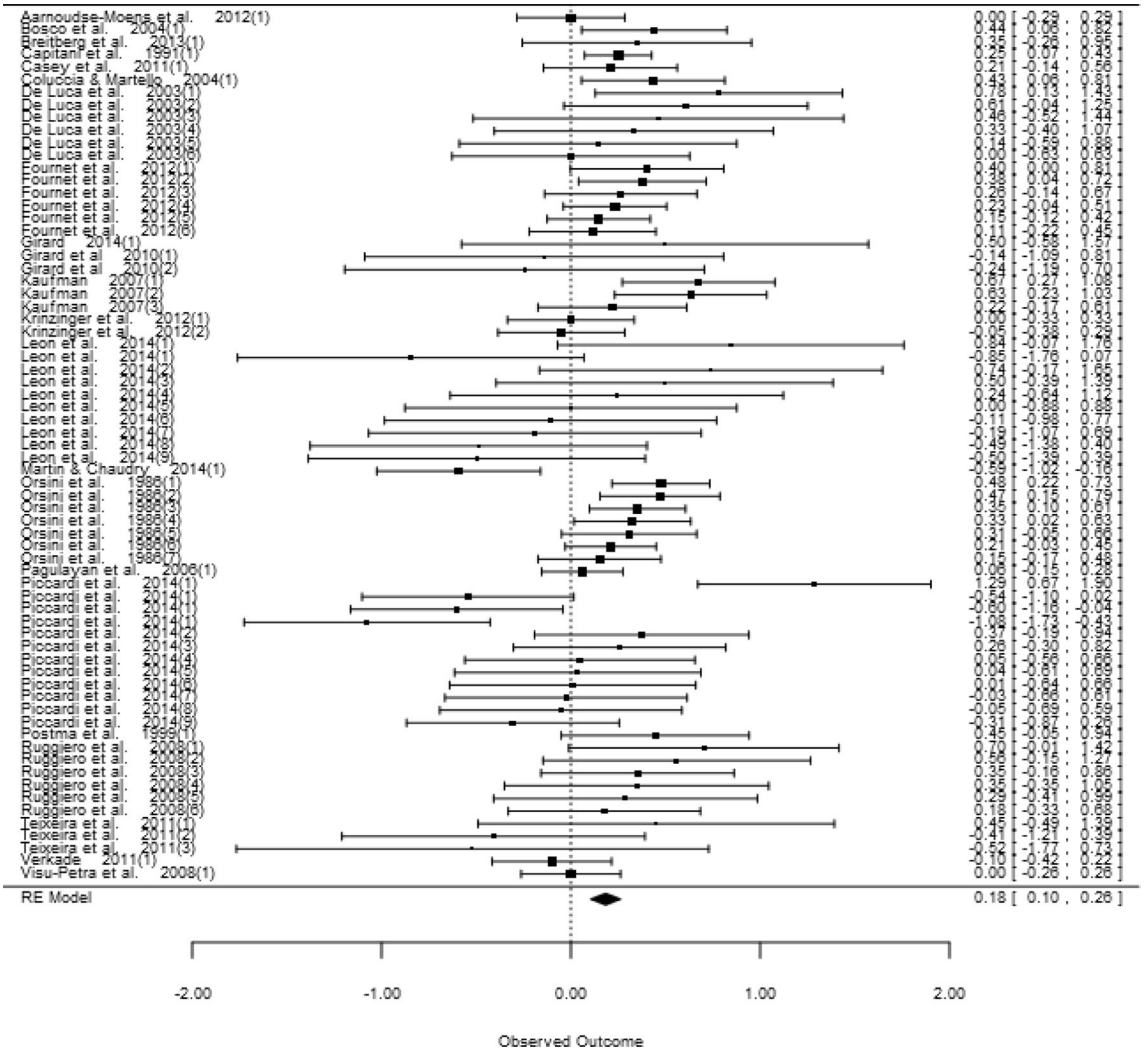

Fig. 1 Forest plot of effect sizes as a function of authors and year of publication for Corsi Block tasks. The square for each study represents the Cohen's d and the size of the square reflects its precision. The error bars reflect the 95 % confidence interval. The random effect model estimate is presented at the bottom of the plot for information purpose, because this estimate does not take the non-independence of effect sizes into account (seen in multiple entries for many studies)

by David Wilson (http://www.campbellcollaboration.org/escalc/html/EffectSizeCalculator-Home.php). A small sample correction was applied to the effect sizes, as recommended by Hedges and Becker (1986). As suggested by Rosenthal (1991), when an effect size was reported as not significant without any other information, it was coded as zero. However, before applying this approach, we contacted by e-mail authors of work published since 2004 with a request for more information but this still left us with 7 of 182 effect sizes in this situation. These were preserved in the sample to provide a representative picture of the available literature. Note, however, that four of the effect sizes presented as zero in Table 1 were actually due to equal performance in females and males (Colom et al., 2013; De Luca et al., 2003; Krinzinger et al., 2012; León et al., 2014).

Data analysis

Multilevel meta-analysis The present meta-analysis was designed to determine whether sex differences in visual-spatial working memory are significant for the overall sample and to identify the variables that moderate them. Examining these two questions would normally be quite straightforward as method to estimate overall effects and performing moderator analyses in the context of meta-analysis are well-established (Lipsey & Wilson, 2001). However, examination of the specific task used in assessing visual-spatial working memory requires consideration of multiple effect sizes obtained from the same sample of participants. These are non-independent effect sizes and this component violates an assumption of fixed and random effects meta-

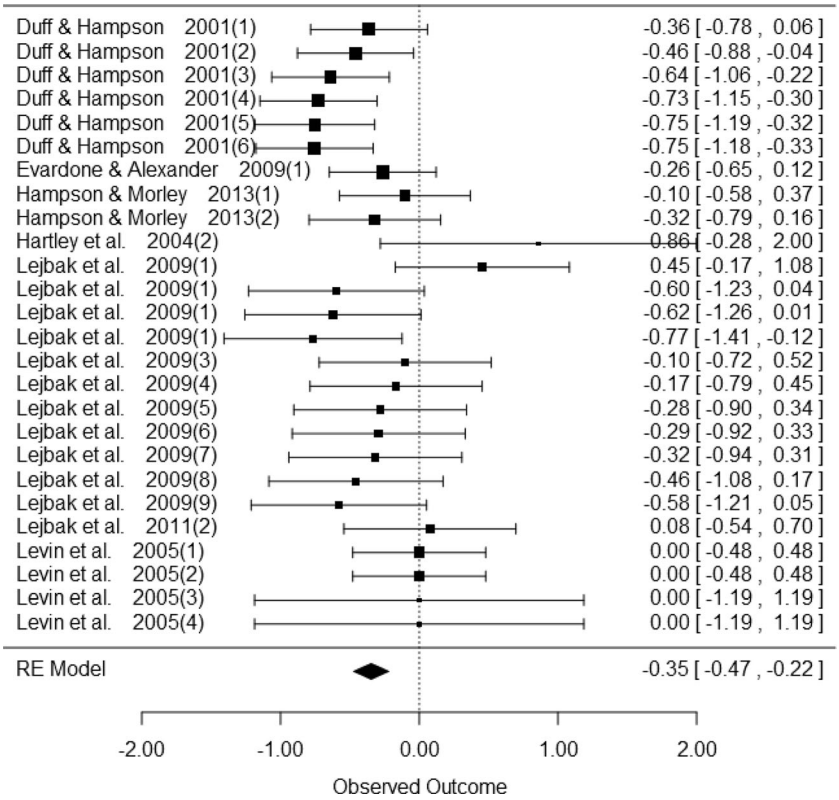

Fig. 2 Forest plot of effect sizes as a function of authors and year of publication for Location tasks. The square for each study represents the Cohen's *d* and the size of the square reflects its precision. The error bars reflect the 95 % confidence interval. The random-effect model estimate is presented at the bottom of the plot for information purpose, because this estimate does not take the non-independence of effect sizes into account (seen in multiple entries for many studies)

analysis (Borenstein et al., 2009). Reliance on one of these approaches to meta-analysis despite the violation of the assumption of independence among effect sizes would distort the statistical analyses, particularly the estimation of the standard errors (Bateman & Jones, 2003). In contrast, multilevel linear modeling (MLM) was designed to handle the type of hierarchical design represented in most meta-analyses without requiring independence of the effect sizes (Raudenbush & Bryk, 2002). In fact, it offers many advantages over fixed and random meta-analysis (Hox & de Leuw, 2003; Hox, 2008). Accordingly, MLM was used as the meta-analytic method here to allow a valid examination of the overall and moderator analyses.

The variables task, year of publication, publication status, mean age of sample, age coded categorically, testing medium, stimulus type, dependent variable, and type of memory were considered in the moderator analysis. Finally, as the standard error calculated for each effect size in a meta-analysis reflects an estimate of the variance for individual effect sizes (Borenstein et al., 2009), it was possible to compute "variance-known" (or V-known) hierarchical linear models. Effect sizes were treated as random effects whereas moderators were treated as fixed effects in what amounted to mixed models. It should be noted that the V-known modeling results in the precision weighted estimates of effect sizes typical of other approaches to meta-analysis (Raudenbush & Bryk, 2002).

Categorical independent variables were dummy coded into k - 1 dichotomous vectors (where k represents the number of categories) for consideration in the multilevel analysis. Following this form of coding, the intercept represents the mean estimated effect size for the category coded as zero in all vectors (reference category) and its test of significance indicates whether it is significantly different from zero. The coefficient for each vector represents the difference between the mean for the category coded as "1" in that vector and the intercept. However, estimated effect sizes (i.e., the sum of the intercept and the coefficient for each coded vector) are reported to simplify the presentation of results. All significance

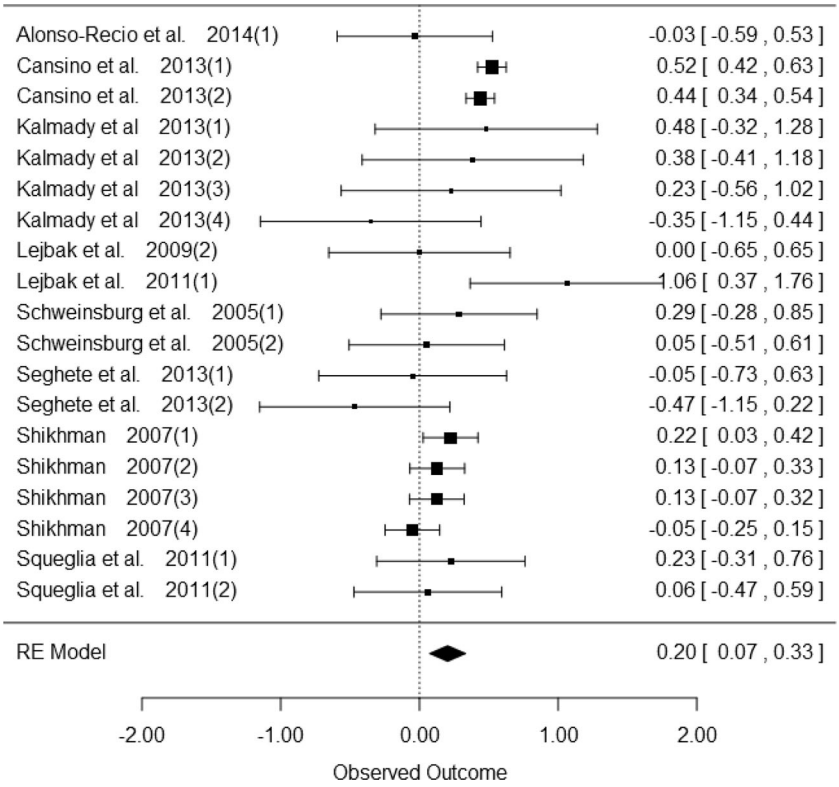

Fig. 3 Forest plot of effect sizes as a function of authors and year of publication for N-back tasks. The square for each study represents the Cohen's *d* and the size of the square reflects its precision. The error bars reflect the 95 % confidence interval. The random-effect model estimate is presented at the bottom of the plot for information purpose, because this estimate does not take the non-independence of effect sizes into account (seen in multiple entries for many studies)

testing was based on the robust estimates of standard errors that are routinely computed by HLM7.

