Gender differences in infant mental rotation

Alexander Enge[1,2,*], Shreya Kapoor[1], Anne-Sophie Kieslinger[1], and Michael A. Skeide[1]

[1] Research Group Learning in Early Childhood, Max Planck Institute for Human Cognitive and Brain Sciences, Stephanstraße 1a, Leipzig, Germany

[2] Department of Psychology, Humboldt-Universität zu Berlin, Rudower Chaussee 18, Berlin, Germany

[*] Corresponding author (enge@cbs.mpg.de)

Abstract

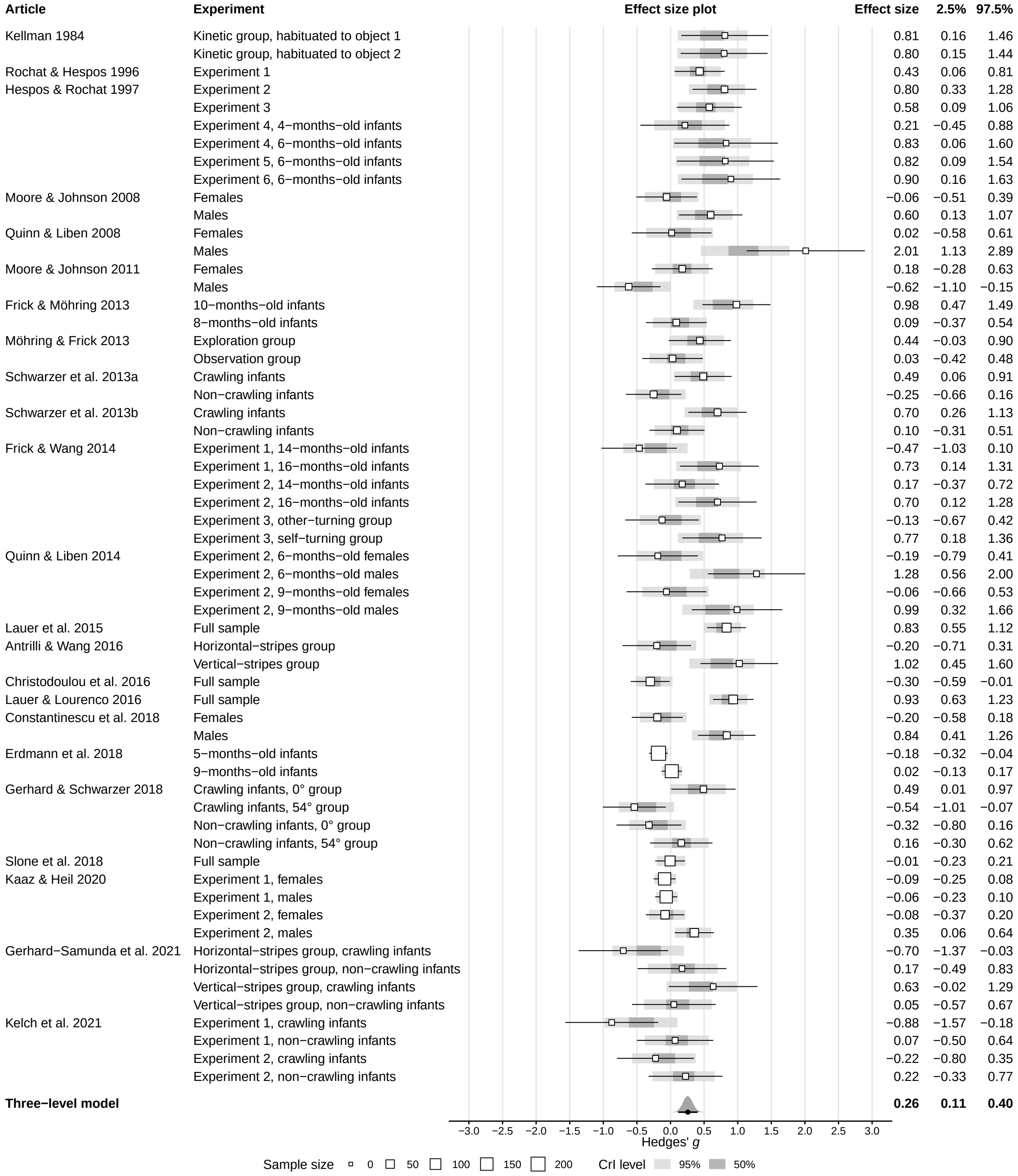
Mental rotation, the cognitive process of moving an object in mind to predict how it looks in a new orientation, is tightly coupled to intelligence, learning, and educational achievement[1–3]. On average, males solve mental rotation tasks slightly faster than females[4–7]. When such behavioral differences emerge during development, however, remains poorly understood[8]. Here we analyzed effect sizes derived from 59 experiments conducted in 1,798 infants aged 3–16 months. We robustly found that male infants recognized rotated objects slightly more reliably than female infants. These findings indicate that subtle gender differences in mental rotation are already present in the first months of life.

Main

The cognitive ability to move visual object representations in mind for recognition across different orientations, known as mental rotation, emerges in the first three months of life[9,10]. Mental rotation is a key component of intelligence and a powerful predictor of learning outcome and educational achievement[1–3].

Previous meta-analyses revealed that males solve mental rotation tasks slightly faster than females on average[4–7]. Effect sizes of this difference, however, are heterogeneous and often only medium in size (mean weighted $g = 0.37$–0.73). Interestingly, a recent meta-analysis in 3–17-year-old children and adolescents suggests only a small-to-medium difference (mean weighted $g = 0.39$)[8]. Whether gender differences in mental rotation behavior already emerge during infancy remains unknown.

In the present study, we meta-analyzed 59 effect sizes derived from looking times in mental rotation tasks performed by 1,798 infants (47.5% female) aged 3 to 16 months (mean age of 7 months 14 days; Supplementary Table 1)[11–33]. All tasks were embedded either into habituation experiments (40) or violation of expectation experiments (19)[34,35]. These experiments comprised real world stimuli (e.g., toy objects)[24], three-dimensional digital stimuli (e.g., cube figures)[14], or two-dimensional digital stimuli (e.g., digits)[15]. In habituation experiments, infants repeatedly saw an object until their looking times declined before they were presented with a mirror image of the object or with the familiar object at a new angle. Longer looking times at the mirror image were taken as evidence that an infant still recognized the familiar object after rotation through the new orientation. In violation of expectation experiments, infants were habituated to an object that was revolving repeatedly through a certain angle. Then, either the familiar object or an unseen object was shown while they were revolving through a certain angle. Subsequently, the object disappeared behind an occluder. Finally, the occluder was removed and either the object or its mirror object were shown. Larger differences in looking times for the familiar object versus the unseen mirror image (at the new angle) were taken as evidence that an infant still recognized the familiar object after rotation through the new angle. In contrast, looking times for both objects were expected to be similar if an infant did not recognize the familiar object from the new angle.



Article	Experiment	Effect size	2.5%	97.5%
Kellman 1984	Kinetic group, habituated to object 1	0.81	0.16	1.46
	Kinetic group, habituated to object 2	0.80	0.15	1.44
Rochat & Hespos 1996	Experiment 1	0.43	0.06	0.81
Hespos & Rochat 1997	Experiment 2	0.80	0.33	1.28
	Experiment 3	0.58	0.09	1.06
	Experiment 4, 4−months−old infants	0.21	−0.45	0.88
	Experiment 4, 6−months−old infants	0.83	0.06	1.60
	Experiment 5, 6−months−old infants	0.82	0.09	1.54
	Experiment 6, 6−months−old infants	0.90	0.16	1.63
Moore & Johnson 2008	Females	−0.06	−0.51	0.39
	Males	0.60	0.13	1.07
Quinn & Liben 2008	Females	0.02	−0.58	0.61
	Males	2.01	1.13	2.89
Moore & Johnson 2011	Females	0.18	−0.28	0.63
	Males	−0.62	−1.10	−0.15
Frick & Möhring 2013	10−months−old infants	0.98	0.47	1.49
	8−months−old infants	0.09	−0.37	0.54
Möhring & Frick 2013	Exploration group	0.44	−0.03	0.90
	Observation group	0.03	−0.42	0.48
Schwarzer et al. 2013a	Crawling infants	0.49	0.06	0.91
	Non−crawling infants	−0.25	−0.66	0.16
Schwarzer et al. 2013b	Crawling infants	0.70	0.26	1.13
	Non−crawling infants	0.10	−0.31	0.51
Frick & Wang 2014	Experiment 1, 14−months−old infants	−0.47	−1.03	0.10
	Experiment 1, 16−months−old infants	0.73	0.14	1.31
	Experiment 2, 14−months−old infants	0.17	−0.37	0.72
	Experiment 2, 16−months−old infants	0.70	0.12	1.28
	Experiment 3, other−turning group	−0.13	−0.67	0.42
	Experiment 3, self−turning group	0.77	0.18	1.36
Quinn & Liben 2014	Experiment 2, 6−months−old females	−0.19	−0.79	0.41
	Experiment 2, 6−months−old males	1.28	0.56	2.00
	Experiment 2, 9−months−old females	−0.06	−0.66	0.53
	Experiment 2, 9−months−old males	0.99	0.32	1.66
Lauer et al. 2015	Full sample	0.83	0.55	1.12
Antrilli & Wang 2016	Horizontal−stripes group	−0.20	−0.71	0.31
	Vertical−stripes group	1.02	0.45	1.60
Christodoulou et al. 2016	Full sample	−0.30	−0.59	−0.01
Lauer & Lourenco 2016	Full sample	0.93	0.63	1.23
Constantinescu et al. 2018	Females	−0.20	−0.58	0.18
	Males	0.84	0.41	1.26
Erdmann et al. 2018	5−months−old infants	−0.18	−0.32	−0.04
	9−months−old infants	0.02	−0.13	0.17
Gerhard & Schwarzer 2018	Crawling infants, 0° group	0.49	0.01	0.97
	Crawling infants, 54° group	−0.54	−1.01	−0.07
	Non−crawling infants, 0° group	−0.32	−0.80	0.16
	Non−crawling infants, 54° group	0.16	−0.30	0.62
Slone et al. 2018	Full sample	−0.01	−0.23	0.21
Kaaz & Heil 2020	Experiment 1, females	−0.09	−0.25	0.08
	Experiment 1, males	−0.06	−0.23	0.10
	Experiment 2, females	−0.08	−0.37	0.20
	Experiment 2, males	0.35	0.06	0.64
Gerhard−Samunda et al. 2021	Horizontal−stripes group, crawling infants	−0.70	−1.37	−0.03
	Horizontal−stripes group, non−crawling infants	0.17	−0.49	0.83
	Vertical−stripes group, crawling infants	0.63	−0.02	1.29
	Vertical−stripes group, non−crawling infants	0.05	−0.57	0.67
Kelch et al. 2021	Experiment 1, crawling infants	−0.88	−1.57	−0.18
	Experiment 1, non−crawling infants	0.07	−0.50	0.64
	Experiment 2, crawling infants	−0.22	−0.80	0.35
	Experiment 2, non−crawling infants	0.22	−0.33	0.77
Three−level model		**0.26**	**0.11**	**0.40**

Fig. 1 | Mental rotation performance. A Bayesian three-level meta-analysis provided evidence for mental rotation ability in infants. White squares depict the effect sizes (Hedges' g) for infants' mental rotation performance in all individual experiments and black lines depict their 95% confidence intervals. Gray bars indicate the 50% and 95% credible intervals (CrI) from the Bayesian model, which take into account that experiments with smaller sample sizes or more extreme effect sizes provide less reliable information. The last line shows the meta-analytic effect size (black dot) together with its 95% CrI (black line) and its posterior distribution (gray curve).

Following previous work in older populations, we hypothesized that looking time differences in mental rotation experiments are on average longer in male compared to female infants. Effect sizes were assumed to be small.

Mental rotation performance

For our first meta-analytic model, effect sizes were quantified as the standardized mean difference in infants' looking times for novel and familiar rotated objects. Using this effect size index, we ran a Bayesian three-level random-effects model to test if there was evidence that infants did perform mental rotation. Across studies, infants indeed looked longer at novel rotated objects than at familiar rotated objects, with a standardized mean difference of $g = 0.26$, 95% credible interval (CrI) [0.11, 0.40] (Fig. 1). The probability for this effect being greater than zero was $> 99.9\%$. The heterogeneity of effect sizes was $\sigma^2_{\text{experiment}} = 0.17$, 95% CrI [0.09, 0.29], at the experiment level and $\sigma^2_{\text{article}} = 0.03$, 95% CrI [0.00, 0.11] at the article level. Therefore, 86.6% of the heterogeneity between effect sizes was attributable to differences between experiments *within* articles and 13.4% was attributable to differences *between* articles.

