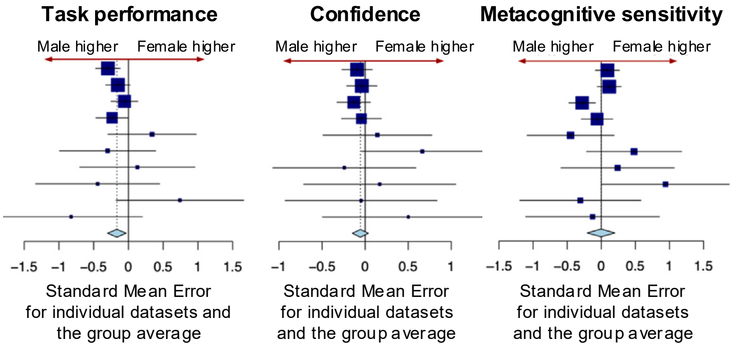
iScience

Article

# No gender difference in confidence or metacognitive ability in perceptual decision-making

## Graphical abstract

**Analysis of 10 Experiments (total N = 1887) Shows No Gender Differences in:**



## Authors

Kai Xue, Yunxuan Zheng,
Christina Papalexandrou,
Kelly Hoogervorst, Micah Allen,
Dobromir Rahnev

## Correspondence

kxue33@gatech.edu

## In brief

Neuroscience; Cognitive neuroscience; Decision science

## Highlights

- Males and females perform similarly in perceptual decision-making tasks
- No gender difference in confidence or metacognition in basic perceptual tasks
- Meta-analysis confirms a lack of gender effects across 10 datasets ($N$ = 1,887)
- Previously reported gender differences in metacognition may be domain-specific

Xue et al., 2024, iScience *27*, 111375
December 20, 2024 © 2024 The Author(s). Published by Elsevier Inc.
https://doi.org/10.1016/j.isci.2024.111375

CellPress

iScience

CellPress
OPEN ACCESS

Article

# No gender difference in confidence or metacognitive ability in perceptual decision-making

Kai Xue,[1,4,*] Yunxuan Zheng,[1] Christina Papalexandrou,[2] Kelly Hoogervorst,[3] Micah Allen,[3] and Dobromir Rahnev[1]
[1]School of Psychology, Georgia Institute of Technology, Atlanta, GA, USA
[2]School of Behavioral and Social Sciences, St. Edward's University, Austin, TX, USA
[3]Institute of Clinical Medicine, Center of Functionally Integrative Neuroscience, Aarhus University, Aarthus, Denmark
[4]Lead contact
*Correspondence: kxue33@gatech.edu

https://doi.org/10.1016/j.isci.2024.111375

## SUMMARY

Prior research has found inconsistent results regarding gender differences in confidence and metacognitive ability. Different studies have shown that men are either more or less confident and have either higher or lower metacognitive abilities than women. However, this research has generally not used well-controlled tasks or used performance-independent measures of metacognitive ability. Here, we test for gender differences in performance, confidence, and metacognitive ability using data from 10 studies from the Confidence Database (total $N$ = 1,887, total number of trials = 633,168). We find an absence of strong gender differences in performance and no gender differences in either confidence or metacognitive ability. These results were further confirmed by meta-analyses of the 10 datasets. These findings show that it is unlikely that gender has a strong effect on metacognitive evaluation in low-level perceptual decision-making and suggest that previously observed gender differences in confidence and metacognition are likely domain-specific.

## INTRODUCTION

Metacognition refers to the ability to evaluate the accuracy of one's own decisions, and it can be measured via confidence ratings. Accurate metacognition can guide learning,[1] cognitive offloading,[2] information seeking,[3] and social interactions.[4,5] Metacognitive ability has also been related to more applied domains such as educational achievement.[6–10] Within the context of education, there have been calls to carefully assess the potential differences in metacognitive skills between males and females to help educators and parents develop more targeted interventions.[7]

Prior research on gender differences in confidence and metacognitive skills has mostly been conducted in applied settings and has resulted in mixed findings. For example, some studies have shown that men are more confident in domains including leadership behavior, math-related tasks, prediction of future outcomes, fraction knowledge, general knowledge, college exams, and reaction time assessment,[11–18] but others show that women are more confident in the domain of literacy skills.[19] Similarly, different studies have found that metacognitive ability—i.e., the degree to which confidence predicts one's performance on a task—is higher in women,[7,20,21] higher in men,[22] or not significantly different between men and women.[8,23] Thus, the literature has been unable to reach a consensus regarding the existence of gender differences in confidence or metacognitive ability. In terms of gender differences in perceptual confidence, inconsistent patterns also emerge. Specifically, Seow and Gillan[18] and Hoogervorst et al.[24] found that males have higher confidence, but Rouault et al.[25] found that neither confidence level nor metacognitive efficiency was associated with gender.

To properly address the question of whether there are differences in metacognition between the genders, it is important to use appropriate measures. One issue with most studies to date is that they either did not use an objective task evaluating subjects' performance[7,17] or did not match subjects' performance when comparing their confidence.[12,16,19–21] Thus, the results of much of the previous literature could be contaminated by first-order performance (i.e., people's ability on the task itself) and first-order bias (i.e., people's propensity to choose one response more frequently than another).[26,27] Therefore, it is critical to measure confidence and metacognitive ability in ways that avoid confounding influences.[26–30]

Here, we tested for the existence of gender difference in confidence and metacognitive ability in 10 datasets featuring perceptual decision-making tasks. All datasets came from the Confidence Database, a large database of open data from experiments that include confidence ratings.[31] The 10 datasets analyzed here are the only ones that included gender information. To anticipate, we found weak and inconsistent effects of gender on performance and no gender differences in confidence judgments or metacognitive ability, with these results holding both for individual datasets and in meta-analyses combining all