The multilevel analysis was computed by examining data organized in two levels: effect sizes nested within samples. This structure resulted in 182 effect sizes (Level 1) nested within 98 samples (Level 2). All analyses were conducted with the HLM 7 software (Raudenbush, Bryk, Cheong, Congdon, & du Toit, 2011) with the significance level set at 0.05. Only moderators that produced significant results are elaborated upon.

Task subgroups analyses The overall multilevel analysis combined statistical power and the ability to examine tasks differences as moderators in the analysis, so it was a necessary step in a thorough examination of the data retrieved for the present analysis. However, as we mentioned previously, in many cases moderator categories were confounded with specific task and this precluded a test of some potentially meaningful interactions. The multilevel meta-analysis examined only main effects as a matter of necessity. However, to circumvent problems associated with confounded moderators, after establishing that the tasks indeed differed in the magnitude of sex differences, we proceeded with a moderator analysis within each specific task to provide more fine-grained conclusions. When non-independent effect sizes were observed for a specific task, the analysis relied on multilevel meta-analysis. However, as a small number of level-2 units precludes calculations of robust standard errors (Raudenbush & Bryk, 2002), this approach could only be applied to task with sufficient numbers of level 2 units. For the remainder, mixed-method meta-analysis (treating effect sizes as random variables and moderators as fixed effects) was applied as recommended by Borenstein et al. (2009). Non-independence of effect sizes still existed in these samples (as noted when relevant), but they were kept as is to preserve a complete data set. Because this non-independence is most likely to affect computation of standard errors, confidence intervals will have to be interpreted with caution. Analyses performed in this second step were conducted either with HLM-7 or with the SPSS macros developed by Wilson (2005). All of the moderators examined for

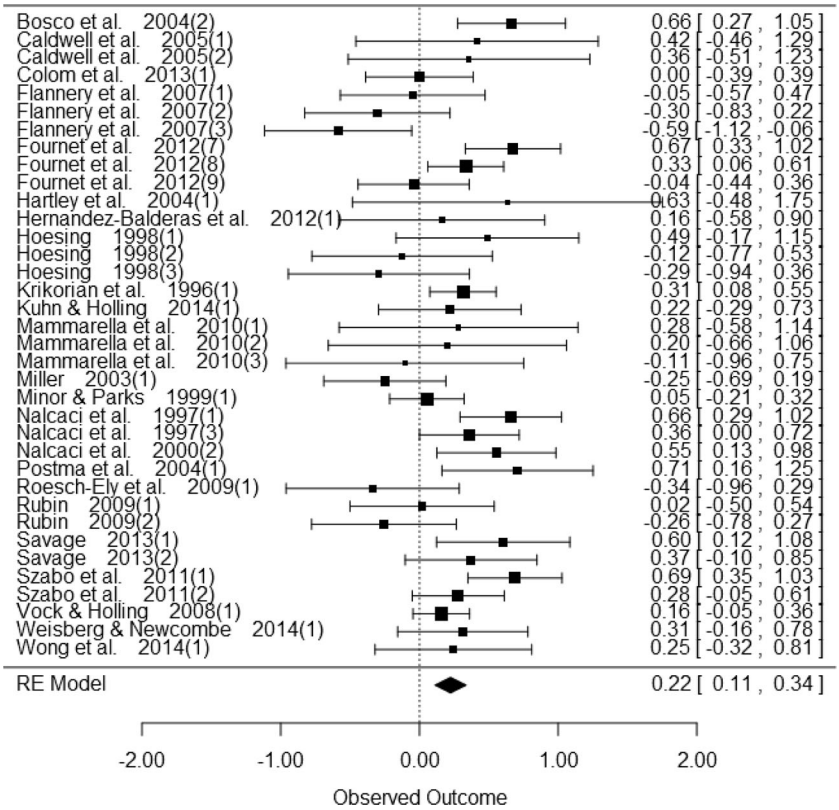

Fig. 4 Forest plot of effect sizes as a function of authors and year of publication for Pattern tasks. The square for each study represents the Cohen's *d* and the size of the square reflects its precision. The error bars reflect the 95 % confidence interval. The random-effect model estimate is presented at the bottom of the plot for information purpose, because this estimate does not take the non-independence of effect sizes into account (seen in multiple entries for many studies)

the overall analysis (except specific task, of course) were considered in this subgroups analysis as well, with the significance level set at 0.05.

Results

A preliminary data analysis was conducted to identify outliers, defined as effect sizes that were more than 3.29 standard deviations[2] from the grand mean (as recommended by Tabachnick and Fidell, 2007). Two such outliers were identified. Data analyses conducted with and without these effect sizes affected the results in terms of significance. Therefore, these outliers were excluded in all data analyses, although they are identified by a star (*) in Table 1. The final sample consisted of 180 effect sizes drawn from 98 independent samples, reflecting combined results from 5,035 males and 5,693 females.

Multilevel meta-analysis

Overall sex differences in visual-spatial working memory Examination of the combined effect size was performed by computing a null model where the test of significance for the intercept is examined (Raudenbush & Bryk, 2002). Results of this analysis revealed a mean estimated *d* of 0.155 (95 % confidence interval (CI) = 0.087-0.223), indicating that males significantly outperformed females on visual-spatial working memory tasks, $t(97) = 4.54$, $p < 0.001$.

The variance component in the full sample was examined to determine whether significant variation in effect sizes exist between samples. Results of this analysis showed that it was the case, $\chi^2(97) = 391.34$, $p < 0.001$. This indicates that significant heterogeneity exists in the effect sizes. Strictly speaking, it appears that the overall estimate of effect size did not provide a representative summary of the sample of effect sizes. The examination

[2] The specific criterion of 3.29 proposed by Tabachnick and Fidell (2007) is based on the notion that scores that deviate from the mean by that amount are linked with a probability of 0.001 in the normal distribution. This criterion is used throughout their book to define outliers. It was adopted, because their approach seemed as reasonable as any other arbitrary criteria that have been adopted by others.

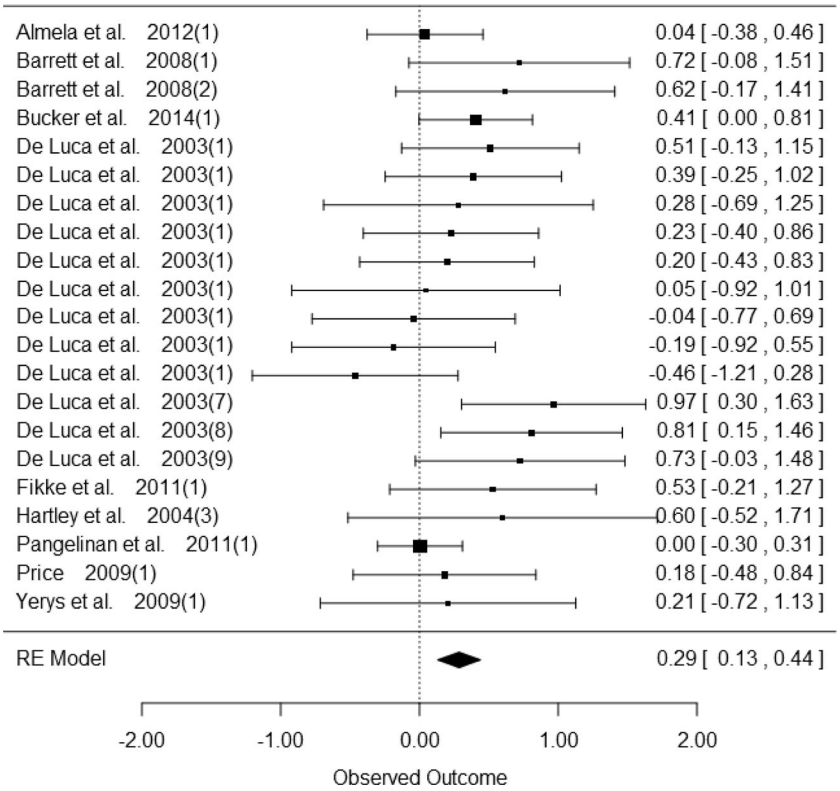

Fig. 5 Forest plot of effect sizes as a function of authors and year of publication for Token tasks. The square for each study represents the Cohen's *d* and the size of the square reflects its precision. The error bars reflect the 95 % confidence interval. The random-effect model estimate is presented at the bottom of the plot for information purpose, because this estimate does not take the non-independence of effect sizes into account seen in multiple entries for many studies)

of potential moderators might therefore shed light on factors accounting for this variability.

Moderators of sex differences in the overall sample The moderator analysis proceeded through a comparison of the deviance observed with a given moderator in the model compared with the deviance observed for the null model using a full maximum likelihood approach, as recommended by Raudenbush and Bryk (2002). As such, moderators were examined one at a time in models. This essentially allowed an assessment of the moderators accounting for significant variance in effect sizes. Proceeding in this manner, the moderator analysis revealed that task accounted for significant variance in effect sizes, $\chi^2(5) = 27.17$, $p < 0.001$. Estimated effect sizes for this variable are presented in Table 2. Considering that none of the 95 % confidence intervals contain zero, this indicates that a significant male advantage was observed for all task categories except memory for location, where a female advantage was in evidence. In addition, the estimated effect size for memory for location was significantly larger than for the reference category (Corsi blocks), $t(77) = -5.29$, $p < 0.001$, whereas the estimated effect size for the reference category did not differ significantly from that obtained in all the other tasks (all $p > 0.277$).

Testing medium also was found to contribute significantly to variance in effect sizes, $\chi^2(2) = 6.82$, $p = 0.032$, with estimated effect sizes also presented in Table 2. In this case, only computer testing produced effect sizes that are significantly different from zero based on the 95 % confidence intervals. Unreported medium and physical medium produced nonsignificant sex differences, although only the latter was significantly smaller than for the reference category of computer testing ($p = 0.026$ for physical medium, $p = 0.102$ for unreported medium).