Effects of gender, age, and task type

As a next step, we conducted a meta-regression analysis to test if the gender of the infants, their age, or the type of mental rotation task was related to mental rotation performance (Fig. 2). Indeed, experiments with all-male samples revealed larger looking time differences than experiments with all-female samples, $b = 0.49$ (where $b = \Delta g$), 95% CrI [0.07, 0.90]. The probability for this effect being larger than zero was 98.9%. We found no difference between mixed-gender and all-female samples, $b = 0.14$, 95% CrI [0.52, -0.24] (76.8% probability of an effect larger than zero). Additionally, mean age was not related to mental rotation performance, with a change per year of $b = 0.05$, 95% [-0.58, 0.67]. We also did not detect an interaction between gender and age ([females - mixed] × age: $b = 0.01$, 95% CrI [-0.81, 0.83]; [males - females] × age: $b = 0.06$, 95% CrI [-0.84, 0.97]). Finally, there was weak evidence that violation of expectation tasks yielded larger effects than habituation tasks, $b = 0.25$, 95% [-0.08, 0.58]. The probability for this effect being greater than zero was 93.5%.

To confirm the gender difference between males and females, we set up another Bayesian meta-analysis, this time focusing on the looking-time contrasts between male and female infants reported *within* each experiment by the original authors. Our additional analysis revealed a meta-analytic effect size of $g = 0.14$, 95% CrI [-0.01, 0.31] (Supplementary Fig. 1) and a probability of this effect being greater than zero of 96.4%. The heterogeneity of effect sizes was $\sigma^2_{\text{experiment}} = 0.02$, 95% CrI [0.00, 0.10], at the experiment level and $\sigma^2_{\text{article}} = 0.04$, 95% CrI [0.00, 0.17] at the article level. Therefore, 44.0% of the heterogeneity between effect sizes was attributable to differences between experiments *within* articles and 56.0% was attributable to differences *between* articles.

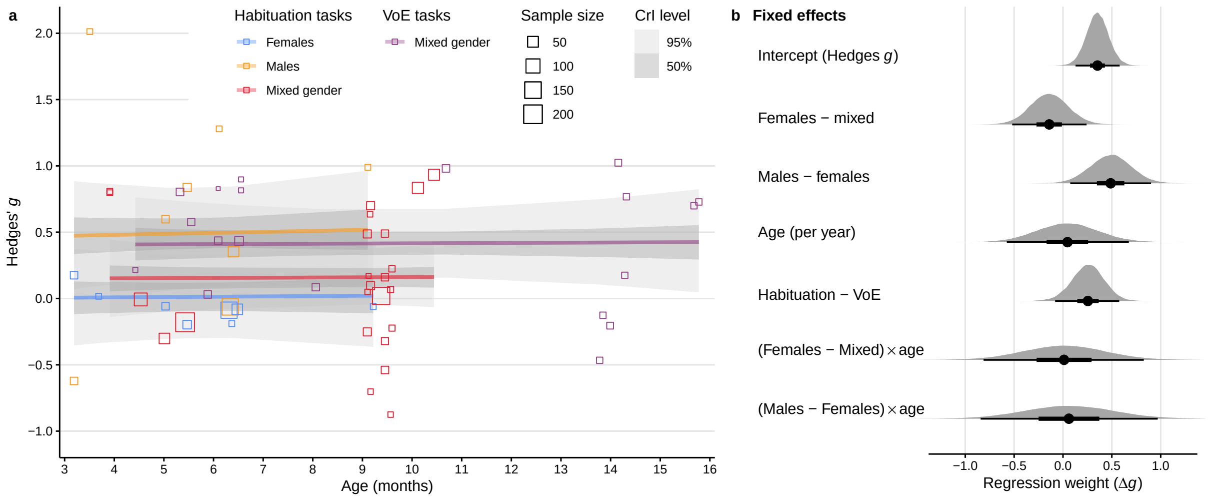


Fig. 2 | Effects of gender, age, and task. a, Squares show the effect size (Hedges' g) for infants' mental rotation performance in each of the 59 individual experiments. Squares are color-coded according to the type of habituation task and the gender of the infants (blue = habituation task, all-female sample, yellow = habituation task, all-male sample, red = habituation task, mixed-gender sample, purple = violation of expectation [VoE] task, mixed-gender sample). Lines indicate the best-fit regression estimates according to a Bayesian three-level meta-regression model and gray ribbons indicate their corresponding 50% and 95% credible interval (CrI). **b,** Fixed effect estimates obtained from the Bayesian three-level meta-regression model are depicted as black dots together with their 50% CrI (thick black lines) and 95% CrI (thin black lines). Gray curves indicate the posterior distribution for each effect.

The results obtained from our Bayesian analyses were reproduced by classical frequentist three-level meta-analysis and meta-regression models (Supplementary Table 2).

Publication bias assessment

We inspected funnel plots and performed Egger regression tests to examine the possibility of publication bias in the literature included here. For the meta-analysis of mental rotation performance (Fig. 3a), we observed a slight asymmetry in the funnel plot, indicating a small publication bias. This was confirmed by an Egger regression test that showed a reliable association between effect sizes and their corresponding standard errors, $b = 1.85$, $t(57) = 3.37$, $P = 0.001$, 95% confidence interval (CI) [0.75, 2.95] (two-sided test). Nevertheless, a jackknife (leave-one-out) analysis confirmed that the current results are robust to the effects of individual outlier experiments (Supplementary Table 3). For the meta-analysis of gender differences in mental rotation performance (Fig. 3b), the asymmetry in the funnel plot was less pronounced and the slope of the Egger regression test was not statistically significant, $b = 0.32$, $t(28) = 0.65$, $P = 0.522$, 95% CI [-0.69, 1.32] (two-sided test).

Discussion

We analyzed looking times during mental rotation in 1,798 infants ranging from 3 to 16 months of age. To this end, we scrutinized the robustness of 59 experimental effect sizes. We found that on average, male infants looked longer at novel rotated objects compared to female infants. This effect was small to medium and unrelated to age in the current range.

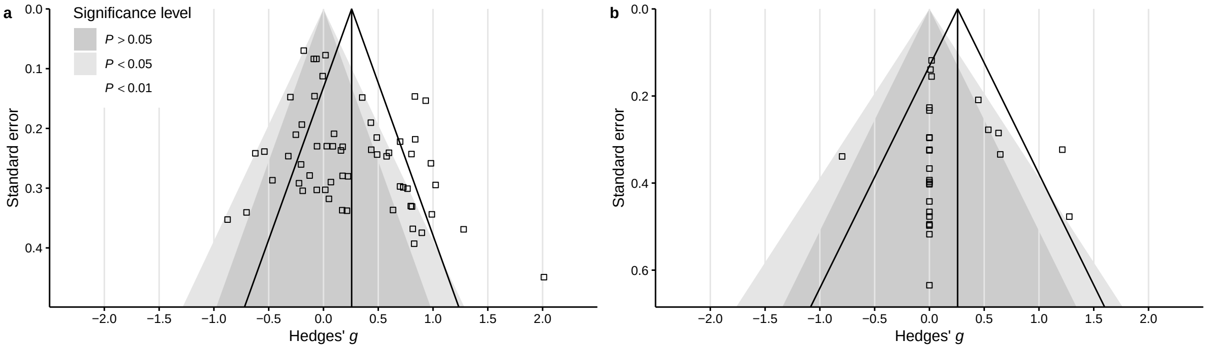


Fig. 3 | Evaluation of publication bias. Funnel plots for the meta-analysis of mental rotation performance (**a**) and for the meta-analysis of gender differences in mental rotation performance (**b**). The plots show the standard error and effect size for each of the individual experiments as a black square. The funnel contours (diagonal black lines) depict a 95% pseudo-confidence interval around the meta-analytic effect size (vertical black line). Gray shades indicate a 95% pseudo-confidence interval (dark gray) and a 99% pseudo-confidence interval (light gray) under the null hypothesis. These shades thus illustrate which of the original experiments observed a significant effect. For the meta-analysis of mental rotation performance, a slight asymmetry induced by the underrepresentation of experiments with high standard errors and small effect sizes suggests a small publication bias.

We interpret the meta-regression-based estimate of the gender difference ($b = 0.49$, where $b = \Delta g$) as an upper bound of the true effect size since this estimate is based on experiments that reported only separate effect sizes for males and females. The gender effect of these experiments can be considered as positively biased when assuming that authors who observe a statistically significant gender difference would be more likely to report separate effect sizes for males and females. In contrast, the meta-analytic estimate derived from the observed gender differences within each experiment ($g = 0.14$) can be viewed as a lower bound of the true effect size. This view is plausible because the unknown effect sizes of experiments without significant gender differences were set to zero although non-zero differences were likely also observed but just not reported because of missing power to render these effects statistically significant.

To facilitate the interpretability of our models, the effect size of the looking time difference between novel and familiar rotation events was coded in a linear fashion. Accordingly, a positive looking time difference (i.e., a novelty preference) was taken as stronger evidence for mental rotation than a looking time difference of zero (i.e., no systematic preference). A zero difference in turn was taken as stronger evidence for mental rotation than a negative looking time difference (i.e., a familiarity preference). This approach can be corroborated by the established view that novelty preference is generally considered as the paradigmatic mental rotation behavior (for related discussion, see[36–38]). Nevertheless, some authors have argued that familiarity preference should also be taken as evidence for mental rotation ability[9,10,28]. Interpreting novelty preference *and* familiarity preference as successful mental rotation would have increased the effect sizes reported here.

The gender differences observed here remain to be explained by interacting genetic and environmental factors that are largely unknown. To the best of our knowledge, there are currently no genetic association or gene-environment interaction studies with a focus on mental rotation. Nevertheless, it is documented that genetic contributions to behavioral variance in mental rotation are substantially smaller than unique non-shared environmental contributions both in male and female adults[39]. Whether this observation also applies to infants remains to be explored.

One recent study on 5–6-month-old female infants provided preliminary evidence for possible social-environmental effects related to parental attitudes towards gender which might partly explain the results of our present work[27]. As far as we know, potentially mediating and moderating factors that could already be operational in infancy, however, are not yet empirically established. In a similar vein, while mental rotation training has small-to-medium post-test effects in children, it is unclear whether it can remove gender differences and be adapted to infants[40].

Sex hormone concentration in male infants, especially postnatal testosterone in the first six months of life, could also contribute to gender differences in mental rotation performance[27,41,42]. However, possible biological developmental pathways, bridging the gap from hormonal to behavioral differences, are currently far from understood.

A number of additional factors have been associated with individual differences in infant mental rotation performance. For example, mental rotation is related to previous relevant experience with the particular objects used in the specific task[18,20,30]. This relation also applies to previous experience with manually rotating toys[20]. While these preliminary results require replication, they are in line with the longstanding notion that prior knowledge is the strongest predictor of learning outcomes in a range of cognitive domains (e.g.,[43–45]). Furthermore, there is yet to be confirmed preliminary evidence for possible links between mental rotation performance and several sensory motor skills including fine and gross motor skills, oculomotor control, and crawling skills[19,20].

The present study robustly revealed that male infants look slightly longer at novel rotated objects than female infants. Thus, on average, males show slightly more reliable mental rotation behavior already in the first months of life.

Methods

Protocol

In the present meta-analysis, we followed the established PRISMA 2020 guidelines (Preferred Reporting Items for Systematic reviews and Meta-Analyses)[46]. The PRISMA checklist is provided in Supplementary Table 4 while Supplementary Fig. 2 displays the PRISMA flowchart.

Eligibility criteria

Articles needed to fulfill six criteria for being included in this meta-analysis: (1) The article is written in English or German; (2) The article includes results from a group study with human samples (thus excluding review articles, meta-analyses, case studies, and animal studies); (3) These samples include at least one group of infants (mean age between 0 months and 36 months); (4) Infants were not born preterm and had no clinical diagnosis; (5) Infants performed a mental rotation task; (6) The article contains quantitative scores that can be converted into a standardized mean difference (see Data collection process and items below). We explicitly included works that were not peer reviewed (e.g., dissertations and preprints) to reduce the impact of publication bias.

Information sources and search strategy

We entered the search terms (“mental rotation” OR “mental transformation” OR “spatial rotation” OR “spatial transformation” OR “spatial ability” OR “spatial skills”) AND (“infant” OR “infants” OR “infanthood” OR “toddler” OR “toddlers” OR “toddlerhood” OR “child” OR “children” OR “childhood” OR “month” OR “months”) into four online databases (APA PsycINFO, PubMed/MEDLINE, Scopus, and ProQuest Dissertations & Theses Global). All database queries were completed on December 6, 2021. We configured the databases to check for article titles, abstracts, and keywords while applying no other filters or limits. This yielded 2,616 articles in total, 1,954 of which remained after removing duplicate records (Supplementary Fig. 2). We further identified 76 articles by screening the reference sections of previous reviews and meta-analyses on mental rotation and related skills[4,5,8–10,40,47–49]. Of these, 34 articles had not been covered by the database search. We also identified 94 articles by screening the reference sections of all publications that had been included after the first pass of the selection process. Of these, 49 articles had not been covered by the database search. Accordingly, we screened 2,037 unique articles in total.