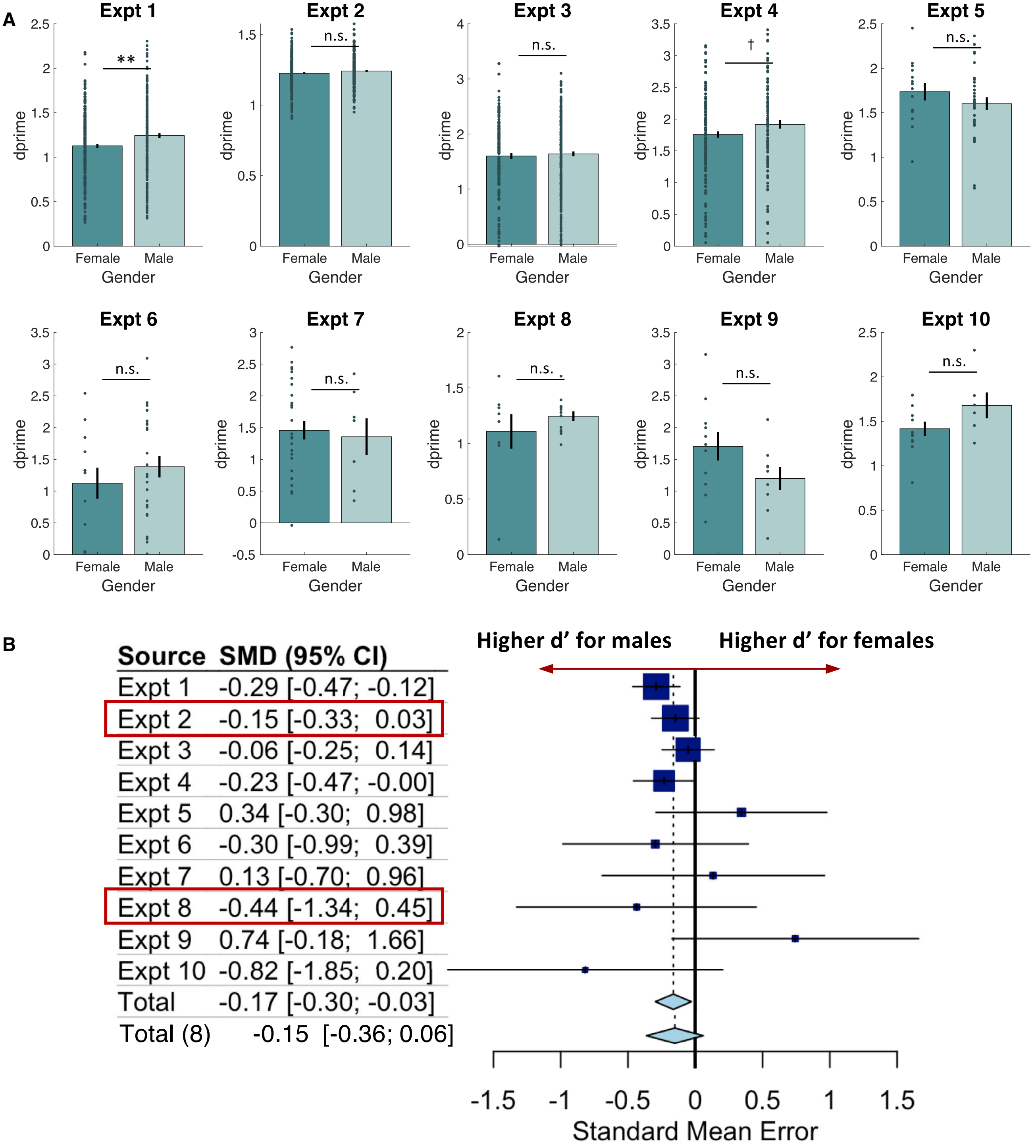


**Figure 1. Males exhibit slightly (but inconsistently) higher task performance (d') than females**

(A) Mean d' values of each gender. Expt 1 shows a significant gender difference even after correction for multiple comparisons, whereas Expt 4 showed a statistically significant gender difference, which, however, disappears after correction for multiple comparisons. Dots represent individual subjects; error bars represent SEM; n.s., not significant; †, $p_{\text{uncorrected}} < 0.05$ but $p_{\text{Bonferroni-corrected}} > 0.05$, **, $p_{\text{Bonferroni-corrected}} < 0.01$.

(B) Meta-analysis results show the standard mean difference (SMD) of d' for each experiment, as well as the meta-analytic average across experiments. The squares on the right correspond to Experiments 1–10, arranged vertically from top to bottom. The x axis displays the standardized mean difference for meta-d' in each experiment. The two experiments circled in red squares are the ones that employed a staircase procedure. Overall, males showed slightly higher d' than females. A negative standard error of the mean (SEM) values indicates that the performance of males is higher than that of females. The size of the squares is

*(legend continued on next page)*

datasets. Overall, these results suggest that gender does not strongly influence performance, confidence, or metacognitive ability in basic perceptual decision-making tasks.

## RESULTS

To investigate the existence of gender differences in confidence or metacognitive ability, we examined all 10 available datasets from the Confidence Database that include gender information for each subject (Table 1). For each dataset, we computed d', confidence, and M-Ratio scores for each subject and compared whether any of these measures is different between the two genders.

### Weak and inconsistent differences in task performance (d') between females and males

In line with previous research,[32] we found that performance (d') was slightly higher for males than females (Figure 1). Independent samples t tests showed no significant difference between d' for female vs. male subjects in 8 of the 10 experiments [Expt2: t(495) = 1.69, $p$ = 0.093, $BF_{01}$ = 2.52; Expt 3: t(419) = 0.57, $p$ = 0.568, $BF_{01}$ = 7.81; Expt 5: t(43) = −1.08, $p$ = 0.288, $BF_{01}$ = 2.03; Expt 6: t(35) = 0.87, $p$ = 0.389, $BF_{01}$ = 2.23; Expt 7: t(33) = −0.31, $p$ = 0.757, $BF_{01}$ = 2.53; Expt 8: t(19) = 1.03, $p$ = 0.318, $BF_{01}$ = 1.74; Expt 9: t(18) = −1.72, $p$ = 0.102, $BF_{01}$ = 0.93; Expt 10: t(16) = 1.73, $p$ = 0.103, $BF_{01}$ = 0.90]. However, both of the remaining two experiments exhibited higher d' for males than females, though the second of these effects could not survive a correction for multiple comparisons [Expt1: t(496) = 3.29, $p$ = 0.001, $BF_{01}$ = 0.06; Expt 4: t(293) = 1.99, $p$ = 0.047, $BF_{01}$ = 1.17]. To increase the power of these analyses, we performed a meta-analysis across these 10 datasets that examined the standard mean difference in d' across the two genders. The heterogeneity assumption was not violated ($x_9^2$ = 12.39, $I^2$ = 27%), suggesting that there is no evidence of significant heterogeneity among all the studies. The meta-analysis showed that males have slightly higher d' than females, but the size of this effect was small (g = −0.17, $p$ = 0.021; Figure 1B). Crucially, when we repeated the meta-analysis excluding the two experiments that employed staircase procedures (Expt 2 and Expt 8 where one would not expect any systematic difference in performance because even true differences in ability should be removed by the staircase), the small gender effect on task performance disappeared entirely (g = −0.15, $p$ = 0.134). Thus, removing the two experiments that employed a staircase procedure turned the originally weak-but-just-significant effect into a slightly weaker and no longer significant effect. These results highlight that the gender effects on performance were small and fragile and thus should be interpreted with caution, as noise in each dataset may push them above or below the threshold for statistical significance. Overall, these results indicate an absence of strong gender differences in basic task performance.