Finally, even though age defined categorically failed to account for significant variance, $\chi^2(4) = 7.24$, $p = 0.123$, mean age of the participants considered as a grand mean centered continuous variable produced a significant contribution to variance, $\chi^2(1) = 5.81$, $p = 0.015$. The positive level 2 coefficient ($\gamma = 0.004$, $SE = 0.001$) indicates a significant increase in the magnitude of sex differences in visual-spatial working memory with age, $t(96) = 3.73$, $p < 0.001$. Despite the lack of significance for the categorical analysis for age, relevant summary values are presented in Table 2 as a means to facilitate discussion of this finding. The remaining moderators

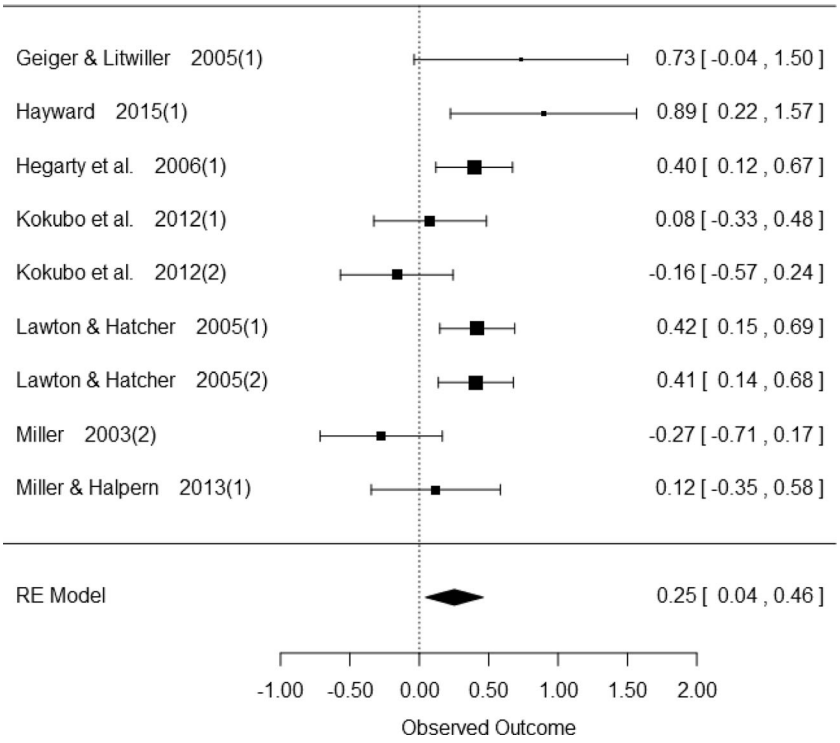

Fig. 6 Forest plot of effect sizes as a function of authors and year of publication for Other tasks. The square for each study represents the Cohen's *d* and the size of the square reflects its precision. The error bars reflect the 95 % confidence interval. The random-effect model estimate is presented at the bottom of the plot for information purpose, because this estimate does not take the non-independence of effect sizes into account (seen in multiple entries for many studies)

examined in this analysis (year of publication, stimuli, dependent variable, and type of memory) failed to achieve significance (all *p*s > 0.071).

Task subgroup analysis

In the task subgroup analyses, it was only possible to use multilevel modeling with Corsi Blocks tasks and memory for patterns, because they had enough level 2 units (samples) to allow computation of reliable robust standard errors (Raudenbush & Bryk, 2002). For the other tasks, an examination of the fit of the fixed model was first performed as reflected in the homogeneity statistic. Such a fit would essentially indicate that no variance remained to be explained by moderators and only sampling error accounts for variability in effect sizes (Borenstein et al., 2009). A moderator analysis is not warranted unless the fixed effects model is rejected in favor of the random-effect model. When this occurred, the moderator analysis was conducted by means of a mixed-effects model. However, because all tasks subgroups included non-independent effect sizes, collapsing across these effect sizes to preserve only independent effect sizes would remove the moderator categories and would defeat the purpose of such an analysis. Accordingly, all effect sizes were preserved as is. In view of the effect of non-independence on standard errors, confidence intervals will have to be interpreted with caution in such analyses.

Corsi blocks With 48 samples (69 effect sizes), analysis of the data obtained with Corsi Blocks tasks and equivalent proceeded with multilevel modeling. A significant variance component was obtained, $\chi^2(47) = 100.10$, $p < 0.001$, suggesting that the examination of moderators might shed light on factors accounting for variability among effect sizes.

As part of this analysis, an additional moderator specific to the Corsi Blocks task was considered because this task often is implemented with either forward recall of order (i.e., in the same order in which the blocks were tapped; $k = 51$ effect sizes) or with backward recall ($k = 11$). Therefore, these formed two possible categories for this moderator, labelled "order." Two other categories had to be coded as some authors combined both orders ($k = 4$) or simply did not report the order that they used ($k = 3$).

Results of the moderator analysis with Corsi Blocks and equivalent tasks are summarized in Table 3. They showed that age defined categorically accounted for a

Table 2 Summary for significant moderators in the multilevel meta-analysis

Moderator	Sample size (*k*)	Estimated mean *d*	95 % confidence interval
Task			
Corsi blocks	69	0.170	0.088, 0.252
n-back	19	0.200	0.020, 0.372
Memory for patterns	36	0.242	0.108, 0.376
Memory for location	26	−0.339	−0.528, −0.150
Memory for token	21	0.258	0.100, 0.416
Other	9	0.272	0.056, 0.488
Testing medium			
Computer	115	0.224	0.158, 0.290
Not reported	10	0.017	−0.228, 0.262
Physical	55	0.058	−0.085, 0.201
Age categories (yr)			
<13	29	0.033	−0.069, 0.135
13-17	10	0.229	0.043, 0.42
18-29	36	0.134	−0.033, 0.302
30-49	10	0.289	0.056, 0.522
≥50	13	0.264	0.141, 0.387

Sample size (*k*) for task and testing medium is based on 180 effect sizes, whereas it is based on 98 samples for age categories

significant proportion of variance, $\chi^2(4) = 14.25$, $p = 0.007$, with summary values for this finding presented in Table 3. Confidence intervals indicate that a significant male advantage was found at all ages, except younger than age 13 years and between ages 13 and 17 years, where the confidence interval includes zero. In addition, the reference category (<13) produced significantly smaller sex differences than the 18-29, 30-49, and ≥50 categories (all $p < 0.03$), and it produced a marginal difference with the 13-17 category ($p = 0.053$). The categorical effect of age was echoed in a significant effect of grand mean centered age as a continuous variable, $\chi^2(1) = 5.14$, $p = 0.022$, reflecting a significant increase in the magnitude of sex differences with age ($\gamma = 0.004$, $SE = 0.001$).

Finally, year of publication accounted for significant variance in effect sizes obtained with Corsi Blocks tasks, $\chi^2(1) = 7.85$, $p = 0.005$. This finding was due to a significant decrease in the magnitude of sex differences as year of publication increases in value (γ = -0.011, SE = 0.003). The apparent role of year of publication might reflect an underlying correlation between mean age of the sample and year of publication, $r(69) = -0.49$, $p < 0.001$. However, year of publication remained a significant moderator of effect sizes when age was entered before that variable in the analysis, $\chi^2(1) = 3.82$, $p = 0.048$. This suggests that the year effect accounts for variance in effect sizes over and above its relation with age of participants.

N-back With eight samples (19 effect sizes), analysis of the data obtained with n-back tasks proceeded with a mixed model moderator analysis after fit of the fixed effect model was rejected. Specifically, the effect sizes showed significant overall heterogeneity, $Q(18) = 59.68$, $p < 0.001$, suggesting that sources of variation other than random sampling (i.e., moderators) might account for variability in effect sizes in this sample.

Accordingly, the moderator analysis for n-back tasks showed that age as a categorical factor accounted for a significant proportion of variance, $Q(2) = 15.04$,

Table 3 Results of the moderator analysis for Corsi blocks and n-back tasks

Moderator	Corsi blocks		n-back	
	k	*d* (95 % CI)	*k*	*d* (95 % CI)
Age of sample (yr)				
<13	21	−0.006 (−0.119,0.107)	0	-
13-17	3	0.325 (−0.001,0.652)	6	0.044 (−0.204,0.291)
18-29	9	0.283 (0.035,0.532)	10	0.137 (0.018,0.256)
30-49	5	0.370 (0.179,0.560)	3	0.453 (0.320,0.588)
≥49	10	0.261 (0.124,0.398)		

Table presents the number of effect sizes (*k*) and the mean weighted *d* for each moderator category with the 95 % confidence interval (CI) in parentheses. The mean weighted effect size is significantly different from zero with $p < 0.05$ if the 95 % CI for *d* does not include zero. For the n-back task, the last age category with data actually includes 30 and above

$p < 0.001$, even though there were no samples younger than age 13 years, whereas the >50-years category contained only one effect size that was combined into a 30 and above grouping. Summary values for this finding are presented in Table 3. In this case, confidence intervals indicate that a significant male advantage was found for the 18-29 and ≥30 categories, but not between 13 and 17 years, where the confidence interval includes zero. In addition, multiple comparisons with the Z-score method as recommended by Borenstein et al. (2009) showed that the magnitude of the effect was significantly larger for those in the 30-49 age group compared with the 13-17 and 18-29 groups (largest $p = 0.004$). The difference between the last two groups did not achieve significance ($p = 0.5$). Also, the categorical age effect was mirrored in a significant effect of grand mean centered age, $Q(1) = 6.22$, $p = 0.013$, reflecting a significant increase in the magnitude of sex differences with age ($b = 0.094$, $SE = 0.004$).

Finally, year of publication also accounted for significant variance in effect sizes obtained with n-back tasks, $Q(1) = 3.86$, $p = 0.049$. This finding was due to a significant increase in the magnitude of sex differences as year of publication increased in value ($b = 0.034$, $SE = 0.017$). However, in this case, the apparent role of year of publication might reflect more the underlying correlation between mean age of the sample and year of publication, $r(19) = 0.32$, $p = 0.19$. Even though this correlation is not significant, likely due to the small sample size, when age is entered before year of publication in the meta-regression, the year effect becomes nonsignificant ($p = 0.34$).

Memory for location and token Overall analysis in the memory for location (26 effect sizes from 9 samples) and the token (21 effect sizes from 14 samples) subgroups supported fit with the fixed effects model. Specifically, in both these grouping, non-significant homogeneity of effect sizes was observed: $Q(25) = 34.96$, $p = 0.089$ for location; $Q(20) = 22.48$, $p = 0.315$ for token. This suggests that the fixed effects model is appropriate for these data and that sampling error accounts for variability in the effect sizes they comprise. Therefore, moderator analysis is not required or appropriate.

Memory for patterns and other tasks With 25 samples (36 effect sizes), the data obtained with tasks measuring memory for patterns were analyzed with multilevel modeling. However, this analysis failed to reveal any moderator accounting for significant variance in effect sizes (all $p > 0.24$), despite a significant variance component, $\chi^2(24) = 59.02$, $p < 0.001$.

Finally, other tasks included only nine effect sizes from seven samples and required a mixed model meta-analysis as fit for the fixed effects model was rejected: $Q(8) = 19.15$, $p = 0.014$. However, here as well, moderator analysis failed to reveal any significant findings.