Selection process

Two independent raters read the abstract and, if necessary, relevant sections of the full text to check if an article fulfilled the inclusion criteria. Interrater agreement for the binary decision to include versus exclude an article was 98.5% (κ_w [Cohen’s weighted kappa] = 0.67, 95% CI [0.55, 0.78]). Interrater agreement for the specific eligibility criteria was 88.2% ($\kappa_w = 0.72$, 95% CI [0.40, 1.00]). Cases where the two ratings diverged were resolved via discussion among all raters until a consensus was reached. One article[50] was excluded because the authors used a unique mental rotation paradigm that was not comparable to the paradigms used in the other articles. Another article[51] was excluded because the average age of the infants studied (30.7 months) was almost twice as high as the average age of the next article (15.8 months; $Z = 6.86$ compared to all articles). We therefore decided to narrow our analysis from the first 3 years of life to the first 16 months of life. This procedure led to a total of 23 articles being included in the meta-analysis (Supplementary Fig. 2).

Many of these articles consisted of multiple experiments, e.g., using different variations of the mental rotation task or different subsamples of infants. We included all of these experiments in the meta-analysis and accounted for the dependencies between them by means of multilevel modeling with by-article random effects (see Bayesian meta-analysis below). However, we excluded experiments if there was insufficient information to compute a standardized effect size (see Data collection process and items below). Whenever an article reported separate effect sizes for males and females—or other subgroups like crawling and non-crawling infants—but also an effect size combining these groups, we only included the combined effect size. Moreover, we disregarded effect sizes that were clearly based on the same data but reported in different articles. This procedure led to a total of 59 experiments being included in the meta-analysis of mental rotation performance (Supplementary Table 1) and 30 experiments being included in the meta-analysis of gender differences in mental rotation performance.

Data collection process and items

Outcome measures and other relevant variables were extracted from each article by one of three raters and verified by a second rater. For the meta-analysis of mental rotation performance, outcome measures were any summary statistic (Supplementary Table 5) that could be used to determine the standardized mean difference between novel/unexpected rotation events and familiar/expected rotation events (see Introduction). Other extracted variables included, if available, the sample size, the number of males and females, the mean age and its standard deviation, the minimum and maximum age, the type of mental rotation task (habituation or violation of expectation), the modality of stimulus presentation (real objects or objects on a computer screen), and the dimensionality of the stimuli (2D or 3D; Supplementary Table 1).

We also conducted a meta-analysis of the gender differences in mental rotation performance observed within the original articles. For this analysis, the outcome measures were any summary statistic (Supplementary Table 6) that could be used to determine the standardized mean difference between male infants' mental rotation performance and female infants' mental rotation performance. Other extracted variables included the sample size, the mean age and its standard deviation, and the minimum and maximum age of each gender group.

No investigators were contacted for obtaining or confirming additional data and no automated tools were used in the data collection process.

Effect measures

For the meta-analysis of mental rotation performance, one outcome measure per experiment was converted into a standardized mean difference with small sample correction (Hedges' *g*) using the formulas provided in Supplementary Table 5[52–56]. The standard error of Hedges' *g* for each experiment was computed using the formula provided by[52]:

$$SE_{\text{rotation}} = \sqrt{\frac{df}{df-2}\frac{2(1-r)}{n}\left(1+g^2_{\text{rotation}}\frac{n}{2(1-r)}\right)-\frac{g^2_{\text{rotation}}}{J(df)^2}}$$

where n is the sample size of the experiment, df are the degrees of freedom (with $df = 2(n-1)$), r is the correlation between the two dependent measures in the experiment, and $J(df)$ is the correction factor for small samples as described in Supplementary Table 5. The correlation r was not reported in any of the original articles[57]. We therefore always assumed a correlation of $r = 0.50$ to make our analysis comparable to standard (between-group) meta-analyses[54,58,59] and because we were able to infer an average correlation of $r \approx 0.50$ from a subsample of articles which provided sufficient information (Supplementary Methods 1). A sensitivity analysis indicated that changing the assumed correlation to values from $r = -0.90$ via $r = 0.00$ to $r = 0.90$ had no meaningful impact on the meta-analytic effect sizes (Supplementary Tables 7 and 8).

For the meta-analysis of gender differences in mental rotation performance, one outcome measure per contrast between male and female infants was converted into a standardized mean difference with small sample correction (Hedges' g) using the formulas provided in Supplementary Table 6. In some cases where the authors of the original articles did not observe a statistically significant gender difference, they did not report a precise outcome measure. In these cases, we assumed an effect size of $g = 0.00$, thus rendering our meta-analytic effect size of the gender difference more conservative. The standard error of Hedges' g for each gender contrast was computed using the formula provided by[52]:

$$SE_{\text{gender}} = \sqrt{\frac{df}{df-2}\frac{2}{\tilde{n}}\left(1+g^2_{\text{gender}}\frac{\tilde{n}}{2}\right)-\frac{g^2_{\text{gender}}}{J(df)^2}}$$

where df are the degrees of freedom (with $df = n_{\text{female}} + n_{\text{male}} - 2$), $\tilde{n}$ is the harmonic mean of the group sizes (i.e., $\tilde{n} = 2/(n^{-1}_{\text{female}} + n^{-1}_{\text{male}})$), and $J(df)$ is the correction factor for small samples as described in Supplementary Table 6.

Bayesian meta-analysis

We synthesized the effect sizes and their sampling variances using a Bayesian multilevel model. This model had three levels, with infant participants nested in experiments and experiments nested in articles[60]. We used a weakly-informative $\mathcal{N}(0, 1)$ (normal) prior for the meta-analytic effect size and a weakly-informative $\mathcal{HC}(0, 0.3)$ (half-Cauchy) prior for the two standard deviations (i.e., the random effects of experiments and articles)[61]. A prior sensitivity analysis indicated that making these priors either more informative or less informative did not have a strong influence on the meta-analytic results (Supplementary Tables 7, 8, and 9). For the meta-analysis of mental rotation performance, the dependent variable was the standardized mean difference (Hedges' g) between the novel and familiar rotation condition, weighted by its standard error (see Effect measures above). For the meta-analysis of gender differences, the dependent variable was the standardized mean difference (Hedges' g) between male infants' mental rotation performance and female infants' mental rotation performance, weighted by its standard error. All Bayesian models were fitted using the brms package (Version 2.16.3)[62,63] in R (Version 4.1.2)[64] and the Stan language (Version 2.21.3)[65]. Markov Chain Monte Carlo (MCMC) sampling was used with four parallel chains, each sampling 20,000 draws (including 2,000 warm-up draws) from the posterior distribution. To verify the convergence of the Markov chains, we examined rank plots as well as the $\widehat{R}$ and N_{eff} statistics (Supplementary Fig. 3)[66]. For reporting, the credible interval (CrI) for each model parameter was computed as the 95% equal-tailed interval (ETI) of its posterior distribution, although replacing this with the 95% highest density interval (HDI) yielded highly similar results[67].

Bayesian meta-regression

We examined the influence of three moderator variables on the mental rotation outcomes across experiments, namely (a) the gender of the sample of infants, (b) the age of the sample of infants, and (c) the type of mental rotation task. Gender was coded as a categorical predictor (mixed-gender sample, all-female sample, all-male sample) and contrast-coded using two successive difference contrasts[68] so that we could compare all-female samples versus mixed samples and all-male samples versus all-female samples. Age was coded as a continuous predictor in years and centered by subtracting the average across all experiments. Task type was coded as a categorical predictor (habituation task, violation of expectation task) and contrast-coded using a scaled sum contrast[68]. We then included these predictors for gender, age, and task type as well as two predictors for the interaction between gender and task type (i.e., [female - mixed] × age, [male - female] × age) into a Bayesian meta-regression model. This model was based on the same random effects structure, sampling parameters, and prior specification as described above, but adding a weakly-informative $\mathcal{N}(0, 0.5)$ (normal) prior for all slope parameters.

Frequentist meta-analysis

We verified the results obtained from our Bayesian analyses using classical frequentist meta-analysis and meta-regression. To this end, we used the metafor package (Version 3.0.2)[69] in R to specify the same three-level models as described above but without the Bayesian priors (Supplementary Table 2). These models were fitted using restricted maximum likelihood estimation (REML). To verify that this procedure converged on the most probable estimates, we examined profile likelihood plots[69,70] for the two variance components in the model (i.e., the random effects of experiments and articles; Supplementary Fig. 4).

Publication bias assessment

Publication bias was evaluated based on funnel plots and Egger regression tests[71,72]. Funnel plots visualize the relationship between standard errors and effect sizes. They were created by adapting code from the R package metaviz (Version 0.3.1)[73]. The Egger regression test is a formal statistical test for this relationship between standard errors and effect sizes, probing if the weighted linear regression weight of the effect sizes on the standard errors is significantly different from zero. This test was performed using the metafor package and applying a two-sided false-positive error rate of $\alpha = 0.05$.

To scrutinize the robustness of the meta-analytic effect size against the influence of any individual experiment (which may or may not be a false positive), we conducted a jackknife (leave-one-out) analysis for the meta-analysis of mental rotation performance[74]. To this end, we refitted the Bayesian three-level model repeatedly while leaving out one of the original experiments on every iteration. We then checked if the meta-analytic effect size and heterogeneity remained constant or if it was sensitive against the influence of any individual experiment (Supplementary Table 3).

We also performed trim-and-fill analyses and tested selection models to confirm the robustness of the results obtained from the frequentist meta-analyses against publication bias (Supplementary Methods 2; Supplementary Tables 10 and 11; Supplementary Fig. 5)[75–77].

Data availability

The datasets generated and analyzed during the current study are available in an Open Science Framework repository[78], https://osf.io/k3wdg/?view_only=06a4a0ac3d5f4681baab11751c1498b8. This link is intended for peer review only and will be replaced with a public DOI link upon publication.

Code availability

The analysis code for the current study is available in an Open Science Framework repository[78], https://osf.io/k3wdg/?view_only=06a4a0ac3d5f4681baab11751c1498b8. This link is intended for peer review only and will be replaced with a public DOI link upon publication.

References

1. Shepard, R. N. & Metzler, J. Mental rotation of three-dimensional objects. *Science* **171**, 701–703 (1971).

2. Hegarty, M. & Kozhevnikov, M. Types of visual–spatial representations and mathematical problem solving. *J. Educ. Psychol.* **91**, 684–689 (1999).

3. Johnson, W. & Bouchard Jr., T. J. The structure of human intelligence: it is verbal, perceptual, and image rotation (VPR), not fluid and crystallized. *Intelligence* **33**, 393–416 (2005).

4. Linn, M. C. & Petersen, A. C. Emergence and characterization of sex differences in spatial ability: a meta-analysis. *Child Dev.* **56**, 1479–1498 (1985).

5. Voyer, D., Voyer, S. & Bryden, M. P. Magnitude of sex differences in spatial abilities: a meta-analysis and consideration of critical variables. *Psychol. Bull.* **117**, 250–270 (1995).

6. Voyer, D. Time limits and gender differences on paper-and-pencil tests of mental rotation: a meta-analysis. *Psychon. Bull. Rev.* **18**, 267–277 (2011).

7. Maeda, Y. & Yoon, S. Y. A meta-analysis on gender differences in mental rotation ability measured by the Purdue Spatial Visualization Tests: Visualization of Rotations (PSVT:R). *Educ. Psychol. Rev.* **25**, 69–94 (2013).

8. Lauer, J. E., Yhang, E. & Lourenco, S. F. The development of gender differences in spatial reasoning: a meta-analytic review. *Psychol. Bull.* **145**, 537–565 (2019).

9. Moore, D. S. & Johnson, S. P. The development of mental rotation ability across the first year after birth. *Adv. Child Dev. Behav.* **58**, 1–33 (2020).

10. Johnson, S. P. & Moore, D. S. Spatial thinking in infancy: origins and development of mental rotation between 3 and 10 months of age. *Cogn. Res. Princ. Implic.* **5**, 10 (2020).

11. Kellman, P. J. Perception of three-dimensional form by human infants. *Percept. Psychophys.* **36**, 353–358 (1984).

12. Rochat, P. & Hespos, S. J. Tracking and anticipation of invisible spatial transformations by 4- to 8-month-old infants. *Cogn. Dev.* **11**, 3–17 (1996).

13. Hespos, S. J. & Rochat, P. Dynamic mental representation in infancy. *Cognition* **64**, 153–188 (1997).

14. Moore, D. S. & Johnson, S. P. Mental rotation in human infants: a sex difference. *Psychol. Sci.* **19**, 1063–1066 (2008).

15. Quinn, P. C. & Liben, L. S. A sex difference in mental rotation in young infants. *Psychol. Sci.* **19**, 1067–1070 (2008).

16. Moore, D. S. & Johnson, S. P. Mental rotation of dynamic, three-dimensional stimuli by 3-month-old infants. *Infancy* **16**, 435–445 (2011).

17. Frick, A. & Möhring, W. Mental object rotation and motor development in 8- and 10-month-old infants. *J. Exp. Child Psychol.* **115**, 708–720 (2013).

18. Möhring, W. & Frick, A. Touching up mental rotation: effects of manual experience on 6-month-old infants' mental object rotation. *Child Dev.* **84**, 1554–1565 (2013).

19. Schwarzer, G., Freitag, C., Buckel, R. & Lofruthe, A. Crawling is associated with mental rotation ability by 9-month-old infants. *Infancy* **18**, 432–441 (2013).

20. Schwarzer, G., Freitag, C. & Schum, N. How crawling and manual object exploration are related to the mental rotation abilities of 9-month-old infants. *Front. Psychol.* **4**, 97 (2013).