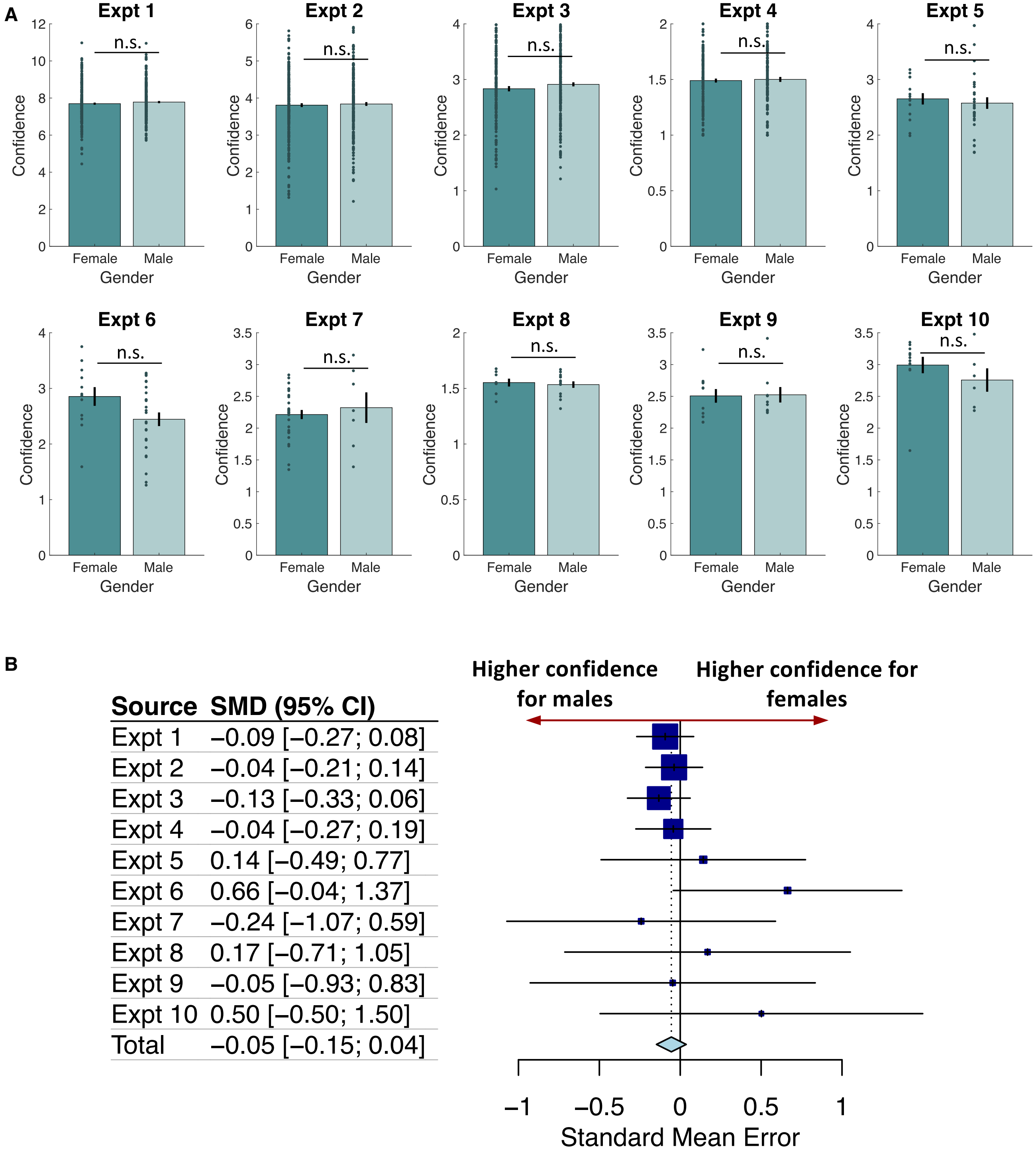
### No gender effect on confidence

Having established the existence of a slight difference in task performance, our primary analysis delved into examining the gender effect in confidence across experiments. We found no effect of gender on confidence (Figure 2A). Specifically, independent samples t tests showed no significant difference in confidence for female vs. male subjects in any of the 10 experiments, with the first four experiments exhibiting Bayes factors above three [Expt 1: t(496) = 1.04, $p$ = 0.300, $BF_{01}$ = 5.94; Expt 2: t(495) = 0.42, $p$ = 0.671, $BF_{01}$ = 9.19; Expt 3: t(419) = 1.34, $p$ = 0.180, $BF_{01}$ = 3.83; Expt 4: t(293) = 0.37, $p$ = 0.715, $BF_{01}$ = 7.23; Expt 5: t(43) = −0.45, $p$ = 0.655, $BF_{01}$ = 2.95; Expt 6: t(35) = −1.93, $p$ = 0.062, $BF_{01}$ = 0.74; Expt 7: t(33) = 0.58, $p$ = 0.563, $BF_{01}$ = 2.32; Expt 8: t(19) = −0.39, $p$ = 0.670, $BF_{01}$ = 2.38; Expt 9: t(18) = 0.11, $p$ = 0.915, $BF_{01}$ = 2.50; Expt 10: t(16) = −1.05, $p$ = 0.308, $BF_{01}$ = 1.62]. To increase the power of these analyses, we again performed a meta-analysis across the 10 datasets that examined the standard mean difference in confidence across the two genders. The heterogeneity assumption was not violated ($x_9^2$ = 0.66, $I^2$ = 0%), suggesting that there is no evidence of significant heterogeneity among all the studies. The results showed that females were slightly less confident than males, but the gender difference in confidence was not significant and the effect size was very small (g = −0.06, $p$ = 0.209; Figure 2B). Note that even this very small effect of higher confidence for males may simply reflect the fact that males also had slightly higher d' values, showing that there is no evidence for males exhibiting overconfidence compared to females in these studies.

To further confirm these results, we performed regression analyses for each dataset where we sought to predict confidence from gender while controlling for accuracy and age. We again found that gender had no significant effect on confidence in any of the 10 datasets (Table S1). In addition, age did not emerge as a significant predictor of confidence levels, except for Expt 3 where the result was only significant without multiple comparison correction (t = −2.36, $p$ = 0.02). This suggests that the observed effects of gender on confidence cannot be attributed to age differences between gender groups. Finally, examination of the age distributions for each dataset revealed only one instance (Expt 2) of a significant age difference between genders (Figure S3).

It is important to note that the interpretation of average confidence should be approached with caution. Although average confidence is often used and reported in the literature, its interpretation can be complex. Theoretically, average confidence might increase with higher accuracy or better metacognitive insight.[30] In our study, we addressed this complexity by first comparing task performance (d') between the two genders, which allowed us to confirm that the performance of both genders was near equivalent. This comparison provided a solid foundation for interpreting the confidence results. Furthermore,

determined by the weight of the effect size: studies with a larger weight (larger sample size) are represented by larger squares. Error bars show 95% confidence intervals (CI). The diamond at the top represents the average effect for all 10 experiments, and the diamond at the bottom represents the average effect after excluding two experiments that employed a staircase procedure. The length of the diamond symbolizes the 95% confidence interval of the pooled effect size. The last line shows the meta-analysis results for d', excluding two experiments that employed a staircase procedure. After the exclusion, the gender effect for d' was no longer significant.



| Source | SMD (95% CI) |
|---|---|
| Expt 1 | −0.09 [−0.27; 0.08] |
| Expt 2 | −0.04 [−0.21; 0.14] |
| Expt 3 | −0.13 [−0.33; 0.06] |
| Expt 4 | −0.04 [−0.27; 0.19] |
| Expt 5 | 0.14 [−0.49; 0.77] |
| Expt 6 | 0.66 [−0.04; 1.37] |
| Expt 7 | −0.24 [−1.07; 0.59] |
| Expt 8 | 0.17 [−0.71; 1.05] |
| Expt 9 | −0.05 [−0.93; 0.83] |
| Expt 10 | 0.50 [−0.50; 1.50] |
| Total | −0.05 [−0.15; 0.04] |

**Figure 2. No gender difference in confidence**

(A) Mean confidence values of each gender. None of the experiments showed a significant gender difference in confidence. Dots represent individual subjects; error bars represent SEM; n.s., not significant.

(B) Meta-analysis results show the standard mean difference (SMD) of confidence for each experiment, as well as the meta-analytic average across experiments. The squares on the right correspond to Experiments 1–10, arranged vertically from top to bottom. The x axis displays the standardized mean difference for meta-d' in each experiment. Overall, there were no significant effects of gender on confidence. A negative standard error of the mean (SEM) values indicates that the confidence of males is higher than that of females.

we also introduced M-Ratio as a measure of metacognitive efficiency, which controls for performance differences and provides a more nuanced understanding of how confidence relates to accuracy.

### No gender effect on metacognitive efficiency or metacognitive sensitivity

Finally, we examined whether there was a gender effect on metacognitive efficiency, M-Ratio. This measure allows us to compare subjects' metacognitive ability controlling for the first-order task performance. Independent samples t tests showed no significant difference between M-Ratio for female vs. male subjects in 8 of the 10 experiments, with three of the experiments having Bayes factors in support of the null hypothesis greater than three [Expt 1: $t(496) = -1.00$, $p = 0.318$, $BF_{01} = 6.18$; Expt 2: $t(495) = -1.30$, $p = 0.194$, $BF_{01} = 4.41$; Expt 4: $t(293) = 0.53$, $p = 0.599$, $BF_{01} = 6.75$; Expt 5: $t(43) = 1.42$, $p = 0.161$, $BF_{01} = 1.44$; Expt 6: $t(35) = -1.40$, $p = 0.171$, $BF_{01} = 1.42$; Expt 7: $t(33) = -0.58$, $p = 0.567$, $BF_{01} = 2.32$; Expt 9: $t(18) = 0.71$, $p = 0.485$, $BF_{01} = 2.10$; Expt 10: $t(16) = -0.29$, $p = 0.792$, $BF_{01} = 2.28$] (Figure 3A). Only 2 of the 10 experiments showed a statistically significant difference in M-Ratio. In one study, males had higher M-Ratio [Expt 3: $t(419) = 2.85$, $p = 0.005$, $BF_{01} = 0.19$], and the results remained statistically significant even after Bonferroni correction. However, the other study showed the opposite pattern such that females had higher M-Ratio [Expt 8: $t(19) = -2.18$, $p = 0.042$, $BF_{01} = 0.52$], though these results were no longer significant after Bonferroni correction.

To address the potential concerns that the empirical false-positive rate is below 5% when significance testing was performed using t tests,[33] we conducted additional analyses using the Mann-Whitney U-test across all 10 experiments. The results of these analyses were consistent with our original findings. Only two experiments showed statistically significant differences in M-Ratio between genders: Experiment 3, where males had higher M-Ratio ($W = 5.65 \text{ X } 10^4$, $p = 4.7 \text{ X } 10^{-5}$; this result remained significant after Bonferroni correction), and Experiment 8, where females had higher M-Ratio ($W = 113.5$, $p = 0.036$; this result was no longer significant after Bonferroni correction). These additional analyses corroborate our initial conclusions and show that the small and inconsistent effects on M-Ratio are not due to an overly low false-positive rate in the statistical test we used.