Publication bias and the file drawer problem

Considering that the present meta-analysis consists mostly of data obtained from published studies and that our literature search strategy might have biased the retrieved material in favor of research that showed a significant sex difference, it is possible that the final sample might not be representative of the entire population of studies in existence (Rosenthal, 1979). Such a situation would potentially be the *result* of an inherent publication bias, reflecting the fact that studies producing nonsignificant results have a lower probability of publication. This problem, called the "file-drawer problem" (Sterling, 1959) has potential to bias meta-analytic results. Essentially, by considering mostly published studies, a meta-analysis might exaggerate the magnitude of the effect under consideration.

As a simple way to examine the potential influence of the file-drawer problem on our results, we compared the mean estimated effect sizes for samples obtained from published ($k = 88$ samples) and unpublished research ($k = 10$ samples). This analysis showed no significant influence of publication status: χ^2 (1) = 0.001, $p > 0.5$. This suggests no evidence of a publication bias in the present sample.

As an additional source of information, the most recently developed approach to an examination of the publication bias also was considered. Specifically, Ioannidis and Trikalinos (2007) proposed a test based on the rationale that a publication bias in a set of effect sizes should produce an excess of observed positive findings when compared to what is expected from the power of individual studies. As a conservative measure, Ioannidis and Trikalinos recommended use of 0.10 as the significance level for such test to reduce the risk of Type II errors.

Accordingly, we examined publication bias with the method proposed by Ioannidis and Trikalinos (2007). As a starting point, we defined a positive finding as a result showing a male advantage, because it would support the potentially pervasive expectations from reviewers and editors leading to a publication bias. Having established this component, we then determined whether the distribution of effect sizes was asymmetrically biased toward those reflecting a male advantage compared with what would be expected by chance under null hypothesis statistical testing. This approach

relies exclusively on the logic of hypothesis testing and makes no further assumption as is required, for example, in the commonly used Egger's test (Egger, Davey Smith, Schneider, & Minder, 1997). In fact, unlike the Egger's test, the Ioannidis and Trikalinos method is not tied to a specific meta-analytic model (e.g., fixed or random effect model). Use of this approach in the overall data set suggests that the number of positive findings is as expected from the power of retrieved studies, reflecting no significant publication bias (Observed = 47, Expected = 43.42, $\chi^2 = 0.39$, $p = 0.53$).

Despite its perceived advantages, the Ioannidis and Trikalinos (2007) method has been the object of many criticisms (see the special issue of *Journal of Mathematical Psychology*, Volume 57, Issue 5, but also Francis' rebuttal of these criticisms in the same issue). Therefore, as a means to obtain converging evidence, we also examined publication bias with two of the most common methods used for that purpose, as suggested by Stern and Egger (2005). Specifically, despite potential drawbacks, the Begg and Mazumdar (1994) and the Egger et al. (1997) approaches were used as additional tests for potential publication bias in the present sample. Both methods are based on the notion that studies with a small sample and a small effect size are less likely to be published (Borenstein et al., 2009). Therefore, if a publication bias is present, a plot of precision (the inverse of the standard error; y axis) against effect size (x axis) would produce an asymmetrical distribution with few values on the bottom left hand side of the plot, where small samples and negative effects would belong. Accordingly, the present data are shown in such a funnel plot in Fig. 7. A visual inspection of Fig. 7 reveals no sign of asymmetry. However, assessing the presence of a bias by visual examination of the plot is quite subjective and the methods proposed by Beggs and Mazumdar and Egger et al. objectify this evaluation.

In particular, the logic of the Begg and Mazumdar (1994) test is fairly straightforward: If it is the case that studies with a small sample and a small effect size are less likely to get published, then a sample exhibiting a publication bias should be replete with studies that have small samples but large effect sizes. Considering that the sampling variance of effect sizes is computed from sample size and large samples produce smaller variance, then this would predict a significant negative correlation between sampling variance and effect size if a bias is present. Of course, this could be examined simply through a Pearson correlation (Pearson $r = 0.028$, $p = 0.706$, N = 180 in the present sample). However, the Beggs and Mazumdar method improves the precision of the correlation estimates by first standardizing

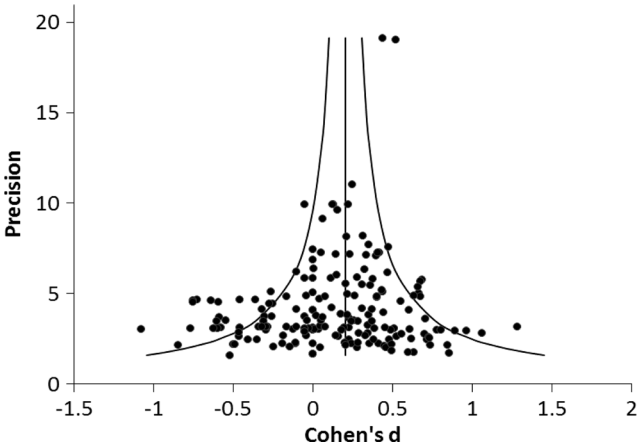

Fig. 7 Funnel plot representing precision (1/SE) as a function of Cohen's *d* for the whole sample. A biased sample is likely to be asymmetrical in such a function

the effect sizes as a function of the weight given to each effect size in the data. A Kendall's rank correlation is then computed between the resultant measure of effect size and sampling variance. A significant correlation is interpreted as reflecting publication bias. However, in the present sample, this approach revealed a correlation of 0.0007 (corrected for ties) that is not significant ($p = 0.989$). This method likely could be seen as appropriate even in a data set with non-independent effect sizes considering that it makes no assumptions about the underlying meta-analytic model.

Concerning the Egger et al. (1997) method, the standard normal deviate for the effect size is regressed on precision. In an unbiased sample, the regression line should run through the origin, so that the intercept of the regression equation should not be significantly different from zero in a statistically symmetrical funnel plot. Egger et al. recommended a significance level of 0.10 to maximize power. Results of the Egger et al. test showed that the intercept was significantly larger than zero at $p < 0.10$, with an intercept estimate of −0.87 (90 % CI: −1.25 to −0.50). However, the negative intercept is somewhat puzzling, because it would suggest that it is the result of what amounts to a positive correlation between precision and effect sizes, contrary to what is the underlying assumption of the Egger's test. This puzzling issue can be better understood when examining the funnel plot more closely (Fig. 7). Specifically, two points with a larger precision than the rest of the sample are particularly obvious and clearly distort the results. Reexamination of the Egger's test with these data points removed showed a nonsignificant intercept (−0.33, 90 % CI: −0.76, 0.10). The Egger's test therefore is consistent with the other publication bias tests conducted so far when this statistical issue is considered. In fact, the finding that two particularly

large sample sizes have such a profound effect on the results emphasizes issues with Egger's test. In contrast, the Begg and Mazumdar (1994) and the Ioannidis and Trikalinos (2007) are not affected as profoundly by such outliers.

As a summary, the four criteria adopted here show no support for a publication bias despite the inclusion of *sex* or *gender* as a search term. Of course, it is not possible to completely exclude a bias on the basis of these results. However, the data show strong converging evidence against the presence of such a bias.

Discussion

The purpose of the present study was to conduct a meta-analysis of sex differences in visual-spatial working memory to quantify the overall findings as well as to examine potential moderators of these sex differences. The importance of examining sex differences in visual-spatial working memory arose from the broad usage of such tasks in clinical and research settings. In addition, quantifying these sex differences has implications for the hypothesis that they might account for sex differences in spatial abilities. Essentially, even though many have made the claim that sex differences in spatial abilities, especially mental rotation, could stem from sex differences in visual-spatial working memory (Kaufman, 2007; Loring-Meier & Halpern, 1999; Wang & Carr, 2014), a comprehensive meta-analysis examining visual-spatial working memory in this context was missing until now.

The present analysis of 180 effects sizes from 98 samples therefore provided an examination of the sex differences in visual-spatial working memory by means of multilevel analysis in the overall sample and multilevel or mixed effect models meta-analyses in the separate tasks that were identified. Results showed a small but significant male advantage in the overall sample. Analysis in the overall sample also showed an effect of task, with all tasks showing a significant male advantage, except memory for location, where a female advantage emerged. Interestingly, the male advantage was only found for computer tasks, not when physical media (cards, blocks, etc.) were used. Finally, age was a significant moderator, showing an increase in the magnitude of sex differences with age. In fact, age was also a significant moderator in two task subgroups (Corsi Blocks, n-back). An effect of year of publication also emerged for these two subgroups. These results show minimal influence from moderators. In fact, memory for location and for token required no moderator analysis as the results showed better fit for a fixed effects model in the subgroups. In contrast, memory for patterns and other tasks showed no significant moderator despite a better fit for the random effects model. Keeping these results in mind, we will discuss their implications, starting with a consideration of the overall effect and then examining the results of the moderator analysis.

Overall results

Results of the overall analysis clearly show a male advantage in visual-spatial working memory with an estimated effect size of 0.155 that is significantly different from zero. However, this also reflects a small effect that essentially fits with the notion that most sex differences are small and that similarities tend to be the norm (Hyde, 2005; 2014). To put this finding in context, an effect size of 0.155 would require a total sample of 654 participants to achieve 80 % power at the 0.05 level of significance. This is much larger than the typical individual study sample size as presented in Table 1. This suggests that open archiving of data might provide an interesting avenue for researchers in that it would provide a cumulative record of available data that would likely improve our understanding of the underlying factors.

Despite their small size, the direction of sex differences is fairly consistent in the present analysis, with only one category reflecting a significant female advantage (memory for location in Table 2). This consistency emphasizes the importance of not disregarding completely any small effect as it might have implications in specific contexts. For example, the mere existence of sex differences in visual-spatial working memory should warn clinicians to be cautious when comparing male and female patients. In fact, if the test publisher feels that there is a need to provide separate norms for the Wechsler Adult Intelligence Scale—Third Edition (WAIS-III; Wechsler, 1997) despite controversial data concerning the existence of sex differences in intelligence (Halpern, 2013), a demonstrated case as we found here should clearly warrant separate norms for males and females. Similarly, it might be advisable to account for the sex difference in building models of visual-spatial working memory. Essentially, much empirical research is required before we can state without a doubt that males and females process visual-spatial working memory tasks in the same way. In the meantime, it might be advisable to examine model fit separately for male and female samples.

In broad terms, likely the most obvious implication of the present results is that the large amount of research that simply ignores sex as a relevant variable in visual-spatial working memory research overlooks a possible source of variation. At the very least, researchers should state the number of females and males they tested and routinely examine sex as a

factor in preliminary analyses. Full report of sample characteristics is always required even when no sex differences have been documented in the past (APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008). The present results suggest that full report of sample composition for studies of visual-spatial working memory is even more imperative than mere adherence to reporting standards.