21. Frick, A. & Wang, S.-H. Mental spatial transformations in 14- and 16-month-old infants: effects of action and observational experience. *Child Dev.* **85**, 278–293 (2014).

22. Quinn, P. C. & Liben, L. S. A sex difference in mental rotation in infants: convergent evidence. *Infancy* **19**, 103–116 (2014).

23. Lauer, J. E., Udelson, H. B., Jeon, S. O. & Lourenco, S. F. An early sex difference in the relation between mental rotation and object preference. *Front. Psychol.* **6**, 558 (2015).

24. Antrilli, N. K. & Wang, S. Visual cues generated during action facilitate 14-month-old infants' mental rotation. *J. Cogn. Dev.* **17**, 418–429 (2016).

25. Christodoulou, J., Johnson, S. P., Moore, D. M. & Moore, D. S. Seeing double: 5-month-olds' mental rotation of dynamic, 3D block stimuli presented on dual monitors. *Infant Behav. Dev.* **45**, 64–70 (2016).

26. Lauer, J. E. & Lourenco, S. F. Spatial processing in infancy predicts both spatial and mathematical aptitude in childhood. *Psychol. Sci.* **27**, 1291–1298 (2016).

27. Constantinescu, M., Moore, D. S., Johnson, S. P. & Hines, M. Early contributions to infants' mental rotation abilities. *Dev. Sci.* **21**, e12613 (2018).

28. Erdmann, K., Kavšek, M. & Heil, M. Infants' looking times in a dynamic mental rotation task: clarifying inconsistent results. *Cogn. Dev.* **48**, 279–285 (2018).

29. Gerhard, T. M. & Schwarzer, G. Impact of rotation angle on crawling and non-crawling 9-month-old infants' mental rotation ability. *J. Exp. Child Psychol.* **170**, 45–56 (2018).

30. Slone, L. K., Moore, D. S. & Johnson, S. P. Object exploration facilitates 4-month-olds' mental rotation performance. *PLoS One* **13**, e0200468 (2018).

31. Kaaz, T. & Heil, M. Infants' looking times in a 2-D mental rotation task. *Infant Child Dev.* **29**, (2020).

32. Gerhard-Samunda, T. M., Jovanovic, B. & Schwarzer, G. Role of manually-generated visual cues in crawling and non-crawling 9-month-old infants' mental rotation. *Cogn. Dev.* **59**, (2021).

33. Kelch, A., Schwarzer, G., Gehb, G. & Jovanovic, B. How 9-month-old crawling infants profit from visual-manual rotations in a mental rotation task. *Infant Behav. Dev.* **65**, 101642 (2021).

34. Fantz, R. L. Visual experience in infants: decreased attention to familiar patterns relative to novel ones. *Science* **146**, 668–670 (1964).

35. Baillargeon, R., Spelke, E. S. & Wasserman, S. Object permanence in five-month-old infants. *Cognition* **20**, 191–208 (1985).

36. Houston-Price, C. & Nakai, S. Distinguishing novelty and familiarity effects in infant preference procedures. *Infant Child Dev.* **13**, 341–348 (2004).

37. Black, A. & Bergmann, C. Quantifying infants' statistical word segmentation: a meta-analysis. in *Proceedings of the 39th Annual Meeting of the Cognitive Science Society* 124–129 (2017).

38. Cristia, A. Can infants learn phonology in the lab? A meta-analytic answer. *Cognition* **170**, 312–327 (2018).

39. Shakeshaft, N. G. *et al.* Rotation is visualisation, 3D is 2D: using a novel measure to investigate the genetics of spatial ability. *Sci. Rep.* **6**, 30545 (2016).

40. Uttal, D. H. *et al.* The malleability of spatial skills: a meta-analysis of training studies. *Psychol. Bull.* **139**, 352–402 (2013).

41. Toivainen, T. *et al.* Prenatal testosterone does not explain sex differences in spatial ability. *Sci. Rep.* **8**, 13653 (2018).

42. Erdmann, K. *et al.* Sex specific relationships between infants' mental rotation ability and amiotic sex hormones. *Neurosci. Lett.* **707**, 134298 (2019).

43. Ausubel, D. P. *Educational Psychology: A Cognitive View.* (Holt, Rinehart and Winston, 1968).

44. Bradley, L. & Bryant, P. E. Categorizing sounds and learning to read—a causal connection. *Nature* **301**, 419–421 (1983).

45. Halberda, J., Mazzocco, M. M. M. & Feigenson, L. Individual differences in non-verbal number acuity correlate with maths achievement. *Nature* **455**, 665–668 (2008).

46. Page, M. J. *et al.* The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ Brit. Med. J.* **372**, n71 (2021).

47. Frick, A., Möhring, W. & Newcombe, N. S. Development of mental transformation abilities. *Trends. Cogn. Sci.* **18**, 536–542 (2014).

48. Kubicek, C. & Schwarzer, G. On the relation between infants' spatial object processing and their motor skills. *J. Mot. Learn. Dev.* **6**, S6–S23 (2018).

49. Yang, W., Liu, H., Chen, N., Xu, P. & Lin, X. Is early spatial skills training effective? A meta-analysis. *Front. Psychol.* **11**, 1938 (2020).

50. Mash, C., Arterberry, M. E. & Bornstein, M. H. Mechanisms of visual object recognition in infancy: five-month-olds generalize beyond the interpolation of familiar views. *Infancy* **12**, 31–43 (2007).

51. Pedrett, S., Kaspar, L. & Frick, A. Understanding of object rotation between two and three years of age. *Dev. Psychol.* **56**, 261–274 (2020).

52. Goulet-Pelletier, J.-C. & Cousineau, D. A review of effect sizes and their confidence intervals, part I: the Cohen's *d* family. *Quant. Method. Psychol.* **14**, 242–265 (2018).

53. Hedges, L. V. & Olkin, I. *Statistical Methods for Meta-Analysis.* (Academic Press, 1985).

54. Lakens, D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for *t*-tests and ANOVAs. *Front. Psychol.* **4**, (2013).

55. Rosenthal, R. *Meta-Analytic Procedures for Social Research.* (SAGE Publications, 1991).

56. Cumming, G. *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis.* (Routledge, 2012).

57. Harrer, M., Cuijpers, P., Furukawa, T. A. & Ebert, D. D. *Doing Meta-Analysis with R: A Hands-on Guide.* (Chapman & Hall/CRC Press, 2021).

58. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences.* (Routledge, 1988).

59. Morris, S. B. & DeShon, R. P. Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychol. Methods* **7**, 105–125 (2002).

60. Van den Noortgate, W., López-López, J. A., Marín-Martínez, F. & Sánchez-Meca, J. Three-level meta-analysis of dependent effect sizes. *Behav. Res.* **45**, 576–594 (2013).

61. Williams, D. R., Rast, P. & Bürkner, P.-C. Bayesian meta-analysis with weakly informative prior distributions. Preprint at https://doi.org/10.31234/osf.io/7tbrm (2018).

62. Bürkner, P.-C. brms: an R package for bayesian multilevel models using Stan. *J. Stat. Soft.* **80**, (2017).

63. Bürkner, P.-C. Advanced Bayesian multilevel modeling with the R package brms. *The R Journal* **10**, 395 (2018).

64. R Core Team. *R: A Language and Environment for Statistical Computing.* (2021).

65. Stan Development Team. *Stan Modeling Language Users Guide and Reference Manual.* (2022).

66. Vehtari, A., Gelman, A., Simpson, D., Carpenter, B. & Bürkner, P.-C. Rank-normalization, folding, and localization: an improved $\widehat{R}$ for assessing convergence of MCMC. Preprint at https://doi.org/10.48550/arXiv.1903.08008 (2021).

67. Kruschke, J. K. Bayesian approaches to testing a point ('null') hypothesis. in *Doing Bayesian Data Analysis (Second Edition)* 335–358 (Academic Press, 2015).

68. Schad, D. J., Vasishth, S., Hohenstein, S. & Kliegl, R. How to capitalize on a priori contrasts in linear (mixed) models: a tutorial. *J. Mem. Lang.* **110**, 104038 (2020).

69. Viechtbauer, W. Conducting meta-analyses in R with the metafor package. *J. Stat. Soft.* **36**, (2010).

70. Raue, A. *et al.* Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics* **25**, 1923–1929 (2009).

71. Egger, M., Smith, G. D., Schneider, M. & Minder, C. Bias in meta-analysis detected by a simple, graphical test. *BMJ Brit. Med. J.* **315**, 629–634 (1997).

72. Sterne, J. A. C. & Egger, M. Regression methods to detect publication and other bias in meta-analysis. in *Publication Bias in Meta-Analysis* (eds. Rothstein, H. R., Sutton, A. J. & Borenstein, M.) 99–110 (John Wiley & Sons, 2005).

73. Kossmeier, M., Tran, U. S. & Voracek, M. *metaviz: forrest plots, funnel plots, and visual funnel plot inference for meta-analysis.* Software at https://github.com/Mkossmeier/metaviz (2020).

74. Efron, B. & Tibshirani, R. *An Introduction to the Bootstrap.* (Chapman & Hall, 1993).

75. Duval, S. & Tweedie, R. Trim and fill: a simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics* **56**, 455–463 (2000).

76. Vevea, J. L. & Hedges, L. V. A general linear model for estimating effect size in the presence of publication bias. *Psychometrika* **60**, 419–435 (1995).

77. Vevea, J. L. & Woods, C. M. Publication bias in research synthesis: sensitivity analysis using a priori weight functions. *Psychol. Methods* **10**, 428–443 (2005).

78. Enge, A., Kapoor, S., Kieslinger, A.-S. & Skeide, M. A. *Data and code for "Gender differences in infant mental rotation".* https://osf.io/k3wdg/?view_only=06a4a0ac3d5f4681baab11751c1498b8 (2022).

Author contributions

M.A.S. conceived, designed, and obtained funding for the study. A.E., S.K., and A.-S.K. (in consultation with M.A.S.) designed and executed the search and selection strategies, performed data analysis, conducted robustness checks, and visualized the results. A.E. and M.A.S. wrote the manuscript with additional input from S.K. and A.-S.K.

Competing interests

The authors declare no competing interests.

Materials & Correspondence

Supplementary Information is available for this paper. Correspondence and requests for materials should be addressed to Alexander Enge (enge@cbs.mpg.de).

Supplementary information

Supplementary Methods 1 | Correlation between dependent samples

Meta-analytic methods require an estimate of the sampling variance of each included experiment[57]. For within-participant experimental designs, this sampling variance needs to account for the fact that repeated measures are taken from the same individuals in different conditions. Repeated measures from the same participant will be more similar to one another than measures from two randomly selected participants and thus do not provide independent information. This is accounted for by the correlation term r in the formula of the standard error of the effect size SE_{rotation} (see Methods)[52]. However, the correlation between repeated measures tends to go unreported in research articles. In fact, none of the 23 articles in the present meta-analysis provided a direct numerical estimate for the correlation between infants' looking times in the novel and familiar rotation object conditions.

However, there were 15 experiments taken from seven different articles that included enough statistical information to reconstruct this correlation. This was done using the following procedure:

1. Extract an effect size d_{diff} based on a paired t-test, a one-sample t-test of difference scores, or an analysis of variance (ANOVA) (see rows (1) to (4) in Supplementary Table 5). All three of these tests take the correlation between the repeated measures into account.

2. Extract the mean looking time difference m_{diff} (in seconds) between the novel and familiar conditions.

3. Compute the standard deviation (SD) of the mean looking time difference:

$$SD_{\text{diff}} = \frac{m_{\text{diff}}}{d_{\text{diff}}}$$

4. Extract the standard deviation of looking times for the novel condition SD_{novel} and the familiar condition SD_{familiar}.

5. Compute the observed correlation between conditions based on a rearranged equation by[58]:

$$r = \frac{SD_{\text{diff}}^2 - SD_{\text{novel}}^2 - SD_{\text{familiar}}^2}{-2 \cdot SD_{\text{novel}} \cdot SD_{\text{familiar}}}$$

Across these 15 experiments, the average observed correlation was $r = 0.56$ (median $r = 0.64$, range [0.03, 1.00]). However, we decided to assume a correlation of $r = 0.50$ when computing the effect size SE_{rotation} for all 59 experiments. This was to ensure the comparability of our methods to meta-analyses of between-participant studies[54,58,59]. To evaluate the robustness of our Bayesian meta-analytic models against the assumption of a correlation of $r = 0.50$, we ran a sensitivity analysis (Supplementary Tables 7 and 8).

Supplementary Methods 2 | Publication bias correction

We used trim-and-fill analysis and selection models to examine if and how our meta-analytic results would change under different degrees of publication bias. For both of these types of models, we used a simplified two-level version of our frequentist three-level model as a basis. In our first step, we computed a trim-and-fill model by imputing possible effect sizes of non-published experiments[75]. This new model produces a symmetric funnel plot with no association between effect sizes and standard errors (Supplementary Tables 10 and 11; Supplementary Fig. 5).