To increase the power of these analyses, we again performed a meta-analysis across the 10 datasets that examined the standard mean difference in M-Ratio across the two genders. The heterogeneity assumption was not violated ($x_9^2 = 19.41$, $I^2 = 54\%$), suggesting that there is no evidence of significant heterogeneity among all the studies. The results showed that females had slightly lower M-Ratio than males, but the gender difference in M-Ratio was not significant and the effect size was close to zero ($g = -0.004$, $p = 0.97$; Figure 3B). We also repeated these analyses using two alternative measures of metacognition: meta-d' and the difference in confidence between correct and incorrect trials (Figures S1 and S2). In both cases, the results mirrored the effects for M-Ratio, such that there was no overall significant difference in either measure between the two genders.

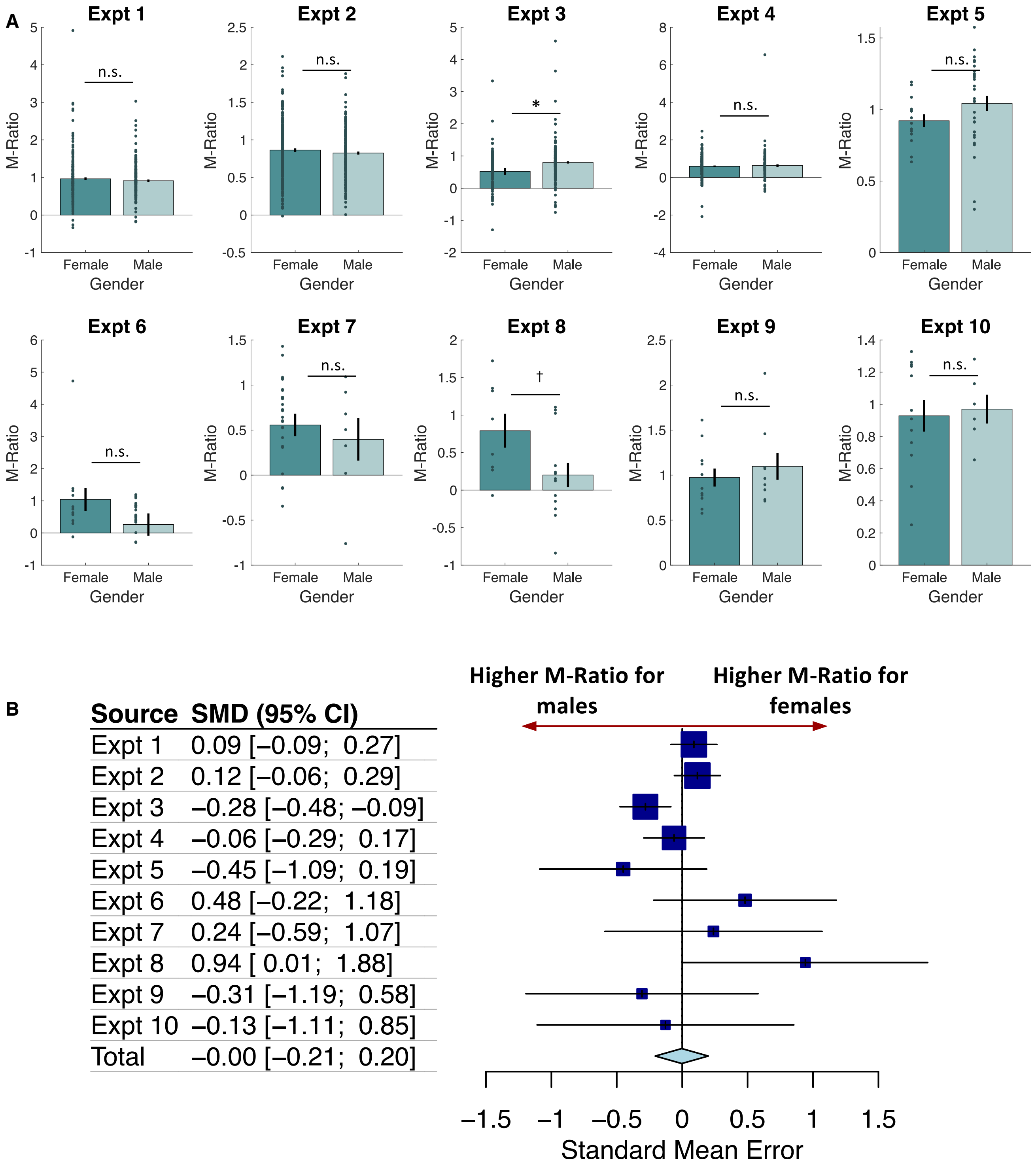
## DISCUSSION

Previous research has found strong gender differences in confidence and metacognition, but the results of different studies have sometimes been in different directions. However, most prior research did not use well-controlled tasks or performance-independent measures of metacognitive ability. Here, we test for gender differences using data from 10 studies from the Confidence Database that featured low-level perceptual decision-making tasks. Our analysis involved a comprehensive examination of task performance, confidence levels, and M-Ratio for each dataset. Additionally, to provide a more holistic perspective, we conducted random-effects meta-analyses across all datasets. Overall, the results failed to reveal consistent gender differences in either confidence or M-Ratio, suggesting that previous examples of metacognitive differences between the genders are likely only domain-specific.

Although gender disparities have been demonstrated in various contexts such as math and literacy skills,[12,16,17,19] research on gender differences in confidence and metacognition has yielded mixed results across different domains. Meta-analyses have shown that males and females can differ in general self-esteem and self-confidence across various domains such as physical appearance, athletics, and self-satisfaction.[34,35] However, these findings are counterbalanced by[36] gender similarities hypothesis that males and females are similar on most psychological variables, with a majority of gender differences being small or negligible.

Our study marks one of the first systematic attempts to explore gender differences in basic perceptual decision-making tasks (but see[24,25,31]). Contrary to findings in some other domains, our results failed to reveal consistent gender differences in either confidence or M-Ratio. These findings suggest that males and females assess the accuracy of their perceptual decisions in a similar manner, providing support for[36] gender similarities hypothesis in the domain of basic perception. On the other hand, the contrast between our findings and those from other domains suggests that prior findings of gender differences in confidence or metacognition may be due to domain-specific mechanisms, such as domain-specific priors or sociocultural factors,[37,38] rather than stemming from a domain-general difference in confidence or metacognitive ability. This underscores the importance of considering domain specificity when studying gender differences in cognitive processes and highlights the need for further research to understand how and why gender differences may emerge in some domains but not others.

Our findings align with the results of the recent study that revealed variations of gender differences in metacognition across domains.[24] Hoogervorst[24] and colleagues examined confidence in four different tasks that involved visual perception, memory, judging countries' GDP, and judging the calorie content of dishes. Males exhibited higher trial-by-trial confidence, with this effect being largest for the GDP task, suggesting possible cultural influences, given that economics is currently a male-dominated field. In contrast with the findings here, males exhibited slightly higher confidence in the vision task. However, this small difference may reflect leakage from the higher self-confidence that males exhibited in the GDP task,[39] given that subjects were introduced to



| Source | SMD (95% CI) |
|---|---|
| Expt 1 | 0.09 [–0.09; 0.27] |
| Expt 2 | 0.12 [–0.06; 0.29] |
| Expt 3 | –0.28 [–0.48; –0.09] |
| Expt 4 | –0.06 [–0.29; 0.17] |
| Expt 5 | –0.45 [–1.09; 0.19] |
| Expt 6 | 0.48 [–0.22; 1.18] |
| Expt 7 | 0.24 [–0.59; 1.07] |
| Expt 8 | 0.94 [ 0.01; 1.88] |
| Expt 9 | –0.31 [–1.19; 0.58] |
| Expt 10 | –0.13 [–1.11; 0.85] |
| Total | –0.00 [–0.21; 0.20] |

**Figure 3. No gender difference in M-Ratio**

(A) Mean M-Ratio values of each gender. Eight of the 10 experiments showed no significant gender difference in M-Ratio. The other two experiments showed the opposite pattern of results, with females having lower M-Ratio in Expt 3 but higher M-Ratio in Expt 8 compared to males. Dots represent individual subjects; error bars represent SEM; n.s., not significant; †, $p_{uncorrected} < 0.05$ but $p_{Bonferroni\text{-}corrected} > 0.05$; *, $p_{Bonferroni\text{-}corrected} < 0.05$.