Visual-spatial working memory and spatial abilities The assumed link between sex differences in spatial abilities and visual-spatial working memory was one of the driving forces behind the present study. From this perspective, the magnitude of the overall effect for sex differences in visual-spatial working memory suggests that such small effects could hardly account for spatial performance sex differences that can be as large as a Cohen's *d* of 0.94 when considering mental rotation (Linn & Petersen, 1985). This suggests that even after taking visual-spatial working memory into account, there is still much variance left to explain in the relation between sex and spatial abilities. In fact, task selection might provide an explanation of Kaufman's (2007) finding that visual-spatial working memory was a complete mediator of sex differences for the DAT-SR but not for the Mental Rotations Test (MRT). Specifically, Voyer et al. (1995) reported a mean *d* of 0.27 for the former and of 0.67 for the latter. If we assume that both tasks have a strong visual-spatial working memory, this would leave much less variance left to explain for the DAT-SR than for the MRT, thereby potentially allowing room for a unique path between sex and scores on the latter. Taking all the evidence into account, the present analysis suggests that sex differences in visual-spatial working memory might partly account for sex differences in spatial abilities, but there also is much variance left to be explained by other factors in future empirical work. The results of the moderator analysis give preliminary notions of variables that might account for some of the remaining variance. Therefore, we will now turn to the implication of these findings.

Moderator analysis

The overall and subgroup analyses overlapped much in terms of the significant moderators. Therefore, this part of the discussion will proceed as a function of specific moderators, starting with task as it was only possible to examine it in the overall analysis, then proceeding to those that emerged most frequently and moving on to those that were unique to a specific subgroup of tasks.

Task as a moderator The findings relevant to task as a moderator (Table 2) suggest that the magnitude of sex differences in visual-spatial working memory is fairly consistent in showing a male advantage in all the tasks we sampled, except for object location memory, where the advantage is in favor of females. In fact, it seems that the effect of task as a moderator is driven completely by the reversal of the effect in memory for location compared to other tasks. Specifically, when memory for location tasks is removed, task does not account for a significant amount of variance in effect sizes anymore (p = 0.69) and the overall effect size goes from 0.155 to 0.208 (95 % CI: 0.149, 0.267). It is thus legitimate to state that sex differences in visual-spatial working memory tapped by Corsi Blocks, n-back, memory for patterns, memory for tokens, and other tasks that fit our theory neutral definition but excluding memory for location produce a small but consistent male advantage. Even though the memory for location tasks included here fit our definition, they reflect the only tasks with a pure location component.

We have argued earlier that the memory for location tasks in our sample are distinct from the ones sampled by Voyer et al. (2007) as their meta-analysis focused on tasks that included a time interval between encoding and retrieval that is too long and the number of objects too numerous to fit our definition. Essentially, the location tasks included here had a short encoding to retrieval interval and typically involved only one object. Therefore, it would appear that the female advantage reported by Voyer et al. (2007) generalizes to the tasks sampled here. In fact, their mean estimated *d* of −0.269 (sign changed to fit our coding) is within the confidence intervals for our mean effect size of −0.339 (Table 2). This strongly supports the existence of a generalizable female advantage in pure object location memory tasks.

As we also noted in the introduction, practically all the visual-spatial working tasks seem to have a location component. However, they are not pure object location memory tasks. Specifically, in addition to location, the Corsi Blocks task has a sequencing component; memory for a pattern requires associating a location with a specific pattern, thereby adding a non-location component; memory for token requires remembering location of the token across several trials so that participants have to remember both the target and distractors; the n-back task requires remembering the location of a target across specific trials, thereby essentially adding an episodic component. Examination of tasks components suggests that, although females are better than males to remember a location, the sex difference reverses in direction as soon as another component is added to the task.

Why is that the case? Obviously, a meta-analysis does not allow an answer to that question as all we can do is compare across studies. Interestingly, only one of the studies we retrieved (Hartley, Elsabagh, & File, 2004) examined memory for location and another visual-spatial working memory task (memory for pattern) in a within-subject design. What is even more interesting is that a male advantage emerged in both tasks in their study. This suggests that more research including visual-spatial working memory tasks that are or are not pure

location measures are required to examine more closely the component processes. A study by Loring-Meier and Halpern (1999) actually examined what they viewed as components tasks reflecting the process of visual-spatial working memory. However, they did not include location as a component. In fact, we could not include their experiment in our sample because, taken separately, the components they studied did not fit our definition of visual-spatial working memory. To our knowledge, no other researchers have taken a sex differences perspective on components of visual-spatial working memory. The present results relevant to the effect of task suggest that it would be a worthwhile avenue for future research.

Age as a moderator Mean age of the sample was found to account for significant variance in the overall analysis as well as in Corsi Blocks tasks and n-back tasks. Age as a continuous variable was positively related to the magnitude of sex differences in these three analyses, but age as a categorical variable was only significant for Corsi Blocks and n-back tasks. Regardless of this aspect, the relevant results presented in Tables 2 and 3 are clear: Sex differences in visual-spatial working memory seem to emerge in young adulthood (18-29 years group), whereas they are not significant in childhood (younger than 13 years) and adolescence (13-17 years). Of course, the estimated effect size for the 18-29 group in the overall analysis (Table 2) seems at variance with this conclusion, because its confidence interval includes zero. However, it is important to keep in mind that all the samples that investigated memory for location belonged in the 18-29 age group. Because this is the only task that showed a clear female advantage, it dropped the magnitude of the effect for that grouping. In fact, when memory for location tasks are removed from the overall analysis, age as a categorical variable becomes significant ($p = 0.008$) and the effect size for the 18-29 age group (based on 31 samples) becomes 0.279 (95 % CI: 0.133, 0.426). It is clear that the sex difference emerges in young adulthood in visual-spatial working memory tasks that show a male advantage.

Aside from pointing out once more the distinct nature of location memory, this reanalysis suggests that when visual-spatial working memory tasks show a male advantage, this advantage appears first in the category that includes the age associated with puberty (13 years)—a finding similar to what Voyer et al. (1995) reported with spatial abilities. In accounting for this finding, hormonal changes, the most obvious factor associated with puberty, provides an interesting option. After all, there is ample evidence of the influence of sex hormones on visual-spatial working memory in humans, although the data remain unclear concerning the respective role of estrogen and progesterone (Duff & Hampson, 2000; Postma, Winkel, Tuiten, & van Honk, 1999) or testosterone (Cherrier et al., 2001) and whether these hormones improve or impair visual-spatial working memory. This is clearly an area that requires much more research.

Another possibility to account for developmental differences in visual-spatial working memory comes from research suggesting that children and adults use different cerebral structures when performing such tasks, especially as memory load increases, which might account in part for poorer performance in children than in adults (Thomason et al., 2009), at least for children aged 7-12 years and adults aged 20-29 years, as was the case in the Thomason et al. study. Inasmuch as changes in activation patterns reflect strategy choice, this would suggest a role for changes in strategy selection with age in the emergence of visual-spatial working memory, although developmental changes in cerebral activation could also reflect brain maturation. However, here as well, data are contradictory with Thomas et al. (1999) suggesting that similar cortical regions are activated in children (8-10 years old) and adults (19-26 years old) during performance of a visual-spatial working memory task. Again, this lack of agreement suggests that more research is required, especially considering that developmental aspect and sex differences in visual-spatial working memory have not been tackled in the same study in the context of research relying on neuroimaging or on the measurement of hormones. Clearly, the present meta-analysis does not allow causal claim to account for age effects on the magnitude of sex differences in visual-spatial working memory. Nevertheless, it suggests that it might be fruitful to investigate hormonal and strategy selection as factors in future work.

Testing medium as a moderator Testing medium was only a significant moderator in the overall analysis. The pattern of results with this moderator showed that sex differences in visual-spatial working memory were only significantly different from zero when computer tasks were used.

Before considering potential explanations for this finding, it is important to remember that we had concerns that medium might be confounded with specific task in our sample. An examination of the distribution of effect sizes and their weighted mean as a function of task and testing medium in Table 4 allows an assessment of this possibility. In particular,

Table 4 Distribution of effect sizes as a function of task and testing medium (mean weighted effect size in parenthesis)

Task	Testing medium		
	Computer	Physical	Not reported
Corsi blocks	23 (0.247*)	36 (0.160*)	10 (0.017)
N-back	19 (0.200*)	-	-
Pattern	34 (0.205*)	2 (0.461*)	-
Location	11 (−0.045)	15 (-0.490*)	-
Token	20 (0.288*)	1 (0.181)	-
Other	8 (0.227*)	1 (0.732)	-

*95 % CI does not include zero

it is clear that all tasks are represented in the computer testing category so that any confounding effect does not lie in that group. In contrast, all instances of "not reported" medium come from studies using the Corsi Blocks task, whereas all but the n-back task had at least one effect size drawn from a physical medium. It is difficult to draw any inferences from the Corsi Blocks tasks data where the testing medium was not reported, because we do not really know how participants were tested. However, when considering the data for physical testing material, more than a quarter of the effect sizes (15 of 55, or 27.2 %) come from memory for location tasks. These tasks, when implemented through physical testing, show the largest effect size across Tables 2, 3, and 4; the fact that this estimate is negative likely accounts in part for the effect of testing medium. Specifically, the large negative effect size for location tasks administered physically reduces the effect size to a large extent, as reflected in the overall estimate reported in Table 2 for physical medium. In fact, it is quite revealing that, here as well, removing memory for location tasks from the data set makes the effect of testing medium non-significant in the overall sample ($p = 0.47$). As a result, physical media now produce a significant male advantage when location tasks are removed (estimated $d = 0$ .197; 95 % CI: 0.064, 0.329), contrary to what was originally found (Table 2).

The moderator analyses discussed so far all have one thing in common: When location for memory tasks are removed from the data set, their influence on accounted variance in effect sizes for the overall sample is affected. Specifically, the effect of age coded categorically became significant, whereas the effect of task and testing medium became nonsignificant. These findings suggest that, in one sense, memory for location tasks are outliers in the present data set, and they essentially do not belong with other visual-spatial memory tasks when considering sex differences. From the present findings with location memory tasks and the results reported by Voyer et al. (2007), it appears that females have the advantage over males in memory for location. However, as soon as another component is added, whether it is sequencing, remembering a pattern, a series of location, etc., then males have the advantage. As suggested earlier, empirical studies examining different tasks and their components are needed to explain the mechanism underlying this dissociation in terms of the direction of sex differences.

Year of publication as a moderator Year of publication was a significant moderator for Corsi Blocks tasks and for n-back tasks. The regression coefficient pointed to a decrease in the magnitude of effect sizes with year of publication in the former and an increase in the latter. However, the year effect could be accounted for by age of the sample in n-back tasks, indicating that the original finding was due to the fact that the younger participants also were found in older studies in n-back tasks. This finding simply suggests that age and year of publication are confounded in the n-back tasks sampled here. This confound precludes the need to interpret the year effect in n-back tasks beyond this statistical explanation. It also would be premature to interpret any age-year of publication correlation as reflecting strong trends in the existing research. For example, this correlation is not significant for n-back tasks.

In the case of Corsi Blocks tasks, however, the effect remained significant after controlling for age. This suggests that, similar to what has been found in some spatial tests (Voyer et al., 1995), the magnitude of sex differences in Corsi Blocks tasks has decreased in recent years. Such a meta-analytic finding is typically interpreted as reflecting the influence of social changes on cognitive sex differences (Feingold, 1988). Such an influence seems to be limited to the Corsi Blocks tasks. One might be tempted to argue that the year of publication effect could reflect an Egger-like small-study effect rather than shrinking effects due to social changes. However, this alternative possibility would predict a significant correlation between year of publication and sample size. This correlation is −0.082, $p = 0.276$ in the present study. The social change interpretation therefore seems like a more plausible account of the year of publication effect.