In our next step, we used selection models to approximate the process by which experiments got selected into the meta-analysis. These models were based on the assumption that experiments yielding significant effects (i.e., small P values) are more likely to get published. Specifically, we ran one selection model assuming a weight function with fixed P value cutoffs[76] at $P = 0.01$, $P = 0.05$, and $P = 0.30$. Additionally, we ran multiple models with a priori selection functions assuming biases of varying severity[77]. These analyses revealed that our results are robust to small to medium publication bias (Supplementary Tables 10 and 11).

Supplementary Table 1 | Experiments included in the main analysis

Article	Experiment	Sample size	Females	Age (M ± SD)[a]	Task	Stimulus type	Stimulus dimensions
Kellman 1984	Kinetic group, habituated to object 1	12	n/a[b]	n/a	Habituation	Digital	3D
	Kinetic group, habituated to object 2	12	n/a	n/a	Habituation	Digital	3D
Rochat & Hespos 1996	Experiment 1	30	11	n/a	VoE[c]	Real	3D
Hespos & Rochat 1997	Experiment 2	21	9	n/a	VoE	Real	3D
	Experiment 3	19	8	n/a	VoE	Real	3D
	Experiment 4, 4-months-old infants	10	n/a	n/a	VoE	Real	3D
	Experiment 4, 6-months-old infants	9	n/a	n/a	VoE	Real	3D
	Experiment 5, 6-months-old infants	10	n/a	n/a	VoE	Real	3D
	Experiment 6, 6-months-old infants	10	n/a	n/a	VoE	Real	3D
Moore & Johnson 2008	Females	20	20	5m 1d ± 10d	Habituation	Digital	3D
	Males	20	0	5m 1d ± 10d	Habituation	Digital	3D
Quinn & Liben 2008	Females	12	12	3m 20d ± 10d	Habituation	Real	2D
	Males	12	0	3m 15d ± 12d	Habituation	Real	2D
Moore & Johnson 2011	Females	20	20	3m 5d ± 12d	Habituation	Digital	3D
	Males	20	0	3m 5d ± 12d	Habituation	Digital	3D
Frick & Möhring 2013	10-months-old infants	20	10	10m 21d ± 20d	VoE	Digital	2D
	8-months-old infants	20	10	8m 1d ± 8d	VoE	Digital	2D
Möhring & Frick 2013	Exploration group	20	10	6m 2d ± 9d	VoE	Digital	2D
	Observation group	20	10	5m 26d ± 8d	VoE	Digital	2D
Schwarzer et al. 2013a	Crawling infants	24	11	n/a	Habituation	Digital	3D
	Non-crawling infants	24	11	n/a	Habituation	Digital	3D
Schwarzer et al. 2013b	Crawling infants	24	12	n/a	Habituation	Digital	3D
	Non-crawling infants	24	10	n/a	Habituation	Digital	3D
Frick & Wang 2014	Experiment 1, 14-months-old infants	14	6	n/a	VoE	Real	3D
	Experiment 1, 16-months-old infants	14	6	n/a	VoE	Real	3D
	Experiment 2, 14-months-old infants	14	4	n/a	VoE	Real	3D
	Experiment 2, 16-months-old infants	14	7	n/a	VoE	Real	3D
	Experiment 3, other-turning group	14	6	n/a	VoE	Real	3D
	Experiment 3, self-turning group	14	6	n/a	VoE	Real	3D
Quinn & Liben 2014	Experiment 2, 6-months-old females	12	12	6m 11d ± 17d	Habituation	Real	2D
	Experiment 2, 6-months-old males	12	0	6m 3d ± 13d	Habituation	Real	2D
	Experiment 2, 9-months-old females	12	12	9m 6d ± 13d	Habituation	Real	2D
	Experiment 2, 9-months-old males	12	0	9m 3d ± 9d	Habituation	Real	2D
Lauer et al. 2015	Full sample	56	28	10m 3d ± 55d	Habituation	Digital	2D
Antrilli & Wang 2016	Horizontal-stripes group	16	8	n/a	VoE	Real	3D
	Vertical-stripes group	16	9	n/a	VoE	Real	3D
Christodoulou et al. 2016	Full sample	48	24	5m 0d ± 7d	Habituation	Digital	3D
Lauer & Lourenco 2016	Full sample	53	25	10m 13d ± 54d	Habituation	Digital	2D
Constantinescu et al. 2018	Females	28	28	5m 14d ± 10d	Habituation	Digital	3D
	Males	26	0	5m 14d ± 10d	Habituation	Digital	3D
Erdmann et al. 2018	5-months-old infants	208	104	5m 13d ± 9d	Habituation	Digital	3D
	9-months-old infants	168	84	9m 11d ± 11d	Habituation	Digital	3D
Gerhard & Schwarzer 2018	Crawling infants, 0° group	19	9	9m 13d ± 8d	Habituation	Digital	3D
	Crawling infants, 54° group	20	9	9m 13d ± 8d	Habituation	Digital	3D
	Non-crawling infants, 0° group	18	8	9m 13d ± 8d	Habituation	Digital	3D
	Non-crawling infants, 54° group	19	9	9m 13d ± 8d	Habituation	Digital	3D
Slone et al. 2018	Full sample	80	40	4m 16d ± 10d	Habituation	Digital	3D
Kaaz & Heil 2020	Experiment 1, females	144	144	6m 9d ± 7d	Habituation	Digital	2D
	Experiment 1, males	144	0	6m 10d ± 8d	Habituation	Digital	2D
	Experiment 2, females	48	48	6m 14d ± 8d	Habituation	Digital	2D
	Experiment 2, males	48	0	6m 12d ± 8d	Habituation	Digital	2D

Supplementary Table 1 continued

Article	Experiment	Sample size	Females	Age ($M \pm SD$)[a]	Task	Stimulus type	Stimulus dimensions
Gerhard-Samunda et al. 2021	Horizontal-stripes group, crawling infants	11	3	9m 5d ± 9d	Habituation	Digital	3D
	Horizontal-stripes group, non-crawling infants	10	5	9m 3d ± 8d	Habituation	Digital	3D
	Vertical-stripes group, crawling infants	11	2	9m 4d ± 9d	Habituation	Digital	3D
	Vertical-stripes group, non-crawling infants	11	6	9m 3d ± 6d	Habituation	Digital	3D
Kelch et al. 2021	Experiment 1, crawling infants	11	10	9m 17d ± 8d	Habituation	Real	3D
	Experiment 1, non-crawling infants	13	5	9m 17d ± 8d	Habituation	Real	3D
	Experiment 2, crawling infants	13	7	9m 18d ± 8d	Habituation	Real	3D
	Experiment 2, non-crawling infants	14	6	9m 18d ± 8d	Habituation	Real	3D

[a] = mean ± standard deviation, [b] = not available, [c] = violation of expectation.

Supplementary Table 2 | Frequentist meta-analyses and meta-regression

Model	Parameter	Estimate	SE^a	Z	P	95% CI^b
Meta-analysis of mental rotation	Hedges' g	0.26	0.07	3.65	< 0.001	[0.12, 0.40]
	$\sigma^2_{\text{article}}$	0.02				
	$\sigma^2_{\text{experiment}}$	0.17				
Meta-regression of mental rotation	Intercept (Hedges' g)	0.38	0.14	2.76	0.006	[0.11, 0.66]
	Female - mixed	-0.26	0.26	-1.02	0.307	[-0.77, 0.24]
	Male - female	0.78	0.32	2.45	0.014	[0.16, 1.41]
	Age (per year)	0.15	0.62	0.23	0.815	[-1.07, 1.36]
	Habituation - VoE	0.29	0.17	1.76	0.079	[-0.03, 0.62]
	(Female - mixed) × age	-0.42	1.26	-0.33	0.738	[-2.88, 2.04]
	(Male - female) × age	1.40	1.71	0.82	0.411	[-1.94, 4.75]
	$\sigma^2_{\text{article}}$	0.02				
	$\sigma^2_{\text{experiment}}$	0.16				
Meta-analysis of gender differences	Hedges' g	0.15	0.08	1.84	0.066	[-0.01, 0.30]
	$\sigma^2_{\text{article}}$	0.05				
	$\sigma^2_{\text{experiment}}$	0.00				

^{a} = standard error, ^{b} = 95% confidence interval.

Supplementary Table 3 | Jackknife (leave-one-out) analysis

Article	Left-out experiment	Hedges' g	$\sigma^2_{\text{article}}$	$\sigma^2_{\text{experiment}}$	ICC^{a}
Kellman 1984	Kinetic group, habituated to object 1	0.25 [0.11, 0.40]	0.03 [0.00, 0.10]	0.17 [0.09, 0.29]	0.12 [0.00, 0.44]
	Kinetic group, habituated to object 2	0.25 [0.11, 0.40]	0.02 [0.00, 0.10]	0.17 [0.09, 0.29]	0.12 [0.00, 0.43]
Rochat & Hespos 1996	Experiment 1	0.25 [0.11, 0.40]	0.03 [0.00, 0.12]	0.17 [0.09, 0.30]	0.14 [0.00, 0.47]
Hespos & Rochat 1997	Experiment 2	0.25 [0.11, 0.39]	0.02 [0.00, 0.10]	0.17 [0.09, 0.29]	0.11 [0.00, 0.44]
	Experiment 3	0.25 [0.11, 0.40]	0.02 [0.00, 0.10]	0.18 [0.09, 0.30]	0.12 [0.00, 0.44]
	Experiment 4, 4-months-old infants	0.26 [0.11, 0.41]	0.03 [0.00, 0.13]	0.17 [0.09, 0.30]	0.15 [0.00, 0.52]
	Experiment 4, 6-months-old infants	0.25 [0.11, 0.40]	0.03 [0.00, 0.12]	0.18 [0.09, 0.30]	0.12 [0.00, 0.46]
	Experiment 5, 6-months-old infants	0.25 [0.11, 0.39]	0.02 [0.00, 0.10]	0.18 [0.09, 0.30]	0.12 [0.00, 0.44]
	Experiment 6, 6-months-old infants	0.25 [0.11, 0.39]	0.02 [0.00, 0.10]	0.18 [0.09, 0.29]	0.11 [0.00, 0.42]
Moore & Johnson 2008	Females	0.26 [0.12, 0.42]	0.03 [0.00, 0.12]	0.17 [0.09, 0.29]	0.14 [0.00, 0.48]
	Males	0.25 [0.10, 0.40]	0.03 [0.00, 0.12]	0.17 [0.09, 0.29]	0.14 [0.00, 0.49]
Quinn & Liben 2008	Females	0.26 [0.12, 0.42]	0.03 [0.00, 0.14]	0.17 [0.08, 0.29]	0.16 [0.00, 0.53]
	Males	0.24 [0.10, 0.38]	0.03 [0.00, 0.11]	0.15 [0.07, 0.26]	0.15 [0.00, 0.50]
Moore & Johnson 2011	Females	0.26 [0.11, 0.40]	0.03 [0.00, 0.11]	0.18 [0.09, 0.30]	0.14 [0.00, 0.47]
	Males	0.27 [0.13, 0.41]	0.02 [0.00, 0.10]	0.16 [0.08, 0.28]	0.13 [0.00, 0.45]
Frick & Möhring 2013	10-months-old infants	0.24 [0.10, 0.39]	0.03 [0.00, 0.11]	0.16 [0.08, 0.28]	0.14 [0.00, 0.48]
	8-months-old infants	0.26 [0.12, 0.41]	0.03 [0.00, 0.12]	0.17 [0.09, 0.30]	0.14 [0.00, 0.49]
Möhring & Frick 2013	Exploration group	0.25 [0.11, 0.40]	0.03 [0.00, 0.11]	0.18 [0.09, 0.30]	0.13 [0.00, 0.46]
	Observation group	0.26 [0.12, 0.41]	0.03 [0.00, 0.12]	0.17 [0.09, 0.30]	0.14 [0.00, 0.48]
Schwarzer et al. 2013a	Crawling infants	0.25 [0.11, 0.40]	0.03 [0.00, 0.12]	0.17 [0.09, 0.30]	0.14 [0.00, 0.49]
	Non-crawling infants	0.27 [0.12, 0.42]	0.03 [0.00, 0.12]	0.17 [0.08, 0.29]	0.14 [0.00, 0.49]
Schwarzer et al. 2013b	Crawling infants	0.25 [0.10, 0.39]	0.03 [0.00, 0.11]	0.17 [0.09, 0.29]	0.14 [0.00, 0.47]
	Non-crawling infants	0.26 [0.11, 0.41]	0.03 [0.00, 0.13]	0.17 [0.09, 0.30]	0.15 [0.00, 0.50]
Frick & Wang 2014	Experiment 1, 14-months-old infants	0.27 [0.12, 0.42]	0.03 [0.00, 0.13]	0.16 [0.08, 0.28]	0.16 [0.00, 0.53]
	Experiment 1, 16-months-old infants	0.25 [0.11, 0.40]	0.03 [0.00, 0.11]	0.17 [0.09, 0.29]	0.14 [0.00, 0.47]
	Experiment 2, 14-months-old infants	0.26 [0.12, 0.40]	0.03 [0.00, 0.11]	0.18 [0.09, 0.30]	0.13 [0.00, 0.44]
	Experiment 2, 16-months-old infants	0.25 [0.11, 0.40]	0.03 [0.00, 0.10]	0.17 [0.09, 0.29]	0.13 [0.00, 0.45]
	Experiment 3, other-turning group	0.26 [0.12, 0.41]	0.03 [0.00, 0.11]	0.17 [0.09, 0.29]	0.14 [0.00, 0.46]
	Experiment 3, self-turning group	0.25 [0.11, 0.40]	0.03 [0.00, 0.11]	0.17 [0.09, 0.29]	0.13 [0.00, 0.46]
Quinn & Liben 2014	Experiment 2, 6-months-old females	0.27 [0.12, 0.41]	0.03 [0.00, 0.13]	0.17 [0.08, 0.29]	0.16 [0.00, 0.51]
	Experiment 2, 6-months-old males	0.24 [0.10, 0.39]	0.03 [0.00, 0.10]	0.16 [0.08, 0.28]	0.14 [0.00, 0.47]
	Experiment 2, 9-months-old females	0.26 [0.12, 0.42]	0.03 [0.00, 0.12]	0.17 [0.08, 0.29]	0.15 [0.00, 0.51]
	Experiment 2, 9-months-old males	0.25 [0.10, 0.40]	0.03 [0.00, 0.11]	0.17 [0.08, 0.29]	0.13 [0.00, 0.48]
Lauer et al. 2015	Full sample	0.24 [0.10, 0.39]	0.03 [0.00, 0.11]	0.17 [0.08, 0.29]	0.13 [0.00, 0.46]
Antrilli & Wang 2016	Horizontal-stripes group	0.26 [0.11, 0.42]	0.03 [0.00, 0.14]	0.17 [0.08, 0.29]	0.16 [0.00, 0.54]
	Vertical-stripes group	0.24 [0.10, 0.40]	0.03 [0.00, 0.12]	0.16 [0.08, 0.28]	0.15 [0.00, 0.50]
Christodoulou et al. 2016	Full sample	0.27 [0.13, 0.42]	0.03 [0.00, 0.11]	0.17 [0.09, 0.29]	0.13 [0.00, 0.46]
Lauer & Lourenco 2016	Full sample	0.24 [0.10, 0.40]	0.03 [0.00, 0.11]	0.16 [0.08, 0.28]	0.13 [0.00, 0.46]
Constantinescu et al. 2018	Females	0.27 [0.12, 0.42]	0.03 [0.00, 0.13]	0.17 [0.08, 0.29]	0.16 [0.00, 0.53]
	Males	0.24 [0.10, 0.39]	0.03 [0.00, 0.12]	0.16 [0.08, 0.28]	0.15 [0.00, 0.52]
Erdmann et al. 2018	5-months-old infants	0.27 [0.12, 0.41]	0.03 [0.00, 0.10]	0.17 [0.09, 0.30]	0.12 [0.00, 0.44]
	9-months-old infants	0.26 [0.11, 0.41]	0.03 [0.00, 0.11]	0.18 [0.09, 0.30]	0.12 [0.00, 0.44]
Gerhard & Schwarzer 2018	Crawling infants, 0° group	0.25 [0.10, 0.41]	0.03 [0.00, 0.12]	0.17 [0.08, 0.29]	0.15 [0.00, 0.51]
	Crawling infants, 54° group	0.27 [0.13, 0.41]	0.02 [0.00, 0.10]	0.17 [0.09, 0.28]	0.13 [0.00, 0.45]
	Non-crawling infants, 0° group	0.27 [0.12, 0.41]	0.03 [0.00, 0.10]	0.17 [0.09, 0.29]	0.12 [0.00, 0.44]
	Non-crawling infants, 54° group	0.26 [0.11, 0.40]	0.03 [0.00, 0.11]	0.18 [0.09, 0.30]	0.13 [0.00, 0.46]
Slone et al. 2018	Full sample	0.27 [0.12, 0.42]	0.03 [0.00, 0.13]	0.17 [0.09, 0.30]	0.14 [0.00, 0.50]
Kaaz & Heil 2020	Experiment 1, females	0.26 [0.12, 0.41]	0.03 [0.00, 0.11]	0.18 [0.09, 0.30]	0.12 [0.00, 0.44]
	Experiment 1, males	0.26 [0.12, 0.41]	0.03 [0.00, 0.11]	0.18 [0.09, 0.30]	0.12 [0.00, 0.44]
	Experiment 2, females	0.26 [0.12, 0.41]	0.03 [0.00, 0.10]	0.18 [0.09, 0.30]	0.12 [0.00, 0.44]
	Experiment 2, males	0.25 [0.11, 0.40]	0.03 [0.00, 0.11]	0.17 [0.09, 0.30]	0.14 [0.00, 0.48]