(B) Meta-analysis results show the standard mean difference (SMD) of M-Ratio for each experiment, as well as the meta-analytic average across experiments. The squares on the right correspond to Experiments 1–10, arranged vertically from top to bottom. The x axis displays the standardized mean difference for meta-d' in each experiment. Overall, there were no significant effects of gender on M-Ratio. A negative standard error of the mean (SEM) values indicates that the M-Ratio of males is higher than that of females.

**Table 1. Dataset information**

| Dataset | N | Number of trials/subject | Confidence scale | Task | Number of difficulty levels |
|---|---|---|---|---|---|
| Experiment 1 | 498 | 210 | 11 | Dot number discrimination | 70 |
| Experiment 2 | 487 | 210 | 6 | Dot number discrimination | Continuous staircase |
| Experiment 3 | 443 | 480 | 4 | Letter or color discrimination | 1 |
| Experiment 4 | 315 | 100 | 2 | Dot number discrimination | 1 |
| Experiment 5 | 45 | 1600 | 4 | Dot number discrimination | 8 |
| Experiment 6 | 37 | 1600 | 4 | Dot number discrimination | 8 |
| Experiment 7 | 37 | 400 | 4 | Tilt direction judgment | 1 |
| Experiment 8 | 22 | 600 | 4 | Figure location detection | Continuous staircase |
| Experiment 9 | 22 | 699 | 4 | Tilt direction judgment | 1 |
| Experiment 10 | 19 | 400 | 4 | Change Detection | 1 |

*Note.* The number of subjects (N), number of trials per subject, confidence scale, task, and number of difficulty levels for each experiment are listed in the table. Order arranged by decreasing the number of subjects.

all tasks in the very beginning of the experiment and rated their self-belief in their performance abilities on all tasks before proceeding to the individual tasks. Importantly, Hoogervorst[24] and colleagues observed a strong global confidence effect: both before and after completing the task, males indicated higher confidence in their abilities to perform all four tasks.

It is likely that confidence and metacognition in the real world are influenced by domain-specific factors and cultural expectations.[40] We argue that perceptual decision-making tasks may serve as a platform for probing intrinsic gender differences in confidence and metacognition. Unlike domains such as mathematics or language, perceptual tasks lack significant cultural expectations or gender-based stereotypes. The lack of gender difference in such tasks might indicate that the previously reported gender difference was not biological. Nevertheless, it is important to note that we did not empirically test whether gender-related cultural stereotypes about visual perception abilities exist, and there is no direct evidence of the existence of such stereotypes in either Western or non-Western cultures. Indeed, phenomena such as ‘‘refrigerator blindness’’ suggest that some level of cultural influence on visual perception cannot be fully ruled out.[41] This limitation highlights an essential area for future research. Subsequent studies should aim to quantify the strength of gender-related cultural stereotypes across different domains. To systematically investigate the potential cultural origins of gender differences in confidence and metacognition, researchers could quantify the degree to which gender stereotypes exist in different domains. Such analyses should directly compare gender differences across fields that have been empirically verified as either strongly embodying cultural stereotypes about gender abilities versus fields lacking such stereotypes. Additionally, to investigate biological differences in metacognition, future research should examine neural correlates of metacognition in each gender, with a special focus on regions in the prefrontal cortex previously identified as important for metacognition.[42–46] Overall, a comprehensive understanding of gender differences in confidence and metacognition should consider the interplay between domain-specific influences, cultural expectations, and biological differences.

One limitation of the current study is that six of the 10 datasets had relatively small sample sizes, potentially limiting the statistical power and generalizability of our results. Thus, our conclusions here should be seen as preliminary and should be replicated in additional large datasets. Another important question is whether our results may show variability across different countries or cultures. The current datasets were collected in five different countries (the US, Australia, France, Poland, and Taiwan) but because of the differences in sample size between studies, we cannot make conclusions about the cross-cultural generalizability of our results. Future studies should examine how confidence and metacognition in perceptual decision-making may vary across cultures.

In conclusion, our study found no difference between males and females in confidence or metacognitive ability in the context of perceptual decision-making tasks. This finding suggests that there may be no or limited inherent metacognitive disparities between the genders.

## Limitations of the study

One limitation of the current study is that 6 of the 10 datasets had relatively small sample sizes, potentially limiting the statistical power and generalizability of our results. Thus, our conclusions here should be seen as preliminary and should be replicated in additional large datasets. Another important question is whether our results may show variability across different countries or cultures. The current datasets were collected in five different countries (the US, Australia, France, Poland, and Taiwan) but because of the differences in sample size between studies, we cannot make conclusions about the cross-cultural generalizability of our results. Future studies should examine how confidence and metacognition in perceptual decision-making may vary across cultures.

## RESOURCE AVAILABILITY

### Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Kai Xue (kxue33@gatech.edu).

### Materials availability

This study did not generate new unique reagents or materials.

### Data and code availability

- All data used in this study are available on OSF (http://ost.io/s46pr/). The datasets were obtained from the Confidence Database (https://osf.io/s46pr/) and include Rouault_2018_Expt1, Rouault_2018_Expt2, Haddara_2022_Expt1, Zheng_2023, VanBoxtel_2019_Expt1, VanBoxtel_2019_Expt2, Siedlecka_2019, Martin_unpub, Gajdos_2019, and Skora_2016.
- All analysis code used in this study is available on OSF (http://ost.io/7agj6/).
- The code includes all scripts used for data analysis and generation of figures presented in this paper.

## ACKNOWLEDGMENTS

This work was supported by the National Institutes of Health (award: R01MH119189) and the Office of Naval Research (award: N00014-20-1-2622). We thank Marion Rouault for sharing gender information for two of her studies shared on the Confidence Database.

## AUTHOR CONTRIBUTIONS

K.X., C.P., and D.R. conceived the idea. K.X. and Y.Z. did the analysis. K.X. and D.R. wrote the paper. K.H. and M.A. reviewed and edited the paper.

## DECLARATION OF INTERESTS

The authors declare no competing interests.

## STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - ○ Experiment 1
  - ○ Experiment 2
  - ○ Experiment 3
  - ○ Experiment 4
  - ○ Experiment 5
  - ○ Experiment 6
  - ○ Experiment 7
  - ○ Experiment 8
  - ○ Experiment 9
  - ○ Experiment 10
- METHOD DETAILS
- QUANTIFICATION AND STATISTICAL ANALYSIS

## SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.111375.

Received: May 2, 2024
Revised: July 12, 2024
Accepted: November 8, 2024
Published: November 12, 2024