Memory for location, token, pattern, and others Findings of the subgroup analysis showed that the fixed-effects model could not be rejected in the memory for location and memory for token tasks. This testifies to the consistency of the female advantage in location tasks and male advantage for token tasks. The results are particularly interesting in memory for location, because they seem not to fit with the rest of the data, as discussed earlier. It is legitimate to state that such tasks produce significant and consistent sex differences in favor of females.

In memory for pattern and other tasks, no significant moderator emerged. In the case of other tasks, low power due to the small number of effect sizes in that grouping as well as random heterogeneity in what amounted to a catch-all category can plausibly explain this finding. However, the lack of significant moderator in memory for pattern is puzzling, because this grouping has the second most effect sizes in the subgroups analysis. This lack of clear explanatory variable limits the interpretations that we can draw from this type of tasks. However, the magnitude of sex differences that they produce (Table 2) fits nicely with the remaining tasks (except memory for location, of course). Therefore, we can at least conclude that it produces sex differences consistent with what is found in other visual-spatial working memory tasks.

Limitations

The present meta-analysis is the first one to tackle the question of sex differences in visual-spatial working memory. Of course, it also has limitations that require some consideration.

What many readers might view as the most obvious limitation is our shift in literature search strategy as we included *sex* or *gender* as search terms to maximize ready accessibility to data. However, it is important to remember that we were compelled to use this limited search strategy as the literature retrieved from a broader search generally did not report data separately by sex and often did not even mention whether males and females were tested. Obviously, a large majority of researchers interested in the study of visual-spatial working memory overlook sex of participants in their design. Hopefully, the present findings will make them reconsider this view.

One might be tempted to argue that we should have contacted each and every author in the hope of receiving the data needed for our meta-analysis, even though this meant writing to possibly more than 2,000 authors. In all likelihood, assuming that we received timely replies, we would have had a sample biased in other ways. For example, we can reasonably expect that only authors of relatively recent studies would have access to the data, so our sample would be biased toward newer research, which this would potentially affect the results of the year of publication analysis. In addition, only authors that actually noted sex of their participants would be able to send data, so that would be an additional source of bias similar to our selection of sex and gender in the search terms. Of course, we could not expect a 100 % response rate, if only because some authors might have moved, be deceased, or simply choose not to reply. Therefore, it is clear that we would face a different form of bias, and we also would have to delay publication of this work with no guarantee that results would be different.

In fact, the most likely argument against including *sex* or *gender* as search terms is that their inclusion might promote a publication bias, whereby only studies that obtain the expected results (a male advantage?) would get published. It is possible that some authors only mentioned sex differences in their results, because they found them to be significant in hope of having their paper published. This would potentially exacerbate a publication bias. However, it is equally plausible to argue that other authors found significant sex differences in preliminary analyses and failed to report them, because they viewed them as noise. It is simply not possible to determine whether an overreporting or an underreporting bias might exist. However, it is clear that our results showed no support for a publication bias either when comparing the magnitude of sex differences in published and unpublished studies or when considering converging evidence from three other statistical approaches to the examination of publication bias. Therefore, in the context of the existing statistical and methodological tools, it seems reasonable to assume that publication bias is not an issue. However, a more definite answer could be provided by recent initiatives for better scientific practices, including mandatory usage of open data repositories for all published studies. Open data archives would help all researchers conducting meta-analyses, because it would make all relevant data readily available. In the meantime, despite our use of limiting search terms, we are confident that our results are a valid reflection of the state of affairs on sex differences in visual-spatial working memory.

Another limitation requiring discussion is that we had to rely on analytic methods that are not optimal to compute the moderator analysis in four out of six tasks subgroups. This approach reflected a balance between the need to consider all possible level 1 (measure level) moderators to provide as complete a picture as possible of the data and the statistical impossibility to obtain valid results with multilevel meta-analysis in these subgroups due to the small number of samples represented. In view of the questionable validity of the moderator analysis in these subgroups, one might be tempted to argue that we should have performed only the overall multilevel analysis. However, the subgroup analysis was necessary to estimate potential confounds between the specific tasks and some of the moderators. The subgroup analysis established that these confounds likely had minimal effects on the multilevel analysis results, thereby strengthening their validity. In reality, the most serious issue that we need to consider is that the confidence intervals for the effect might be narrower than they should be as standard errors would be underestimated. However, the moderator analysis produced results that paralleled those obtained in the overall analysis, especially when considering age effects. This corroboration suggests that use of a less than optimal analytic method for some of the tasks subgroups had minimal influence on the results.

Conclusions

The present meta-analysis summarized findings pertaining to sex differences in visual-spatial working memory by means of two complementary statistical approaches. It showed that a small but significant male advantage exists in such tasks and that it is fairly consistent across tasks. However, contrary to the hypothesis proposed by some researchers, these sex differences are too small to account fully for sex differences in spatial abilities, where effect sizes can be relatively large, especially in mental rotation. The only exception to the male advantage was in pure memory for location tasks, where a female advantage was observed. The importance of these location tasks in affecting the results of moderator analyses prompted us to suggest that they do not really belong with the remainder of the tasks sampled here, at least in the context of sex differences.

The finding that sex differences in visual-spatial working memory seem to emerge around puberty suggests the need to pursue research investigating possible mechanisms

of this maturation effect. In the meantime, we hope that the present results will make researchers aware of the existence of sex differences in visual-spatial working memory and compel them to at least report the sex composition of their sample if not to consider sex as another relevant variable in this research area.

Acknowledgments This research was made possible by research grants awarded by the Natural Sciences and Engineering Research Council of Canada (NSERC) to D. Voyer and J. Saint-Aubin.

References

Starred articles were included in the meta-analysis

*Aarnoudse-Moens, C., Duivenvoorden, H. J., Weisglas-Kuperus, N., van Goudoever, J. B., & Oosterlaan, J. (2012). The profile of executive function in very preterm children at 4 to 12 years. *Developmental Medicine & Child Neurology, 54*(3), 247–253. doi:10.1111/j.1469-8749.2011.04150.x

*Almela, M., van der Meij, L., Hidalgo, V., Villada, C., & Salvador, A. (2012). The cortisol awakening response and memory performance in older men and women. *Psychoneuroendocrinology, 37*(12), 1929–1940. doi:10.1016/j.psyneuen.2012.04.009

*Alonso-Recio, L., Martín-Plasencia, P., Loeches-Alonso, Á., & Serrano-Rodríguez, J. M. (2014). Working memory and facial expression recognition in patients with Parkinson's disease. *Journal of the International Neuropsychological Society, 20*(5), 496–505. doi:10.1017/S1355617714000265

APA Publications and Communications Board Working Group on Journal Article Reporting Standards. (2008). Reporting Standards for Research in Psychology: Why Do We Need Them? What Might They Be? *American Psychologist, 63,* 839–851. doi:10.1037/0003-066X.63.9.839

*Barrett, S. L., Kelly, C., Bell, R., & King, D. J. (2008). Gender influences the detection of spatial working memory deficits in bipolar disorder. *Bipolar Disorders, 10*(5), 647–654. doi:10.1111/j.1399-5618.2008.00592.x

Bateman, I., & Jones, L. P. (2003). Contrasting conventional with multilevel modeling approaches to meta-analysis: Expectation consistency in U.K. woodland recreation values. *Land Economics, 79,* 235–258.

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics, 50,* 1088–1101. doi:10.2307/2533446

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. (2009). *Introduction to meta-analysis*. doi:10.1002/9780470743386

*Bosco, A., Longoni, A. M., & Vecchi, T. (2004). Gender effects in spatial orientation: Cognitive profiles and mental strategies. *Applied Cognitive Psychology, 18*(5), 519–532. doi:10.1002/acp.1000

*Breitberg, A., Drevets, W. C., Wood, S. E., Mah, L., Schulkin, J., Sahakian, B. J., & Erickson, K. (2013). Hydrocortisone infusion exerts dose- and sex-dependent effects on attention to emotional stimuli. *Brain and Cognition, 81*(2), 247–255. doi:10.1016/j.bandc.2012.10.010

*Bücker, J., Popuri, S., Muralidharan, K., Kozicky, J., Baitz, H. A., Honer, W. G., Yatham, L. N. (2014). Sex differences in cognitive functioning in patients with bipolar disorder who recently recovered from a first episode of mania: Data from the systematic treatment optimization program for early mania (STOP-EM). *Journal of Affective Disorders, 155,* 162–168. doi:10.1016/j.jad.2013.10.044

Cabeza, R., Grady, C. L., Nyberg, L., McIntosh, A. R., Tulving, E., Kapur, S., & Craik, F. I. (1997). Age-related differences in neural activity during memory encoding and retrieval: A positron emission tomography study. *The Journal of Neuroscience, 17,* 391–400.

*Caldwell, L. C., Schweinsburg, A. D., Nagel, B. J., Barlett, V. C., Brown, S. A., & Tapert, S. F. (2005). Gender and adolescent alcohol use disorders on BOLD (blood oxygen level dependent) response to spatial working memory. *Alcohol and Alcoholism, 40*(3), 194–200. doi:10.1093/alcalc/agh134

*Cansino, S., Hernández-Ramos, E., Estrada-Manilla, C., Torres-Trejo, F., Martínez-Galindo, J. G., Ayala-Hernández, M., & Rodríguez-Ortiz, M. D. (2013). The decline of verbal and visuospatial working memory across the adult life span. *Age, 35*(6), 2283–2302. doi:10.1007/s11357-013-9531-1

*Capitani, E., Laiacona, M., & Ciceri, E. (1991). Sex differences in spatial memory: A reanalysis of block tapping long-term memory according to the short-term memory level. *Italian Journal of Neurological Sciences, 12*(5), 461–466. doi:10.1007/BF02335507

*Casey, B. M., Dearing, E., Vasilyeva, M., Ganley, C. M., & Tine, M. (2011). Spatial and numerical predictors of measurement performance: The moderating effects of community income and gender. *Journal of Educational Psychology, 103*(2), 296–311. doi:10.1037/a0022516

Cherrier, M. M., Asthana, S., Plymate, S., Baker, L., Matsumoto, A. M., Peskind, E., & LaTendresse, S. (2001). Testosterone supplementation improves spatial and verbal memory in healthy older men. *Neurology, 57,* 80–88. doi:10.1212/WNL.57.1.80

Christie, G. J., Cook, C. M., Ward, B. J., Tata, M. S., Sutherland, J., Sutherland, R. J., & Saucier, D. M. (2013). Mental rotational ability is correlated with spatial but not verbal working memory performance and P300 amplitude in males. *PLoS One, 8*(2).

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

*Colom, R., Stein, J. L., Rajagopalan, P., Martinez, K., Hermel, D., Wang, Y., & Thompson, P. M. (2013). Hippocampal structure and human cognition: Key role of spatial processing and evidence supporting the efficiency hypothesis in females. *Intelligence, 41*(2), 129–140. doi:10.1016/j.intell.2013.01.002

*Coluccia, E., & Martello, A. (2004). Il ruolo della memoria di lavoro visuo-spaziale nell'orientamento geografico: Uno studio correlazione. (The role of visuospatial working memory in geographical orientation: A correlation study). *Giornale Italiano Di Psicologia, 31*(3), 523–552.