Supplementary Table 3 continued

Article	Left-out experiment	Hedges' g	$\sigma^2_{\text{article}}$	$\sigma^2_{\text{experiment}}$	ICC[a]
Gerhard-Samunda et al. 2021	Horizontal-stripes group, crawling infants	0.27 [0.13, 0.41]	0.03 [0.00, 0.11]	0.16 [0.08, 0.28]	0.14 [0.00, 0.48]
	Horizontal-stripes group, non-crawling infants	0.26 [0.11, 0.40]	0.03 [0.00, 0.11]	0.18 [0.09, 0.30]	0.13 [0.00, 0.45]
	Vertical-stripes group, crawling infants	0.25 [0.11, 0.40]	0.03 [0.00, 0.12]	0.17 [0.09, 0.29]	0.15 [0.00, 0.49]
	Vertical-stripes group, non-crawling infants	0.26 [0.11, 0.41]	0.03 [0.00, 0.11]	0.18 [0.09, 0.30]	0.13 [0.00, 0.46]
Kelch et al. 2021	Experiment 1, crawling infants	0.27 [0.13, 0.42]	0.02 [0.00, 0.10]	0.16 [0.09, 0.28]	0.12 [0.00, 0.45]
	Experiment 1, non-crawling infants	0.26 [0.11, 0.42]	0.03 [0.00, 0.12]	0.18 [0.09, 0.30]	0.13 [0.00, 0.47]
	Experiment 2, crawling infants	0.26 [0.12, 0.41]	0.02 [0.00, 0.10]	0.18 [0.09, 0.29]	0.12 [0.00, 0.44]
	Experiment 2, non-crawling infants	0.26 [0.11, 0.41]	0.03 [0.00, 0.11]	0.18 [0.09, 0.30]	0.13 [0.00, 0.47]

[a] = intraclass correlation.

Supplementary Table 4 | PRISMA 2020 checklist

Section	Topic	Item #	Checklist item	Location where item is reported
Title	Title	1	Identify the report as a systematic review.	–
Abstract	Abstract	2	See the PRISMA 2020 for Abstracts checklist.	Abstract
Introduction	Rationale	3	Describe the rationale for the review in the context of existing knowledge.	Introduction
	Objectives	4	Provide an explicit statement of the objective(s) or question(s) the review addresses.	Introduction
Methods	Eligibility criteria	5	Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses.	Methods: Eligibility criteria
	Information sources	6	Specify all databases, registers, websites, organisations, reference lists and other sources searched or consulted to identify studies. Specify the date when each source was last searched or consulted.	Methods: Information sources and search strategy
	Search strategy	7	Present the full search strategies for all databases, registers and websites, including any filters and limits used.	Methods: Information sources and search strategy
	Selection process	8	Specify the methods used to decide whether a study met the inclusion criteria of the review, including how many reviewers screened each record and each report retrieved, whether they worked independently, and if applicable, details of automation tools used in the process.	Methods: Selection process
	Data collection process	9	Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and if applicable, details of automation tools used in the process.	Methods: Data collection process and items
	Data items	10a	List and define all outcomes for which data were sought. Specify whether all results that were compatible with each outcome domain in each study were sought (e.g. for all measures, time points, analyses), and if not, the methods used to decide which results to collect.	Methods: Data collection process and items
		10b	List and define all other variables for which data were sought (e.g. participant and intervention characteristics, funding sources). Describe any assumptions made about any missing or unclear information.	Methods: Data collection process and items
	Study risk of bias assessment	11	Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and if applicable, details of automation tools used in the process.	Methods: Selection process
	Effect measures	12	Specify for each outcome the effect measure(s) (e.g. risk ratio, mean difference) used in the synthesis or presentation of results.	Methods: Effect measures
	Synthesis methods	13a	Describe the processes used to decide which studies were eligible for each synthesis (e.g. tabulating the study intervention characteristics and comparing against the planned groups for each synthesis (item #5)).	Methods: Data collection process and items
		13b	Describe any methods required to prepare the data for presentation or synthesis, such as handling of missing summary statistics, or data conversions.	Methods: Effect measures
		13c	Describe any methods used to tabulate or visually display results of individual studies and syntheses.	Results: Fig. 1
		13d	Describe any methods used to synthesize results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to identify the presence and extent of statistical heterogeneity, and software package(s) used.	Methods: Bayesian meta-analysis; Methods: Frequentist meta-analysis
		13e	Describe any methods used to explore possible causes of heterogeneity among study results (e.g. subgroup analysis, meta-regression).	Methods: Bayesian meta-regression
		13f	Describe any sensitivity analyses conducted to assess robustness of the synthesized results.	Methods: Bayesian meta-analysis; Supplementary Methods 1
	Reporting bias assessment	14	Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting biases).	Methods: Publication bias asssessment

Supplementary Table 4 continued

Section	Topic	Item #	Checklist item	Location where item is reported
	Certainty assessment	15	Describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome.	Methods: Publication bias asssessment
Results	Study selection	16a	Describe the results of the search and selection process, from the number of records identified in the search to the number of studies included in the review, ideally using a flow diagram.	Introduction; Results: Mental rotation performance; Supplementary Fig. 2
		16b	Cite studies that might appear to meet the inclusion criteria, but which were excluded, and explain why they were excluded.	Methods: Selection process
	Study characteristics	17	Cite each included study and present its characteristics.	Supplementary Table 1
	Risk of bias in studies	18	Present assessments of risk of bias for each included study.	–
	Results of individual studies	19	For all outcomes, present, for each study: (a) summary statistics for each group (where appropriate) and (b) an effect estimate and its precision (e.g. confidence/credible interval), ideally using structured tables or plots.	Results: Fig. 1
	Results of syntheses	20a	For each synthesis, briefly summarise the characteristics and risk of bias among contributing studies.	Introduction
		20b	Present results of all statistical syntheses conducted. If meta-analysis was done, present for each the summary estimate and its precision (e.g. confidence/credible interval) and measures of statistical heterogeneity. If comparing groups, describe the direction of the effect.	Results: Mental rotation performance
		20c	Present results of all investigations of possible causes of heterogeneity among study results.	Results: Effects of gender, age, and task
		20d	Present results of all sensitivity analyses conducted to assess the robustness of the synthesized results.	Supplementary Tables 7, 8, and 9
	Reporting biases	21	Present assessments of risk of bias due to missing results (arising from reporting biases) for each synthesis assessed.	Results: Publication bias assessment; Supplementary Methods 2; Supplementary Tables 3, 10 and 11
	Certainty of evidence	22	Present assessments of certainty (or confidence) in the body of evidence for each outcome assessed.	Results: Fig. 1; Results: Fig. 2
Discussion	Discussion	23a	Provide a general interpretation of the results in the context of other evidence.	Discussion
		23b	Discuss any limitations of the evidence included in the review.	Discussion
		23c	Discuss any limitations of the review processes used.	Discussion
		23d	Discuss implications of the results for practice, policy, and future research.	Discussion
Other information	Registration and protocol	24a	Provide registration information for the review, including register name and registration number, or state that the review was not registered.	–
		24b	Indicate where the review protocol can be accessed, or state that a protocol was not prepared.	–
		24c	Describe and explain any amendments to information provided at registration or in the protocol.	–
	Support	25	Describe sources of financial or non-financial support for the review, and the role of the funders or sponsors in the review.	–
	Competing interests	26	Declare any competing interests of review authors.	Competing interests
	Availability of data, code and other materials	27	Report which of the following are publicly available and where they can be found: template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review.	Methods: Data availability; Methods: Code availability

Supplementary Table 5 | Effect sizes for the meta-analysis of mental rotation performance

Available statistics	Conversion to standardized mean difference (Hedges' g)
(1) Cohen's d and sample size[52,53]	$df = 2(n-1)$; $J(df) = \frac{\Gamma(\frac{1}{2}df)}{\sqrt{\frac{df}{2}}\Gamma(\frac{1}{2}(df-1))}$; $g_{\text{rotation}} = d \cdot J(df)$
(2) t statistic and sample size[55]	$d = \frac{t}{\sqrt{n}}$; then (1)
(3) F statistic and sample size[54]	$t = \sqrt{F}$; then (2)
(4) Mean difference between conditions[54]	$d = \frac{m_{\text{diff}}}{SD_{\text{diff}}}$; then (1)
(5) Mean looking times$ per condition[56]	$SD_{\text{av}} = \sqrt{\frac{SD^2_{\text{novel}} + SD^2_{\text{familiar}}}{2}}$; $d = \frac{m_{\text{novel}} - m_{\text{familiar}}}{SD_{\text{av}}}$; then (1)
(6) Mean novelty preference score[54]	$d = \frac{m_{\text{pref}}}{SD_{\text{pref}}}$; then (1)