## REFERENCES

1. Guggenmos, M., Wilbertz, G., Hebart, M.N., and Sterzer, P. (2016). Mesolimbic confidence signals guide perceptual learning in the absence of external feedback. J.I. Gold, ed. *5*, e13388. https://doi.org/10.7554/eLife.13388.
2. Gilbert, S.J., Bird, A., Carpenter, J.M., Fleming, S.M., Sachdeva, C., and Tsai, P.C. (2020). Optimal use of reminders: Metacognition, effort, and cognitive offloading. J. Exp. Psychol. Gen. *149*, 501–517. https://doi.org/10.1037/xge0000652.
3. Desender, K., Boldt, A., and Yeung, N. (2018). Subjective Confidence Predicts Information Seeking in Decision Making. Psychol. Sci. *29*, 761–778. https://doi.org/10.1177/0956797617744771.
4. Bahrami, B., Olsen, K., Latham, P.E., Roepstorff, A., Rees, G., and Frith, C.D. (2010). Optimally interacting minds. Science *329*, 1081–1085. https://doi.org/10.1126/science.1185718.
5. Pescetelli, N., and Yeung, N. (2021). The role of decision confidence in advice-taking and trust formation. J. Exp. Psychol. Gen. *150*, 507–526. https://doi.org/10.1037/xge0000960.
6. Fisher, R. (1998). Thinking About Thinking: Developing Metacognition in Children. Early Child. Dev. Care *141*, 1–15. https://doi.org/10.1080/0300443981410101.
7. Liliana, C., and Lavinia, H. (2011). Gender Differences in Metacognitive Skills. A Study of the 8th Grade Pupils in Romania. Procedia Soc. Behav. Sci. *29*, 396–401. https://doi.org/10.1016/j.sbspro.2011.11.255.
8. Mohamed, A.H.H. (2012). The Relationship Between Metacognition and Self-regulation in Young Children. Procedia Soc. Behav. Sci. *69*, 477–486. https://doi.org/10.1016/j.sbspro.2012.11.436.
9. Muawiyah, D., Yamtinah, S., and Indriyanti, N.Y. (2019). Modelling testlet instrument in blended learning design to assess students' metacognition in the environmental chemistry course. J. Phys. Conf. Ser. *1157*, 042012. https://doi.org/10.1088/1742-6596/1157/4/042012.
10. Schraw, G. (1998). Promoting General Metacognitive Awareness. Instr. Sci. *26*, 113–125. https://doi.org/10.1023/A:1003044231033.
11. Brabender, V., and Boardman, S.K. (1977). Sex Differences in Self-Confidence as a Function of Feedback and Social Cues. Psychol. Rep. *41*, 1007–1010. https://doi.org/10.2466/pr0.1977.41.3.1007.
12. Cho, S.Y. (2017). Explaining Gender Differences in Confidence and Overconfidence in Math. SSRN Electron J. https://doi.org/10.2139/ssrn.2902717.
13. Lundeberg, M.A., Fox, P.W., and Punćcohaŕ, J. (1994). Highly confident but wrong: Gender differences and similarities in confidence judgments. J. Educ. Psychol. *86*, 114–121. https://doi.org/10.1037/0022-0663.86.1.114.
14. Pallier, G. (2003). Gender Differences in the Self-Assessment of Accuracy on Cognitive Tasks. Sex. Roles *48*, 265–276. https://doi.org/10.1023/A:1022877405718.
15. Rivers, M.L., Fitzsimmons, C.J., Fisk, S.R., Dunlosky, J., and Thompson, C.A. (2021). Gender differences in confidence during number-line estimation. Metacogn. Learn. *16*, 157–178. https://doi.org/10.1007/s11409-020-09243-7.
16. Ross, J.A., Scott, G., and Bruce, C.D. (2012). The Gender Confidence Gap in Fractions Knowledge: Gender Differences in Student Belief-Achievement Relationships: The Gender Confidence Gap in Fractions Knowledge. Sch. Sci. Math. *112*, 278–288. https://doi.org/10.1111/j.1949-8594.2012.00144.x.
17. Sarsons, H., and Xu, G. (2021). Confidence Men? Evidence on Confidence and Gender among Top Economists. AEA Pap. Proc. *111*, 65–68. https://doi.org/10.1257/pandp.20211086.
18. Seow, T.X.F., and Gillan, C.M. (2020). Transdiagnostic Phenotyping Reveals a Host of Metacognitive Deficits Implicated in Compulsivity. Sci. Rep. *10*, 2883. https://doi.org/10.1038/s41598-020-59646-4.
19. Michalak, R., Rysavy, M.D., and Wessel, A. (2017). Students' perceptions of their information literacy skills: the confidence gap between male and female international graduate students. J. Acad. Librariansh *43*, 100–104. https://doi.org/10.1016/j.acalib.2017.02.003.
20. Al-Hilawani, Y.A. (2001). Examining Metacognition in Hearing and Deaf/Hard of Hearing Students: A Comparative Study. Am. Ann. Deaf *146*, 45–50. https://doi.org/10.1353/aad.2012.0101.
21. Erhan, A. (2016). Examining the relation between metacognitive understanding of what is listened to and metacognitive awareness levels of secondary school students. Educ. Res. Rev. *11*, 390–401. https://doi.org/10.5897/ERR2015.2616.

22. Lemieux, C.L., Collin, C.A., and Watier, N.N. (2019). Gender differences in metacognitive judgments and performance on a goal-directed wayfinding task. J. Cogn. Psychol. *31*, 453–466. https://doi.org/10.1080/20445911.2019.1625905.

23. Chantharanuwong, W., Thatthong, K., Yuenyong, C., and Thomas, G.P. (2012). Exploring the Metacognitive Orientation of the Science Classrooms in a Thai Context. Procedia - Soc Behav Sci. *46*, 5116–5123. https://doi.org/10.1016/j.sbspro.2012.06.393.

24. Hoogervorst, K., Banellis, L., Fardo, F., Xue, K., Rahnev, D., and Allen, M. (2024). Gender differences in metacognition: global and local contrasts in bias and efficiency. Preprint at PsyArXiv. https://doi.org/10.31234/osf.io/2gwky.

25. Rouault, M., Seow, T., Gillan, C.M., and Fleming, S.M. (2018). Psychiatric Symptom Dimensions Are Associated With Dissociable Shifts in Metacognition but Not Task Performance. Biol. Psychiatry *84*, 443–451. https://doi.org/10.1016/j.biopsych.2017.12.017.

26. Fleming, S.M., and Lau, H.C. (2014). How to measure metacognition. Front. Hum. Neurosci. *8*, 443. https://doi.org/10.3389/fnhum.2014.00443.

27. Rahnev, D. (2023). Measuring metacognition: A comprehensive assessment of current methods. Preprint at PsyArXiv. https://doi.org/10.31234/osf.io/waz9h.

28. Desender, K., Vermeylen, L., and Verguts, T. (2022). Dynamic influences on static measures of metacognition. Nat. Commun. *13*, 4208. https://doi.org/10.1038/s41467-022-31727-0.

29. Maniscalco, B., and Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. Conscious. Cogn. *21*, 422–430. https://doi.org/10.1016/j.concog.2011.09.021.

30. Xue, K., Shekhar, M., and Rahnev, D. (2021). Examining the robustness of the relationship between metacognitive efficiency and metacognitive bias. Conscious. Cogn. *95*, 103196. https://doi.org/10.1016/j.concog.2021.103196.

31. Rahnev, D., Desender, K., Lee, A.L.F., Adler, W.T., Aguilar-Lleyda, D., Akdoğan, B., Arbuzova, P., Atlas, L.Y., Balcı, F., Bang, J.W., et al. (2020). The Confidence Database. Nat. Hum. Behav. *4*, 317–325. https://doi.org/10.1038/s41562-019-0813-1.

32. Shaqiri, A., Roinishvili, M., Grzeczkowski, L., Chkonia, E., Pilz, K., Mohr, C., Brand, A., Kunchulia, M., and Herzog, M.H. (2018). Sex-related differences in vision are heterogeneous. Sci. Rep. *8*, 7521. https://doi.org/10.1038/s41598-018-25298-8.

33. Rausch, M., and Zehetleitner, M. (2023). Evaluating false positive rates of standard and hierarchical measures of metacognitive accuracy. Metacogn. Learn. *18*, 863–889. https://doi.org/10.1007/s11409-023-09353-y.

34. Gentile, B., Grabe, S., Dolan-Pascoe, B., Twenge, J.M., Wells, B.E., and Maitino, A. (2009). Gender Differences in Domain-Specific Self-Esteem: A Meta-Analysis. Rev. Gen. Psychol. *13*, 34–45. https://doi.org/10.1037/a0013689.

35. Lirgg, C.D. (1991). Gender Differences In Self-Confidence in Physical Activity: A Meta-Analysis of Recent Studies. J. Sport Exerc. Psychol. *13*, 294–310. https://doi.org/10.1123/jsep.13.3.294.

36. Hyde, J.S. (2014). Gender Similarities and Differences. Annu. Rev. Psychol. *65*, 373–398. https://doi.org/10.1146/annurev-psych-010213-115057.

37. Eagly, A.H. (2009). The his and hers of prosocial behavior: An examination of the social psychology of gender. Am. Psychol. *64*, 644–658. https://doi.org/10.1037/0003-066X.64.8.644.

38. Wood, W., and Eagly, A.H. (2012). Biosocial Construction of Sex Differences and Similarities in Behavior. Adv. Exp. Soc. Psychol. *46*, 55–123. https://doi.org/10.1016/B978-0-12-394281-4.00002-7.