Cooper, J. J. (2006). The digital divide: The special case of gender. *Journal of Computer Assisted Learning, 22,* 320–334. doi:10.1111/j.1365-2729.2006.00185.x

Cowan, N. (2008). What are the differences between long-term, short-term, and working memory? *Progress in Brain Research, 169,* 323–338. doi:10.1016/S0079-6123(07)00020-9

*De Luca, C. R., Wood, S. J., Anderson, V., Buchanan, J. A., Proffitt, T. M., Mahony, K., & Pantelis, C. (2003). Normative data from the CANTAB. I: Development of executive function over the lifespan. *Journal of Clinical and Experimental Neuropsychology, 25*(2), 242–254. doi:10.1076/jcen.25.2.242.13639

Duff, S. J., & Hampson, E. (2000). A beneficial effect of estrogen on working memory in postmenopausal women taking hormone replacement therapy. *Hormones and Behavior, 38,* 262–276. doi:10.1006/hbeh.2000.1625

*Duff, S. J., & Hampson, E. (2001). A sex differences on a novel spatial working memory task in humans. *Brain and Cognition, 47*(3), 470–493. doi:10.1006/brcg.2001.1326

Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal, 315,* 629–634. doi:10.1136/bmj.315.7109.629

*Evardone, M., & Alexander, G. M. (2009). Anxiety, sex-linked behaviors, and digit ratios (2D:4D). *Archives of Sexual Behavior, 38*(3), 442–455. doi:10.1007/s10508-007-9260-6

Feingold, A. (1988). Cognitive gender differences are disappearing. *American Psychologist, 43,* 95–103. doi:10.1037/0003-066X.43.2.95

*Fikke, L. T., Melinder, A., & Landrø, N. I. (2011). Executive functions are impaired in adolescents engaging in non-suicidal self-injury. *Psychological Medicine, 41*(3), 601–610. doi:10.1017/S0033291710001030

*Flannery, B., Fishbein, D., Krupitsky, E., Langevin, D., Verbitskaya, E., Bland, C., & Zvartau, E. (2007). Gender differences in neurocognitive functioning among alcohol-dependent Russian patients. *Alcoholism: Clinical and Experimental Research, 31*(5), 745–754. doi:10.1111/j.1530-0277.2007.00372.x

*Fournet, N., Roulin, J., Vallet, F., Beaudoin, M., Agrigoroaei, S., Paignon, A., & Desrichard, O. (2012). Evaluating short-term and working memory in older adults: French normative data. *Aging & Mental Health, 16*(7), 922–930. doi:10.1080/13607863.2012.674487

*Geiger, J. F., & Litwiller, R. M. (2005). Spatial working memory and gender differences in science. *Journal of Instructional Psychology, 32*(1), 49–57.

*Girard, T.A. (2014). Unpublished raw data.

*Girard, T.A., Christensen, B.K., & Rizvi, S. (2010). Visual-spatial episodic memory in Schizophrenia: A multiple systems framework. *Neuropsychology, 24*(3), 368–378. doi:10.1037/a0018313

Halpern, D. F. (2013). *Sex differences in cognitive abilities*. New York: Psychology Press.

*Hampson, E., & Morley, E.E. (2013). Estradiol concentrations and working memory performance in women of reproductive age. *Psychoneuroendocrinology, 38,* 2897–2904. doi:10.1016/j.psyneuen.2013.07.020

*Hartley, D. E., Elsabagh, S., & File, S. E. (2004). Binge drinking and sex: Effects on mood and cognitive function in healthy young volunteers. *Pharmacology, Biochemistry and Behavior, 78*(3), 611–619. doi:10.1016/j.pbb.2004.04.027

*Hayward, E.N. (2014). *Places and objects within a virtual environment: An ecological investigation of spatial memory* (Unpublished undergraduate thesis). University of Winnipeg: Manitoba, Canada.

Hedges, L. V., & Becker, B. J. (1986). Statistical methods in the meta-analysis of research on gender differences. In J. Hyde & M. C. Linn (Eds.), *The psychology of gender: Advance through meta-analysis* (pp. 14–50). Baltimore: John Hopkins University Press.

*Hegarty, M., Montello, D. R., Richardson, A. E., Ishikawa, T., & Lovelace, K. (2006). Spatial abilities at different scales: Individual differences in aptitude-test performance and spatial-layout learning. *Intelligence, 34*(2), 151–176. doi:10.1016/j.intell.2005.09.005

*Hernández-Balderas, M. Á., Rángel-Félix, G., Zavala-González, J. C., Romero-Romero, H., Silva-Pereyra, J. F., del Rio-Portilla, I. Y., & Bernal-Hernández, J. (2012). Sex differences in the visuospatial sketchpad in scholar children. *Journal of Behavior, Health & Social Issues, 4*(2), 103–115. doi:10.5460/jbhsi.v4.2.34111

*Hoesing, J. M. (1998). An evaluation of sex differences in processing of visual information for features and locations. *Dissertation Abstracts International: Section B: The Sciences and Engineering, 59*(7-), 3725–3725.

Hox, J. J. (2008). Accommodating measurement errors. In E. D. de Leeuw, J. J. Hox, & D. A. Dillman (Eds.), *The International Handbook of Survey Methodology* (pp. 387–402). New York/London: Erlbaum/Taylor & Francis.

Hox, J. J., & de Leuw, E. D. (2003). Multilevel models for meta-analysis. In S. P. Reise & N. Duan (Eds.), *Multilevel modeling: Methodological advances, issues, and applications* (pp. 90–111). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Hyde, J. S. (2014). Gender similarities and differences. *Annual Review of Psychology, 65,* 373–398. doi:10.1146/annurev-psych-010213-115057

Hyun, J. S., & Luck, S. J. (2007). Visual working memory as the substrate for mental rotation. *Psychonomic Bulletin & Review, 14*(1), 154–158. doi:10.3758/BF03194043

Ioannidis, J. P. A., & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials, 4,* 245–253.

Johnson, J. (2013). *Designing with the mind in mind: Simple guide to understanding user interface design guidelines* (2nd ed.). New York: Elsevier.

*Kalmady, S. V., Agarwal, S. M., Shivakumar, V., Jose, D., Venkatasubramanian, G., & Reddy, Y. C. J. (2013). Revisiting Geschwind's hypothesis on brain lateralisation: A functional MRI study of digit ratio (2D:4D) and sex interaction effects on spatial working memory. *Laterality: Asymmetries of Body, Brain and Cognition, 18*(5), 625–640. doi:10.1080/1357650X.2012.744414

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

Kay, R. (2006). Addressing gender differences in computer ability, attitudes and use: The laptop effect. *Journal of Educational Computing Research, 34,* 187–211.

*Kokubo, N., Inagaki, M., Gunji, A., Kobayashi, T., Ohta, H., Kajimoto, O., & Kaga, M. (2012). Developmental change of visuo-spatial working memory in children: Quantitative evaluation through an advanced trail making test. *Brain & Development, 34*(10), 799–805. doi:10.1016/j.braindev.2012.02.001

*Krikorian, R., Bartok, J. A., & Gay, N. (1996). Immediate memory capacity for nonsequential information: The configural attention test. *Neuropsychology, 10*(3), 352–356. doi:10.1037/0894-4105.10.3.352

*Krinzinger, H., Wood, G., & Willmes, K. (2012). What accounts for individual and gender differences in the multi-digit number processing of primary school children? *Zeitschrift Für Psychologie, 220*(2), 78–89. doi:10.1027/2151-2604/a000099

*Kuhn, J., & Holling, H. (2014). Number sense or working memory? The effect of two computer based trainings on mathematical skills in elementary school. *Advances in Cognitive Psychology, 10*(2), 59–67. doi:10.5709/acp-0157-2

*Lawton, C. A., & Hatcher, D. W. (2005). Gender differences in integration of images in visuospatial memory. *Sex Roles, 53*(9-10), 717–725. doi:10.1007/s11199-005-7736-1

*Lejbak, L., Crossley, M., & Vrbancic, M. (2011). A male advantage for spatial and object but not verbal working memory using the N-back task. *Brain and Cognition, 76*(1), 191–196. doi:10.1016/j.bandc.2010.12.002

*Lejbak, L., Vrbancic, M., & Crossley, M. (2009). The female advantage in object location memory is robust to verbalizability and mode of presentation of test stimuli. *Brain and Cognition, 69*(1), 148–153. doi:10.1016/j.bandc.2008.06.006

*León, I., Cimadevilla, J. M., & Tascón, L. (2014). Developmental gender differences in children in a virtual spatial memory task. *Neuropsychology, 28*(4), 485–495. doi:10.1037/neu0000054

*Levin, S. L., Mohamed, F. B., & Platek, S. M. (2005). Common ground for spatial cognition? A behavioral and fMRI study of sex differences in mental rotation and spatial working memory. *Evolutionary Psychology, 3,* 227–254.

Linn, M. C., & Petersen, A. C. (1985). Emergence and characterization of gender differences in spatial abilities: A meta-analysis. *Child Development, 56,* 1479–1498. doi:10.1111/1467-8624.ep7252392

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Logie, R. H. (1995). *Visuo-spatial working memory*. Hillsdale, NJ: Erlbaum.

Loring-Meier, S., & Halpern, D. F. (1999). Sex differences in visuospatial working memory: Components of cognitive processing. *Psychonomic Bulletin & Review, 6*(3), 464–471. doi:10.3758/BF03210836

MacLeod, C. M., & Kampe, K. E. (1996). Word frequency effects on recall, recognition, and word fragment completion tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 132. doi:10.1037/0278-7393.22.1.132

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. doi:10.1007/s10648-012-9215-x

*Mammarella, I. C., Lucangeli, D., & Cornoldi, C. (2010). Spatial working memory and arithmetic deficits in children with nonverbal learning difficulties. *Journal of Learning Disabilities, 43*(5), 455–468. doi:10.1177/0022219409355482

*Martin, G. N., & Chaudry, A. (2014). Working memory performance and exposure to pleasant and unpleasant ambient odor: Is spatial span special? *International Journal of Neuroscience, 124*(11), 806–811. doi:10.3109/00207454.2014.890619

*Miller, C. R. (2003). Individual differences in object-location learning: How is working memory related? *Dissertation Abstracts International: Section B: The Sciences and Engineering, 64*(12-), 6350–6350.