Supplementary Table 6 | Effect sizes for the meta-analysis of gender differences

Available statistics	Conversion to standardized mean difference (Hedges' g)
(1) Cohen's d and group sizes[52,53]	$df = n_{\text{female}} + n_{\text{male}} - 2$; $J(df) = \frac{\Gamma(\frac{1}{2}df)}{\sqrt{\frac{df}{2}}\Gamma(\frac{1}{2}(df-1))}$; $g_{\text{gender}} = d \cdot J(df)$
(2) t statistic and group sizes[54]	$d = t\sqrt{\frac{1}{n_{\text{female}}} + \frac{1}{n_{\text{male}}}}$; then (1)
(3) F statistic and group sizes[54]	$t = \sqrt{F}$; then (2)
(4) Mean differences between conditions or preference scores[54]	$SD_{\text{pooled}} = \sqrt{\frac{(n_{\text{female}}-1)SD^2_{\text{female}} + (n_{\text{male}}-1)SD^2_{\text{male}}}{df}}$; $d = \frac{m_{\text{female}} - m_{\text{male}}}{SD_{\text{pooled}}}$; then (1)

Supplementary Table 7 | Sensitivity analyses for the meta-analysis of mental rotation performance

Manipulation		Hedges' g	$\sigma^2_{\text{article}}$	$\sigma^2_{\text{experiment}}$	ICC[a]
Assumed correlation	r = -0.90	0.23 [0.09, 0.39]	0.04 [0.00, 0.15]	0.03 [0.00, 0.12]	0.58 [0.00, 1.00]
	r = -0.60	0.24 [0.10, 0.39]	0.05 [0.00, 0.15]	0.04 [0.00, 0.15]	0.52 [0.00, 1.00]
	r = -0.30	0.24 [0.10, 0.40]	0.04 [0.00, 0.14]	0.07 [0.00, 0.18]	0.40 [0.00, 0.99]
	r = 0.00	0.25 [0.11, 0.40]	0.03 [0.00, 0.13]	0.10 [0.02, 0.22]	0.24 [0.00, 0.79]
	r = 0.30	0.25 [0.11, 0.40]	0.03 [0.00, 0.11]	0.14 [0.06, 0.26]	0.16 [0.00, 0.55]
	r = 0.60	0.26 [0.12, 0.41]	0.03 [0.00, 0.11]	0.19 [0.10, 0.31]	0.12 [0.00, 0.43]
	r = 0.90	0.26 [0.12, 0.41]	0.03 [0.00, 0.10]	0.23 [0.14, 0.35]	0.10 [0.00, 0.35]
Prior specification	$b \sim \mathcal{U}(-10, 10)$[b]	0.26 [0.12, 0.41]	0.03 [0.00, 0.11]	0.17 [0.09, 0.29]	0.13 [0.00, 0.47]
	$b \sim \mathcal{N}(0, 0.2)$[c]	0.23 [0.09, 0.36]	0.03 [0.00, 0.11]	0.17 [0.09, 0.29]	0.13 [0.00, 0.46]
	$\sigma \sim \mathcal{U}(0, 10)$[d]	0.26 [0.11, 0.41]	0.03 [0.00, 0.14]	0.18 [0.09, 0.31]	0.15 [0.00, 0.51]
	$\sigma \sim$ Student-$t(10, 0, 0.2)$[e]	0.26 [0.12, 0.40]	0.03 [0.00, 0.10]	0.16 [0.08, 0.26]	0.14 [0.00, 0.48]

[a] = intraclass correlation, [b] = an uninformative (uniform) prior on the mean effect, [c] = an informative (normal) prior on the mean effect, [d] = an uninformative (uniform) prior on the standard deviations, [e] = an informative (Student-t) prior on the standard deviations.

Supplementary Table 8 | Sensitivity analyses for the meta-regression of mental rotation performance

Manipulation		Intercept	Female - mixed	Male - female	Age (per year)	Habituation - VoE	[Female - mixed] × age	[Male - female] × age
Assumed correlation	r = -0.90	0.29 [0.06, 0.54]	-0.11 [-0.48, 0.28]	0.33 [-0.03, 0.71]	0.16 [-0.48, 0.80]	0.22 [-0.14, 0.56]	-0.11 [-0.95, 0.72]	-0.03 [-0.94, 0.88]
	r = -0.60	0.30 [0.07, 0.54]	-0.11 [-0.48, 0.27]	0.35 [-0.01, 0.74]	0.15 [-0.49, 0.79]	0.22 [-0.14, 0.55]	-0.10 [-0.93, 0.72]	-0.01 [-0.92, 0.89]
	r = -0.30	0.32 [0.08, 0.55]	-0.12 [-0.49, 0.26]	0.39 [0.02, 0.78]	0.14 [-0.50, 0.78]	0.22 [-0.14, 0.55]	-0.08 [-0.91, 0.75]	0.01 [-0.89, 0.91]
	r = 0.00	0.33 [0.10, 0.56]	-0.13 [-0.50, 0.25]	0.43 [0.04, 0.83]	0.11 [-0.53, 0.74]	0.23 [-0.12, 0.56]	-0.04 [-0.87, 0.78]	0.03 [-0.88, 0.93]
	r = 0.30	0.34 [0.12, 0.57]	-0.14 [-0.51, 0.24]	0.47 [0.06, 0.87]	0.07 [-0.56, 0.70]	0.25 [-0.09, 0.57]	-0.01 [-0.83, 0.81]	0.05 [-0.85, 0.94]
	r = 0.60	0.36 [0.13, 0.59]	-0.14 [-0.52, 0.24]	0.50 [0.09, 0.91]	0.04 [-0.59, 0.66]	0.26 [-0.09, 0.58]	0.02 [-0.82, 0.85]	0.06 [-0.84, 0.96]
	r = 0.90	0.37 [0.14, 0.59]	-0.14 [-0.52, 0.24]	0.52 [0.09, 0.94]	0.01 [-0.60, 0.63]	0.27 [-0.07, 0.59]	0.05 [-0.77, 0.86]	0.07 [-0.85, 0.97]
Prior specification	$b \sim \mathcal{U}(-10, 10)$[a]	0.36 [0.13, 0.59]	-0.13 [-0.52, 0.26]	0.49 [0.08, 0.90]	0.05 [-0.57, 0.68]	0.26 [-0.08, 0.58]	0.02 [-0.80, 0.83]	0.06 [-0.85, 0.97]
	$b \sim \mathcal{N}(0, 0.2)$[b]	0.27 [0.06, 0.46]	-0.22 [-0.59, 0.15]	0.47 [0.06, 0.88]	0.01 [-0.61, 0.63]	0.19 [-0.15, 0.50]	-0.08 [-0.90, 0.74]	0.02 [-0.89, 0.91]
	$\sigma \sim \mathcal{U}(0, 10)$[c]	0.35 [0.12, 0.59]	-0.14 [-0.53, 0.25]	0.49 [0.07, 0.90]	0.05 [-0.58, 0.70]	0.25 [-0.11, 0.58]	0.00 [-0.83, 0.82]	0.05 [-0.85, 0.96]
	$\sigma \sim$ Student-$t(10, 0, 0.2)$[d]	0.35 [0.13, 0.57]	-0.14 [-0.51, 0.23]	0.49 [0.08, 0.89]	0.05 [-0.56, 0.68]	0.26 [-0.07, 0.57]	0.01 [-0.81, 0.82]	0.07 [-0.83, 0.96]

[a] = an uninformative (uniform) prior on the mean effect, [b] = an informative (normal) prior on the mean effect, [c] = an uninformative (uniform) prior on the standard deviations, [d] = an informative (Student-t) prior on the standard deviations.

Supplementary Table 9 | Sensitivity analyses for the meta-analysis of gender differences

Manipulation		Hedges' *g*	$\sigma^2_{\text{article}}$	$\sigma^2_{\text{experiment}}$	*ICC*[a]
Prior specification	$b \sim \mathcal{U}(-10, 10)$[b]	0.14 [-0.02, 0.31]	0.04 [0.00, 0.17]	0.02 [0.00, 0.10]	0.56 [0.00, 1.00]
	$b \sim \mathcal{N}(0, 0.2)$[c]	0.12 [-0.02, 0.27]	0.04 [0.00, 0.17]	0.02 [0.00, 0.10]	0.56 [0.00, 1.00]
	$\sigma \sim \mathcal{U}(0, 10)$[d]	0.14 [-0.02, 0.32]	0.06 [0.00, 0.23]	0.03 [0.00, 0.12]	0.59 [0.00, 1.00]
	$\sigma \sim$ Student-$t(10, 0, 0.2)$[e]	0.14 [-0.01, 0.29]	0.03 [0.00, 0.13]	0.02 [0.00, 0.09]	0.54 [0.00, 1.00]

[a] = intraclass correlation, [b] = an uninformative (uniform) prior on the mean effect, [c] = an informative (normal) prior on the mean effect, [d] = an uninformative (uniform) prior on the standard deviations, [e] = an informative (Student-*t*) prior on the standard deviations.

Supplementary Table 10 | Publication bias correction for the meta-analysis of mental rotation performance

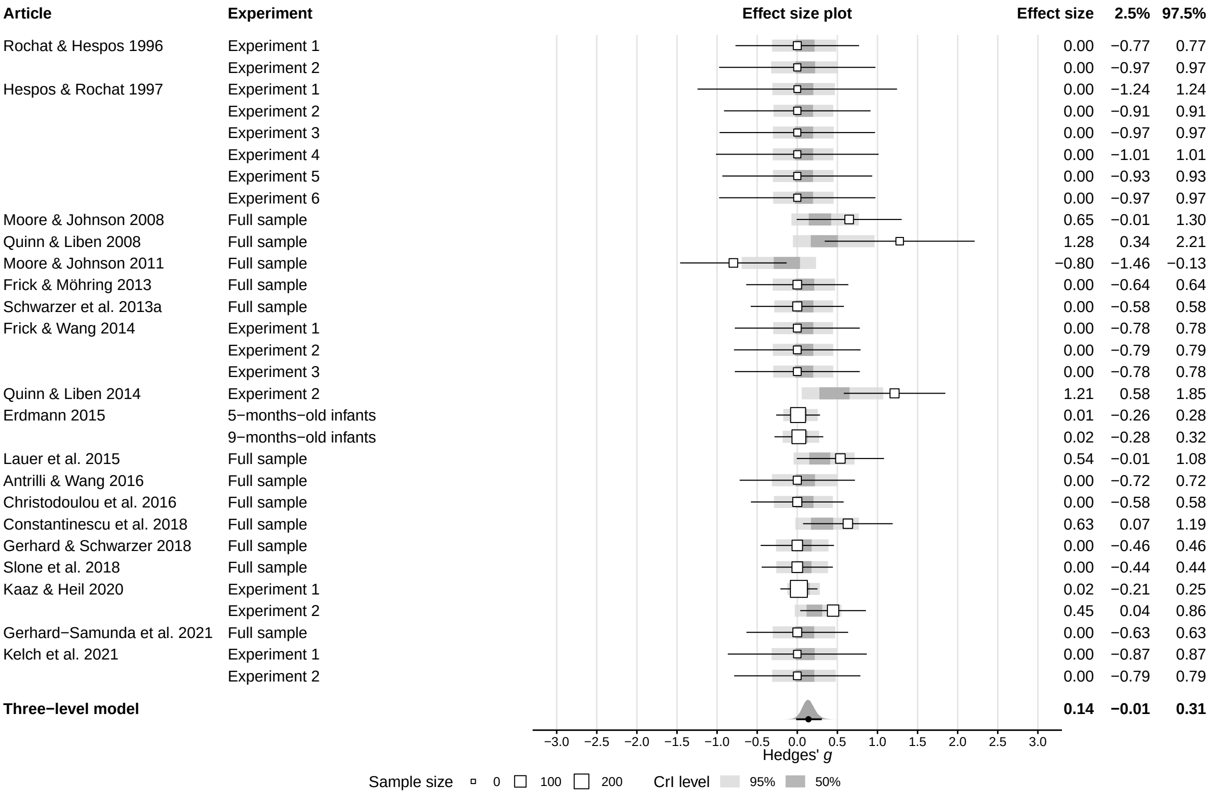
Model	Sub-model	Hedges' *g*	*SE*[a]	*Z*	*P*	95% CI[b]
Two-level model	–	0.25	0.07	3.90	< 0.001	[0.13, 0.38]
Trim-and-fill model	–	0.16	0.07	2.20	0.028	[0.02, 0.30]
Selection models	Weight function, *P* value cutoffs [0.01, 0.05, 0.30]	0.07	0.10	0.67	0.502	[-0.13, 0.26]
	One-tailed, moderate bias	0.13	0.07	1.97	0.049	[0.00, 0.26]
	One-tailed, severe bias	-0.14	0.10	-1.40	0.161	[-0.34, 0.06]
	Two-tailed, moderate bias	0.22	0.06	3.82	< 0.001	[0.11, 0.34]
	Two-tailed, severe bias	0.19	0.05	3.70	< 0.001	[0.09, 0.29]

[a] = standard error, [b] = 95% confidence interval.