39. Rahnev, D., Koizumi, A., McCurdy, L.Y., D'Esposito, M., and Lau, H. (2015). Confidence Leak in Perceptual Decision Making. Psychol. Sci. *26*, 1664–1680. https://doi.org/10.1177/0956797615595037.

40. Eriksson, K. (2020). Gender Differences in the Interest in Mathematics Schoolwork Across 50 Countries. Front. Psychol. *11*, 578092. https://doi.org/10.3389/fpsyg.2020.578092.

41. Macnab, A.J., and Bennett, M. (2005). Refrigerator blindness: selective loss of visual acuity in association with a common foraging behaviour. Can. Med. Assoc. J. *173*, 1494–1495. https://doi.org/10.1503/CMAJ.051393.

42. Fleming, S.M., Huijgen, J., and Dolan, R.J. (2012). Prefrontal Contributions to Metacognition in Perceptual Decision Making. J. Neurosci. *32*, 6117–6125. https://doi.org/10.1523/JNEUROSCI.6489-11.2012.

43. Rounis, E., Maniscalco, B., Rothwell, J.C., Passingham, R.E., and Lau, H. (2010). Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. Cogn. Neurosci. *1*, 165–175. https://doi.org/10.1080/17588921003632529.

44. Saccenti, D., Moro, A.S., Sassaroli, S., Malgaroli, A., Ferro, M., and Lamanna, J. (2024). Neural correlates of metacognition: Disentangling the brain circuits underlying prospective and retrospective second-order judgments through noninvasive brain stimulation. J. Neurosci. Res. *102*, e25330. https://doi.org/10.1002/jnr.25330.

45. Shekhar, M., and Rahnev, D. (2018). Distinguishing the Roles of Dorsolateral and Anterior PFC in Visual Metacognition. J. Neurosci. *38*, 5078–5087. https://doi.org/10.1523/JNEUROSCI.3484-17.2018.

46. Xue, K., Zheng, Y., Rafiei, F., and Rahnev, D. (2023). The timing of confidence computations in human prefrontal cortex. Cortex *168*, 167–175. https://doi.org/10.1016/j.cortex.2023.08.009.

47. Haddara, N., and Rahnev, D. (2022). The Impact of Feedback on Perceptual Decision-Making and Metacognition: Reduction in Bias but No Change in Sensitivity. Psychol. Sci. *33*, 259–275. https://doi.org/10.1177/09567976211032887.

48. Orchard, E.R., Dakin, S.C., and van Boxtel, J.J.A. (2022). Internal noise measures in coarse and fine motion direction discrimination tasks and the correlation with autism traits. J. Vis. *22*, 19. https://doi.org/10.1167/jov.22.10.19.

49. Siedlecka, M., Wereszczyński, M., Paulewicz, B., and Wierzchoń, M. (2020). Visual awareness judgments are sensitive to accuracy feedback in stimulus discrimination tasks. Conscious. Cogn. *86*, 103035. https://doi.org/10.1016/j.concog.2020.103035.

50. Gajdos, T., Fleming, S.M., Saez Garcia, M., Weindel, G., and Davranche, K. (2019). Revealing subthreshold motor contributions to perceptual confidence. Neurosci. Conscious. *2019*, niz001. https://doi.org/10.1093/nc/niz001.

51. Skóra, Z., and Wierzchoń, M. (2016). The level of subjective visibility at different stages of memory processing. J Cogn Psychol. *28*, 965–976. https://doi.org/10.1080/20445911.2016.1225745.

52. Zheng, Y., Xue, K., Shekhar, M., and Rahnev, D. (2024). Similar computational noise for perceptual decision making with confidence, expectation, and reward. Preprint at PsyArXiv. https://doi.org/10.31234/osf.io/ydx6z.

53. Rahnev, D., and Fleming, S.M. (2019). How experimental procedures influence estimates of metacognitive ability. Neurosci. Conscious. *2019*, niz009. https://doi.org/10.1093/nc/niz009.

54. Arnold, D.H., Johnston, A., Adie, J., and Yarrow, K. (2023). On why we lack confidence in some signal-detection-based analyses of confidence. Conscious. Cogn. *113*, 103532. https://doi.org/10.1016/j.concog.2023.103532.

55. Murad, M.H., Wang, Z., Chu, H., and Lin, L. (2019). When continuous outcomes are measured using different scales: guide for meta-analysis and interpretation. BMJ *364*, k4817. https://doi.org/10.1136/bmj.k4817.

56. Bradley, M.T., and Brand, A. (2013). Alpha Values as a Function of Sample Size, Effect Size, and Power: Accuracy over Inference. Psychol. Rep. *112*, 835–844. https://doi.org/10.2466/03.49.PR0.112.3.835-844.

# STAR★METHODS

## KEY RESOURCES TABLE

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| Deposited data | | |
| Confidence data | https://osf.io/s46pr/ | NA |
| Software and algorithms | | |
| Analysis code | https://osf.io/7agj6/ | NA |

## EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

To assess the gender difference in confidence or metacognitive ability, we explored the datasets available in the Confidence Database.[31] The Confidence Database is a large repository of open data from experiments that include confidence ratings (171 datasets were included as of July 2023). We examined all available datasets and found that only eight of them included information about the gender of each individual subject. After submitting an initial preprint with eight datasets, Dr. Marion Rouault shared gender information for two more datasets already shared on the Confidence Database, thus bringing the total count to 10 datasets. We call these datasets "Experiments 1–10" and arrange them in decreasing order of their sample size. In all experiments, confidence was elicited without any incentive schemes in place (e.g., payoffs) that could differentially affect confidence ratings between genders. We contacted all authors to obtain detailed information about how gender was determined in each of the 10 experiments. We obtained relevant information for all experiments, except for Experiment 8. Gender was self-reported in eight out of the nine remaining experiments. The only exception was Experiment 9 where gender was identified by the researchers based on name and appearance (we repeated all analyses excluding Experiment 9 and all conclusions remained unchanged). Seven of the nine experiments only used two options (male/female), whereas Experiments 4 and 7 included a third option for "other", although no subject selected that option. Information about other aspects of each experiment including sample size, number of trials per subject, confidence scale granularity, task, and number of difficulty levels is provided in Table 1. Detailed information about all experiments can be found in the original publications as well as the accompanying files on the Confidence Database (https://osf.io/s46pr/). Below, we give a brief summary of each experiment.

### Experiment 1

Experiment 1 is referred to as Rouault_2018_Expt1 in the Confidence Database and reports the data from Experiment 1 in Rouault et al. (2018). Subjects (*N* = 498; 237 males and 261 females) made perceptual judgments concerning which of the two boxes contained a higher number of dots. One box was always half-filled with 313 out of 625 dots, while the other box contained between 1 and 70 more dots compared to the half-filled box, leading to a total of 70 difficulty levels. The stimuli were presented for 300 ms. Confidence was indicated with a separate button press using an 11-point scale. There were 210 trials per subject.

### Experiment 2

Experiment 2 is referred to as Rouault_2018_Expt2 in the Confidence Database and reports the data from Experiment 2 in Rouault et al. (2018). Subjects (*N* = 487; 240 males and 247 females) underwent the same experiment procedure and design as in Experiment 1, except for a continuous 2-down-1-up staircase procedure to maintain a constant level of performance both during the experiment. Confidence was indicated with a separate button press using a 6-point scale.

### Experiment 3

Experiment 3 is referred to as Haddara_2022_Expt1 in the Confidence Database and reports the data from Experiment 1 in.[47] Subjects (*N* = 443; 244 males and 199 females) made perceptual judgments concerning whether the letter X or O (Task 1) or the color red or blue (Task 2) occurred more frequently in a 7x7 grid. The number of the more frequently presented letters was fixed to 30/49, whereas the number of the more frequently presented colors was fixed to 27/49. The stimuli were presented for 500 ms. Confidence was indicated with a separate button press using a 4-point scale. Each subject completed 330 trials from Task 1 and 150 trials from Task 2. Here we combine the data from the two tasks.