*Miller, D. I., & Halpern, D. F. (2013). Can spatial training improve long-term outcomes for gifted STEM undergraduates?. *Learning and Individual Differences, 26,* 141–152. doi:10.1016/j.lindif.2012.03.012

*Minor, K., & Park, S. (1999). Spatial working memory: Absence of gender differences in schizophrenia patients and healthy control subjects. *Biological Psychiatry, 46*(7), 1003–1005. doi:10.1016/S0006-3223(99)00149-3

*Nalçaci, E., Çiçek, M., Kalaycioglu, C., & Yavuzer, S. (1997). Pseudoneglect of males and females on a spatial short-term memory task. *Perceptual and Motor Skills, 84*(1), 99–105. doi:10.2466/PMS.84.1.99-105

*Nalçaci, E., Kalaycioglu, C., Çıçek, M., & Budanur, Ö. E. (2000). Magical ideation and right-sided hemispatial inattention on a spatial working memory task: Influences of sex and handedness. *Perceptual and Motor Skills, 91*(3), 883–892. doi:10.2466/PMS.91.7.883-892

*Orsini, A., Chiacchio, L., Cinque, M., Cocchiaro, C., Schiappa, O., & Grossi, D. (1986). Effects of age, education and sex on two tests of immediate memory: A study of normal subjects from 20 to 99 years of age. *Perceptual and Motor Skills, 63*(2), 727–732. doi:10.2466/pms.1986.63.2.727

*Pagulayan, K. F., Busch, R. M., Medina, K. L., Bartok, J. A., & Krikorian, R. (2006). Developmental normative data for the Corsi block-tapping task. *Journal of Clinical and Experimental Neuropsychology, 28*(6), 1043–1052. doi:10.1080/13803390500350977

*Pangelinan, M. M., Zhang, G., VanMeter, J. W., Clark, J. E., Hatfield, B. D., & Haufler, A. J. (2011). Beyond age and gender: Relationships between cortical and subcortical brain volume and cognitive-motor abilities in school-age children. *Neuroimage, 54*(4), 3093–3100. doi:10.1016/j.neuroimage.2010.11.021

*Piccardi, L., Palermo, L., Leonzi, M., Risetti, M., Zompanti, L., D'Amico, S., & Guariglia, C. (2014). The walking corsi test (WalCT): A normative study of topographical working memory in a sample of 4- to 11-year-olds. *The Clinical Neuropsychologist, 28*(1), 84–96. doi:10.1080/13854046.2013.863976

*Postma, A., Jager, G., Kessels, R. P. C., Koppeschaar, H. P. F., & van Honk, J. (2004). Sex differences for selective forms of spatial memory. *Brain and Cognition, 54*(1), 24–34. doi:10.1016/S0278-2626(03)00238-0

*Postma, A., Winkel, J., Tuiten, A., & van Honk, J. (1999). Sex differences and menstrual cycle effects in human spatial memory. *Psychoneuroendocrinology, 24*(2), 175–192. doi:10.1016/S0306-4530(98)00073-0

*Price, I.L. (2009). Visuospatial reasoning in toddlers: A correlational study of door task performance. *Doctoral Dissertations Available from Proquest.* Paper AAI3359156.

Prime, D. J., & Jolicoeur, P. (2010). Mental rotation requires visual short-term memory: Evidence from human electric cortical activity. *Journal of Cognitive Neuroscience, 22,* 2437–2446. doi:10.1162/jocn.2009.21337

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA: Sage.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., Congdon, R. T., Jr., & du Toit, M. (2011). *HLM 7: Hierarchical linear and nonlinear modeling User's Manual*. Lincolnwood, IL: Scientific Software International.

*Roesch-Ely, D., Hornberger, E., Weiland, S., Hornstein, C., Parzer, P., Thomas, C., & Weisbrod, M. (2009). Do sex differences affect prefrontal cortex associated cognition in schizophrenia? *Schizophrenia Research, 107*(2-3), 255–261. doi:10.1016/j.schres.2008.09.021

Rosenthal, R. (1979). The "file drawer problem" and the tolerance for null results. *Psychological Bulletin, 86,* 638–641.

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (revth ed.). Beverly Hills, CA: Sage.

*Rubin, L. H. (2009). Effects of sex steroid hormones on cognition in schizophrenia. *Dissertation Abstracts International: Section B: The Sciences and Engineering, 71*(2-), 1353–1353.

*Ruggiero, G., Sergi, I., & Iachini, T. (2008). Gender differences in remembering and inferring spatial distances. *Memory, 16*(8), 821–835. doi:10.1080/09658210802307695

*Savage, L. (2013). *Near and far transfer of working memory training related gains in healthy adults* (Unpublished master's thesis). University of Calgary, Alberta, Canada.

Scherer, R., & Siddiq, F. (2015). Revisiting teachers' computer self-efficacy: A differentiated view on gender differences. *Computers in Human Behavior, 53,* 48–57. doi:10.1016/j.chb.2015.06.038

*Schweinsburg, A. D., Nagel, B. J., & Tapert, S. F. (2005). fMRI reveals alteration of spatial working memory networks across adolescence. *Journal of the International Neuropsychological Society, 11*(5), 631–644. doi:10.1017/S1355617705050757

*Seghete, K. L. M., Cservenka, A., Herting, M. M., & Nagel, B. J. (2013). Atypical spatial working memory and task-general brain activity in adolescents with a family history of alcoholism. *Alcoholism: Clinical and Experimental Research, 37*(3), 390–398. doi:10.1111/j.1530-0277.2012.01948.x

*Shikhman, M. (2007). Age, gender, general intelligence and educational level influences on working memory. *Dissertation Abstracts International: Section B: The Sciences and Engineering, 68*(10-), 6994–6994.

Sieverding, M., & Koch, S. C. (2009). (Self-) Evaluation of computer competence: How gender matters. *Computers & Education, 52,* 696–701. doi:10.1016/j.compedu.2008.11.016

Silverman, I., & Eals, M. (1992). Sex differences in spatial abilities: Evolutionary theory and data. In J. Barkow, I. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 533–549). New York: Oxford University Press.

*Squeglia, L. M., Schweinsburg, A. D., Pulido, C., & Tapert, S. F. (2011). Adolescent binge drinking linked to abnormal spatial working memory brain activation: Differential gender effects. *Alcoholism: Clinical and Experimental Research, 35,* 1831–1841. doi:10.1111/j.1530-0277.2011.01527.x

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance— or vice versa. *Journal of the American Statistical Association, 54,* 30–34.

Stern, J. A. C., & Egger, M. (2005). Regression methods to detect publication and other bias in meta-analysis. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis* (pp. 99–110). West Sussex, UK: Wiley.

*Szabo, A. N., McAuley, E., Erickson, K. I., Voss, M., Prakash, R. S., Mailey, E. L., & Kramer, A. F. (2011). Cardiorespiratory fitness, hippocampal volume, and frequency of forgetting in older adults. *Neuropsychology, 25,* 545–553. doi:10.1037/a0022733

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston, MA: Allyn & Bacon.

Techentin, C. L., Voyer, D., & Voyer, S. D. (2014). Spatial abilities and aging: A Meta-analysis. *Experimental Aging Research, 40,* 395–425. doi:10.1080/0361073X.2014.926773

*Teixeira, R. A., Zachi, E. C., Roque, D. T., Taub, A., & Ventura, D. F. (2011). Memory span measured by the spatial span tests of the Cambridge Neuropsychological Test Automated Battery in a group of Brazilian children and adolescents. *Dementia & Neuropsychologia, 5*(2), 129–134

Thomas, K. M., King, S. W., Franzen, P. L., Welsh, T. F., Berkowitz, A. L., Noll, D. C., & Casey, B. J. (1999). A developmental functional MRI study of spatial working memory. *NeuroImage, 10,* 327–338. doi:10.1006/nimg.1999.0466

Thomason, M. E., Race, E., Burrows, B., Whitfield-Gabrieli, S., Glover, G. H., & Gabrieli, J. D. (2009). Development of spatial and verbal working memory capacity in the human brain. *Journal of Cognitive Neuroscience, 21,* 316–332. doi:10.1162/jocn.2008.21028

Torgrimson, B. N., & Minson, C. T. (2005). Sex and gender: What is the difference? *Journal of Applied Physiology, 99,* 785–787. doi:10.1152/japplphysiol.00376.2005

Unsworth, N., Redick, T. S., Heitz, R. P., Broadway, J. M., & Engle, R. W. (2009). Complex working memory span tasks and higher-order cognition: A latent-variable analysis of the relationship between processing and storage. *Memory, 17,* 635–654. doi:10.1080/09658210902998047

Vandenberg, S., & Kuse, A. (1978). Mental rotation, a group test of 3-D spatial visualization. *Perceptual and Motor Skills, 47,* 599–604. doi:10.2466/pms.1978.47.2.599

Vecchi, T., Phillips, L. H., & Cornoldi, C. (2001). Individual differences in visuo-spatial working memory. In M. Denis, R. H. Logie, C. Cornoldi, M. de Vega, & J. Engelkamp (Eds.), *Imagery, language and visuospatial thinking* (pp. 29–58). New York, NY, US: Psychology Press.

*Verkade, E, Dorrestijn, M., & Plat, M. (2011). Genderverschillen in het werkgeheugen [Translated title: Gender differences in working memory]. Bachelor Thesis.

*Visu-Petra, L., Cheie, L., & Benga, O. (2008). Short-term memory performance and metamemory judgments in preschool and early school-age children: A quantitative and qualitative analysis. *Cogniţie Creier Comportament, 12,* 71–101.

*Vock, M., & Holling, H. (2008). The measurement of visuo–spatial and verbal–numerical working memory: Development of IRT-based scales. *Intelligence, 36*(2), 161–182. doi:10.1016/j.intell.2007.02.004

Voyer, D., Postma, A., Brake, B., & Imperato-McGinley, J. (2007). Gender differences in object location memory: A meta-analysis. *Psychonomic Bulletin and Review, 14,* 23–38. doi:10.3758/BF03194024

Voyer, D., & Voyer, S. D. (2014). Gender differences in scholastic achievement: A meta-analysis. *Psychological Bulletin, 140,* 1174–1204. doi:10.1037/a0036620 [NSERC]

Voyer, D., & Voyer, S. D. (2015). *Psychology: The science of undergraduate women.* Paper presented at the meeting of the Psychonomic Society, Chicago, IL, November 2015.

Voyer, D., Voyer, S., & Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. *Psychological Bulletin, 117,* 250–270. doi:10.1037/0033-2909.117.2.250

Wang, L., & Carr, M. (2014). Working Memory and Strategy Use Contribute to Gender Differences in Spatial Ability. *Educational Psychologist, 49*(4), 261–282. doi:10.1080/00461520.2014.960568

Wechsler, D. (1997). *WAIS-III administration and scoring manual*. San Antonio, TX: Psychological Corporation.

*Weisberg, S. M., & Newcombe, N. S. (2014). *How do (some) people make a cognitive map? Routes, places, and working memory.* Manuscript submitted for publication.

Wilson, D. B. (2005). *Meta-analysis macros for SAS, SPSS, and Stata*. Retrieved from http://mason.gmu.edu/~dwilsonb/ma.html

*Wong, L. M., Riggins, T., Harvey, D., Cabaral, M., & Simon, T. J. (2014). Children with chromosome 22q11.2 deletion syndrome exhibit impaired spatial working memory. *American Journal on Intellectual and Developmental Disabilities, 119*(2), 115–132. doi:10.1352/1944-7558-119.2.115

*Yerys, B. E., Wallace, G. L., Sokoloff, J. L., Shook, D. A., James, J. D., & Kenworthy, L. (2009). Attention deficit/hyperactivity disorder symptoms moderate cognition and behavior in children with autism spectrum disorders. *Autism Research*, *2*(6), 322–333. doi:10.1002/aur.103