Supplementary Table 11 | Publication bias correction for the meta-analysis of gender differences

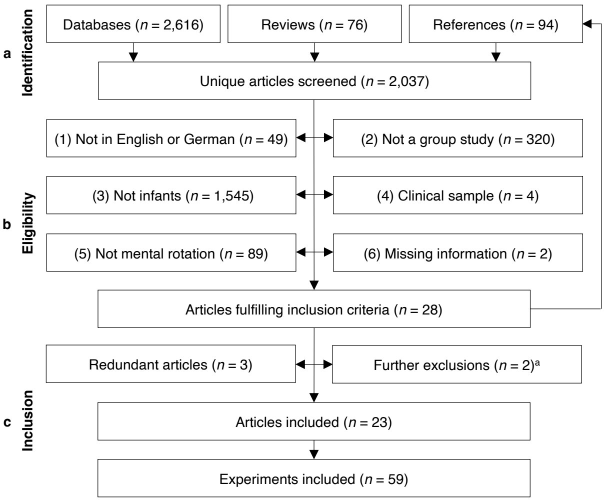
Model	Sub-model	Hedges' *g*	*SE*[a]	*Z*	*P*	95% CI[b]
Two-level model	–	0.13	0.06	2.02	0.044	[0.00, 0.26]
Trim-and-fill model	–	0.13	0.06	2.02	0.044	[0.00, 0.26]
Selection models	Weight function, *P* value cutoffs [0.01, 0.05, 0.30]	0.03	0.06	0.55	0.583	[-0.08, 0.14]
	One-tailed, moderate bias	0.06	0.05	1.24	0.214	[-0.04, 0.16]
	One-tailed, severe bias	-0.08	0.10	-0.74	0.459	[-0.28, 0.13]
	Two-tailed, moderate bias	0.10	0.05	2.02	0.043	[0.00, 0.19]
	Two-tailed, severe bias	0.08	0.04	1.80	0.072	[-0.01, 0.16]

[a] = standard error, [b] = 95% confidence interval.



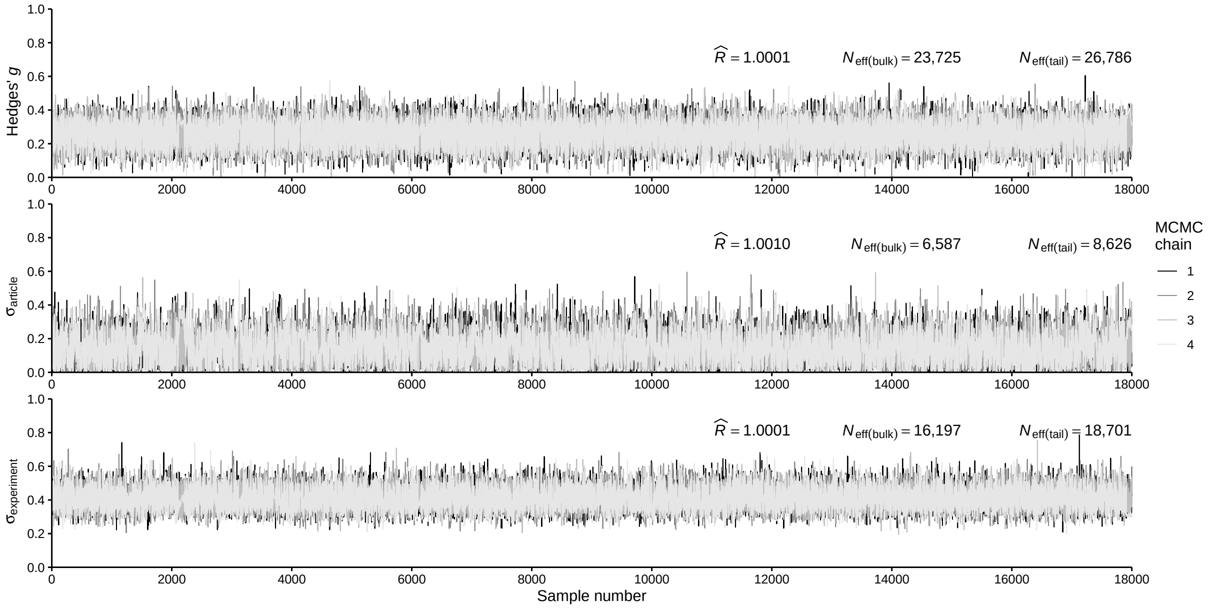
Article	Experiment	Effect size	2.5%	97.5%
Rochat & Hespos 1996	Experiment 1	0.00	−0.77	0.77
	Experiment 2	0.00	−0.97	0.97
Hespos & Rochat 1997	Experiment 1	0.00	−1.24	1.24
	Experiment 2	0.00	−0.91	0.91
	Experiment 3	0.00	−0.97	0.97
	Experiment 4	0.00	−1.01	1.01
	Experiment 5	0.00	−0.93	0.93
	Experiment 6	0.00	−0.97	0.97
Moore & Johnson 2008	Full sample	0.65	−0.01	1.30
Quinn & Liben 2008	Full sample	1.28	0.34	2.21
Moore & Johnson 2011	Full sample	−0.80	−1.46	−0.13
Frick & Möhring 2013	Full sample	0.00	−0.64	0.64
Schwarzer et al. 2013a	Full sample	0.00	−0.58	0.58
Frick & Wang 2014	Experiment 1	0.00	−0.78	0.78
	Experiment 2	0.00	−0.79	0.79
	Experiment 3	0.00	−0.78	0.78
Quinn & Liben 2014	Experiment 2	1.21	0.58	1.85
Erdmann 2015	5−months−old infants	0.01	−0.26	0.28
	9−months−old infants	0.02	−0.28	0.32
Lauer et al. 2015	Full sample	0.54	−0.01	1.08
Antrilli & Wang 2016	Full sample	0.00	−0.72	0.72
Christodoulou et al. 2016	Full sample	0.00	−0.58	0.58
Constantinescu et al. 2018	Full sample	0.63	0.07	1.19
Gerhard & Schwarzer 2018	Full sample	0.00	−0.46	0.46
Slone et al. 2018	Full sample	0.00	−0.44	0.44
Kaaz & Heil 2020	Experiment 1	0.02	−0.21	0.25
	Experiment 2	0.45	0.04	0.86
Gerhard−Samunda et al. 2021	Full sample	0.00	−0.63	0.63
Kelch et al. 2021	Experiment 1	0.00	−0.87	0.87
	Experiment 2	0.00	−0.79	0.79
Three−level model		**0.14**	**−0.01**	**0.31**

Supplementary Fig. 1 | Meta-analysis of gender differences. A Bayesian three-level meta-analysis provided evidence for a small gender difference in mental rotation performance between male and female infants. White squares indicate the effect sizes (Hedges' g) of gender differences in all individual experiments and black lines indicate their 95% confidence intervals. For experiments resulting in a non-significant gender difference and for which the authors did not specify the exact size of this effect, we assumed an effect size of $g = 0.00$. Gray bars depict the 50% and 95% Bayesian credible interval (CrI) based on a Bayesian three-level random-effects model. The last line shows the meta-analytic effect size (black dot) together with its 95% CrI (black line) and its posterior distribution (gray curve).

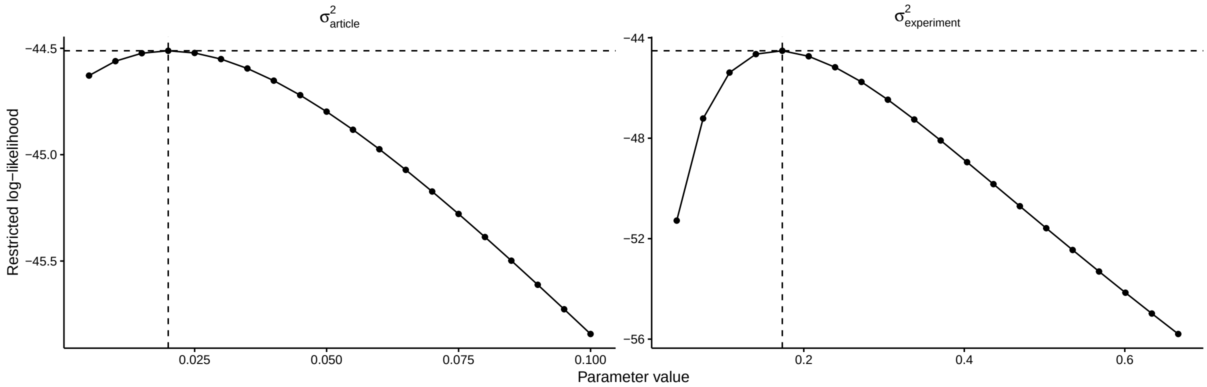


a
Identification
Databases (n = 2,616)
Reviews (n = 76)
References (n = 94)
Unique articles screened (n = 2,037)
b
Eligibility
(1) Not in English or German (n = 49)
(2) Not a group study (n = 320)
(3) Not infants (n = 1,545)
(4) Clinical sample (n = 4)
(5) Not mental rotation (n = 89)
(6) Missing information (n = 2)
Articles fulfilling inclusion criteria (n = 28)
Redundant articles (n = 3)
Further exclusions (n = 2)a
c
Inclusion
Articles included (n = 23)
Experiments included (n = 59)

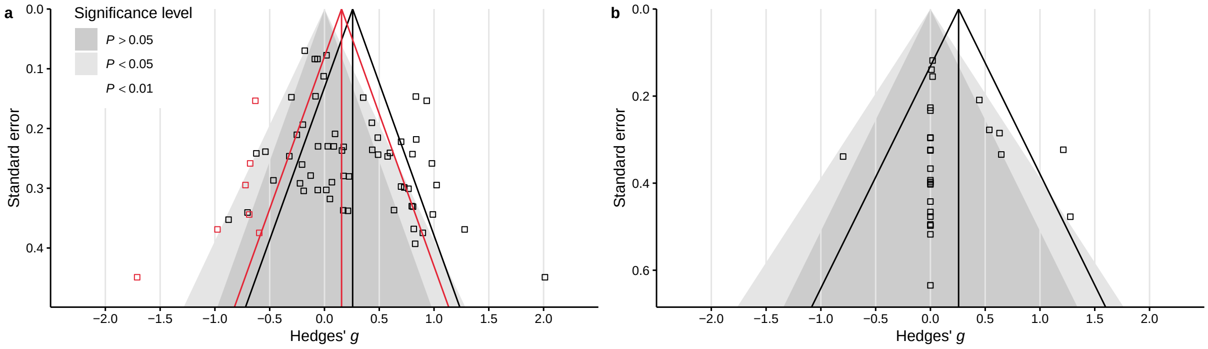
Supplementary Fig. 2 | Literature search and selection process. a, We searched four online databases as well as reviews and reference sections to identify articles in which mental rotation experiments in infants were reported. **b,** Experiments were included in the meta-analysis if they fulfilled six pre-specified inclusion criteria. **c,** Redundant articles comprising the same experiment(s) were excluded. ^{a} = We decided to exclude two additional articles: one article because it was based on an uncommon experimental paradigm, differing substantially from all other articles, and another article because it was based on a sample of infants who were substantially older compared to all other articles.



Supplementary Fig. 3 | Convergence checks for the Bayesian meta-analysis. For each of the three parameters in the three-level meta-analytic model, gray traces show the exploration of the posterior distribution by four independent Markov Chain Monte Carlo (MCMC) chains. Overlap of the gray traces as well as random upward and downward fluctuations (rather than systematic drifts) indicate efficient exploration of the posterior distribution. $\widehat{R}$ = potential scale reduction factor, with values close to 1.00 indicating convergence of the chains, $N_{\text{eff(bulk)}}$ = bulk effective sample size, with larger values indicating better sampling efficiency in the bulk of the distribution (e.g., for estimating the posterior mean), $N_{\text{eff(tail)}}$ = tail effective sample size, with larger values indicating better sampling efficiency in the tails of the distribution (e.g., for estimating the 95% credible interval [CrI]). Details about these convergence criteria can be found in[66].



Supplementary Fig. 4 | Convergence checks for the frequentist meta-analysis. Black curves show the profile of the restricted log likelihood for the two variance components in the frequentist three-level model. The two peaks indicate that restricted maximum likelihood estimation (REML) was able to converge on the most likely parameter estimates[69,70].



Supplementary Fig. 5 | Trim-and-fill analysis. As in Fig. 3 in the main text, black squares indicate the effect sizes (x-axis) and standard errors (y-axis) of the individual experiments included in the meta-analysis of mental rotation performance (**a**) and in the meta-analysis of gender differences within each article (**b**). The funnel contours (diagonal black lines) depict a 95% pseudo-confidence interval around the meta-analytic effect sizes (vertical black lines). Gray shades indicate 95% pseudo-confidence intervals (dark gray) and 99% pseudo-confidence intervals (light gray) under the null hypothesis. Red squares show fictional experiments that were imputed using the trim-and-fill method to compensate for the small publication bias observed in the original funnel plot. Red lines depict the meta-analytic effect size and its 95% pseudo-confidence interval for the trim-and-fill-corrected meta-analysis of mental rotation performance. For the meta-analysis of gender differences, the results of the trim-and-fill analysis suggested that no additional experiments had to be imputed to compensate for publication bias.