### Experiment 4

Experiment 4 is referred to as Zheng_2023 in the Confidence Database and reports the data from Zheng et al. (2024). Subjects (*N* = 315; 123 males and 192 females) judged which of two black squares contained more white dots. This experiment features multiple conditions, and we only analyzed the condition in which subjects gave confidence ratings. The difficulty level was fixed by having the number of white dots in each black square fixed (100 in one and 85 in the other), and the stimulus was presented for 300 ms. Confidence was indicated with a separate button press using a 2-point scale. There were 100 trials per subject.

### Experiment 5

Experiment 5 is referred to as VanBoxtel_2019_Expt1 in the Confidence Database and reports the data from Experiment 1 in.[48] Subjects (*N* = 45; 31 males and 14 females) judged the direction of dot motion. The difficulty level was manipulated by altering the coherence of the dot motion. There are eight coherence levels. The stimulus was presented for 750 ms. Confidence was indicated together with the decision on a 4-point scale. There were 1600 trials per subject.

### Experiment 6

Experiment 6 is referred to as VanBoxtel_2019_Expt2 in the Confidence Database and reports the data from Experiment 2 in.[48] Subjects (*N* = 37; 25 males and 12 females) judged the direction of the dot motion. The difficulty level was manipulated by the coherence of the dot motion, and there were eight levels of difficulty. The stimulus was presented for 750 ms. Confidence was indicated together with the decision on a 4-point scale. There were 1600 trials per subject.

### Experiment 7

Experiment 7 is referred to as Siedlecka_2019 in the Confidence Database and reports the data from.[49] Subjects (*N* = 37; 7 males and 30 females) judged whether the Gabor gratings were tilted toward left or right. The difficulty level was fixed and was determined by a staircase procedure targeting 71% of accuracy. The stimulus was presented for 500 ms. Two within-subject conditions were included in the experiment: with and without accuracy feedback. We analyzed the data across both conditions. In the feedback condition, the feedback was provided after the confidence rating. Confidence was indicated with a separate button press on a 4-point scale. There were 400 trials per subject.

### Experiment 8

Experiment 8 is referred to as Martin_unpub in Confidence Database and the manuscript associated with this dataset is still unpublished. In this experiment, two figures (square or diamond) were presented on each side of the screen. Subjects (*N* = 22; 13 males and 9 females) discriminated whether the square was presented in the left or the right. A staircase procedure was employed to keep performance around 70% correct. Confidence was indicated with a separate button press on a 4-point scale. There were 600 trials per subject.

### Experiment 9

Experiment 9 is referred to as Gajdos_2019 in the Confidence Database and reports the data from.[50] Subjects (*N* = 22; 9 males and 13 females) judged whether the Gabor patch presented was whether vertical or horizontal. The difficulty level was fixed for each subject and was determined by a staircase procedure targeting a 79.4% accuracy. The stimulus was presented for 33 ms. Confidence was indicated with a separate button press on a 4-point scale. There were 699 trials per subject.

### Experiment 10

Experiment 10 is referred to as Skora_2016 in the Confidence Database and reports the data from.[51] Subjects (*N* = 19; 6 males and 13 females) first memorized the displayed items (memory display) and then performed a change detection task (task display). The difficulty level was fixed for each subject and is determined by a staircase procedure targeting 71% accuracy. The memory display was presented for 250 ms. A within-subject manipulation changed the time between the disappearance of the original stimulus and the appearance of the cue. There were three different delay conditions (50 ms, 100 ms, and 1000 ms). The three conditions were combined for analysis purposes. Confidence was indicated with a separate button press on a 4-point scale. There were 400 trials per subject.

## METHOD DETAILS

For all 10 datasets, following standard practice in our lab,[30,45,46,52] we excluded subjects with accuracy lower than 55% or higher than 95% (because floor or ceiling effects on task performance result in noisy estimates of metacognition) and subjects who only used one confidence rating (which makes it impossible to estimate metacognitive ability). Note that no subject used only one confidence rating, so nobody was excluded based on that criterion. These criteria led to the exclusion of 0, 0, 22, 20, 0, 0, 3, 0, 2, and 1 subjects in the 10 experiments, respectively (corresponding to exclusion rates of 0%, 0%, 6.35%, 4.97%, 0%, 0%, 8.11%, 0%, 9.09%, and 5.26%). Note that Experiments 1 and 2 only include subjects that passed somewhat overlapping exclusion criteria used in Rouault et al. (2018).

We computed task performance (d’) using the formula:

$$d' = \phi^{-1}(HR) - \phi^{-1}(FAR)$$

where *HR* and *FAR* are the hit and false alarm rates associated with the decision criterion. We computed M-Ratio using the codes provided by Maniscalco & Lau (2012). In specific, we first calculate meta-d' which is the value of stimulus sensitivity (d') that best describes the observed pattern of confidence responses given the assumptions of signal detection theory. Then, the M-Ratio is computed by dividing the meta-d' by d', and this measure quantifies the extent to which confidence ratings discriminate between correct and incorrect responses while controlling for the first-order task performance. It is worth noting that prior research has found that including multiple difficulty levels can lead to inflated estimates of metacognitive ability compared to using a single difficulty level.[53] To address this potential issue, we conducted analyses on the two experiments in our study that employed an online staircase to examine whether the stimulus variability experienced by each gender differed significantly. We found no significant difference in the degree of stimulus variability experienced by males and females (Expt 2: F(239, 256) = 1.16, $p = 0.25$; Expt8: F(12, 7) = 0.57, $p = 0.37$). This finding confirms that the use of an online staircase does not contaminate the interpretation of the M-Ratio in our study, as it does not introduce uneven inflation to the M-Ratio for each gender.

Additionally, it is important to acknowledge recent discussions in the field regarding potential bias in M-Ratio estimates due to violations of the normality assumption in experiential distributions[54] or because of either explicit or implicit difference response caution.[28] While such distortions may occur, they are expected to affect both genders equally in our study, thus preserving the validity of our comparative results. Accordingly, we confirmed that there are no significant differences in reaction times between genders ($p > 0.05$ for all 10 experiments), supporting the conclusion that there are no systematic differences in response strategies between males and females in our study.

Furthermore, while some experiments included more than one difficulty level, we did not compute separate M-Ratio values for each difficulty condition. This was due to feasibility issues, as the number of difficulty levels could be as high as 70 in some cases (and the number of trials per difficulty level as low as three). As all participants experienced the full range of difficulty levels in the relevant experiments, we reason that this methodological choice would not affect the overall conclusion. However, mixing difficulty levels could increase the overall noise in M-ratio estimation, potentially reducing the chance of detecting significant differences between the two genders.

## QUANTIFICATION AND STATISTICAL ANALYSIS

To investigate the gender difference in the measure scores (d', confidence, and M-Ratio in the main paper, as well as meta-d' and confidence for correct versus incorrect trials in Figures S1 and S2), we performed independent samples t-tests for the measure scores of males and females across each experiment. Further, in order to compare the effect size of gender across experiments, we conducted random-effect meta-analyses. Studies were weighted by their inverse variance which reflects the study sample size. We used Hedge's g as the bias-corrected effect size of mean group difference. We used the R package, meta, to estimate the random-effect inverse-variance weighted mean and standard error of performance within groups. To quantify the heterogeneity in effect sizes across studies, we computed $I^2$, which is the percentage of variance across studies that is due to heterogeneity rather than chance.

In order to ensure the robustness of our meta-analysis, we conducted power analysis using the R package, dmetar, to ascertain whether our sample size was adequate to detect even small effect sizes. Specifically, we examined the power needed to detect an effect where the standard mean difference (SMD), a measure equivalent to Cohen's d, equals 0.2, which is considered a small effect size in a meta-analysis.[55] We set the alpha level at 0.05 and aimed for a power level of 0.80, which follows recommendations for psychological research.[56] The power analysis revealed that with our current sample size of 10 studies with a total number of 1,887 subjects, we achieved a power of 0.99. This indicates that our meta-analysis is sufficiently powered to detect a small effect size of SMD = 0.2 with a probability of 99%, surpassing the conventional threshold of 80%.

Additionally, since age may influence perceptual performance, we conducted multiple regression analyses to predict confidence from accuracy, age, response time, difficulty level, and gender (Table S1). This allowed us to rule out the possibility that any observed gender effects on confidence were confounded by age or accuracy differences.