The statistical fragility of animal cognition findings: a meta-meta-analytic reappraisal

Ayumi Mizuno[1], Pietro Pollo[2,3], Malgorzata Lagisz[1,2], Lauren Guillette[4], Masayo Soma[5], and Shinichi Nakagawa[1,2]

1. Department of Biological Sciences, Faculty of Science, University of Alberta, Edmonton, AB, Canada
2. Evolution & Ecology Research Centre, School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, NSW, Australia
3. School of Environmental and Life Sciences, University of Newcastle, Newcastle, NSW, Australia
4. Department of Psychology, Faculty of Science, University of Alberta, Edmonton, AB, Canada
5. Department of Biology, Faculty of Science, Hokkaido University, Sapporo, Hokkaido, Japan

ORCID iDs
Ayumi Mizuno: 0000-0003-0822-5637
Pietro Pollo: 0000-0001-6555-5400
Malgorzata Lagisz: 0000-0002-3993-6127
Lauren Guillette: 0000-0002-8777-6543
Masayo Soma: 0000-0002-8596-1956
Shinichi Nakagawa: 0000-0002-7765-5182

Corresponding author
Ayumi Mizuno
Department of Biological Sciences, Faculty of Science, University of Alberta, Edmonton, AB, Canada
ayumi.mizuno5@gmail.com/amizuno@alberta.ca

Abstract

How reliable is the evidence in animal cognition research? Concerns are mounting over the statistical robustness of this and other fields. Many primary studies rely on small samples and rarely report null results, while meta-analyses sometimes overlook publication bias, all of which may contribute to unreliable conclusions. We conducted a second-order meta-analysis across 28 published meta-analytical papers in the animal cognition field to evaluate three inferential metrics: statistical power, Type M (magnitude) error, and Type S (sign) error. These were calculated at the primary study and meta-analysis levels. To approximate the true effect, we used the mean effect size from each meta-analysis; when publication bias was detected, we applied a correction to mitigate potential overestimation. Our results indicate low statistical power and inflated effect sizes in both primary studies and meta-analyses. After bias correction, on average, power decreased from 17% to 9% and effect size values decreased from 82% to 45%. Type M errors were common, indicating that statistically significant results often exaggerated the underlying effect sizes. For improving the reliability of animal cognition research, we recommend preregistering and transparently reporting both primary and secondary studies. We also call for the routine application of publication bias correction in meta-analytical syntheses.

Keywords

Comparative psychology, Evidence synthesis, Meta-research, Quantitative synthesis, Replication crisis, Research methodology

1. Introduction

Research on animal cognition has expanded rapidly over recent decades [1]. The number of both review and empirical studies has grown faster than in the broader life sciences [1], revealing diverse cognitive abilities across taxa from invertebrates to mammals. These findings have important implications for comparative psychology and for evolutionary and ecological theory. For instance, cognitive traits have been increasingly recognised as key drivers of behavioural flexibility, niche expansion, and adaptive divergence, central to understanding species persistence in changing environments [2–4]. However, concerns have emerged about the statistical reliability and reproducibility of findings in this field, partly due to small sample sizes, selective reporting of significant results, and methodological heterogeneity [5,6]. The latter, which is a strength of the field because variation often reflects necessary adaptations to species' perceptual and ecological worlds [7], but it also complicates cross-study comparisons. To address these issues, researchers have increasingly adopted quantitative synthesis methods [5,8]. Meta-analysis is a powerful tool for synthesising results across studies and estimating overall effect sizes, often regarded as proxies for "true" effects [9]. Yet, its reliability depends critically on the quality of the included primary studies.

Similar concerns occur in other disciplines [10–14]. For example, Yang et al. [13] showed that adjusting for publication bias substantially lowered effect-size estimates and statistical support for previously reported relationships in ecology and evolution. In psychology, Bartoš et al. [11] found that over half of psychological meta-analyses exaggerated both the evidence for and the magnitude of effects. These findings highlight the need for rigorous meta-analytic practices and better reporting standards in both primary and secondary research.

Here, we evaluate the statistical reliability of findings in animal cognition by estimating three complementary metrics, statistical power, Type M (magnitude) error, and Type S (sign) error [12–16]. We focus on meta-analyses examining learning, memory, and decision-making in non-human animals, excluding those centred on humans, disease models, or purely physiological traits (e.g., visual acuity). These metrics, derived from the estimated true effect size, its standard error, and a significance threshold, provide complementary views of inferential reliability. We calculate them for both meta-analyses and their constituent primary studies and compare uncorrected and bias-corrected estimates to evaluate how publication bias affects evidential strength. We aim to clarify whether current findings offer a reliable foundation for theoretical and empirical work in animal cognition.

2. Materials and methods

Literature inclusion and dataset compilation

This study is pre-registered [17], and Figure 1 outlines the workflow from data extraction and bias correction to the calculation and aggregation of reliability metrics. We limited our analysis to three widely used effect size metrics, i.e., log response ratio (lnRR), standardised mean difference (SMD), and Fisher's *z*-transformed correlation (Zr), due to their cross-study comparability, approximate normality, and analytically derivable sampling variances [18]. We

retained all effect sizes in their original metric (lnRR, SMD, and Zr) and did not convert them to a different metric. Based on this criterion, we included 28 out of 49 meta-analytical studies [19–46] identified in our previous systematic mapping of the animal cognition literature [1]. The remaining 21 studies did not use lnRR, SMD, or Zr as their main effect sizes or provide sufficient information about effect sizes and their variances. The study selection process is illustrated in Figure S1. We retrieved the data from published figures, tables, supplementary files, and online repositories (e.g., OSF, Dryad). Among 28 meta-analytical studies, 17 either reported both effect sizes and corresponding sampling variances extracted from the primary studies or provided scripts detailing how these values were calculated. In such cases, we either directly extracted or computed effect sizes and variances based on the provided information. For three meta-analytical studies, effect sizes and variances were not reported; however, raw descriptive statistics (e.g., means, standard deviations, sample sizes) were available. These were used to compute the effect sizes and variances using standard meta-analytic formulas. When only forest plots (n = 7) or tables reporting sample sizes and effect sizes (n = 1) were available in meta-analyses, we extracted the necessary values using the WebPlotDigitizer (https://apps.automeris.io/wpd4/) and calculated the corresponding variances. When the data were not publicly available, we contacted the corresponding authors. Datasets that required proprietary software were excluded if access could not be obtained (e.g., Comprehensive Meta-Analysis software). For each included meta-analytical study, we recorded whether the authors assessed publication bias, which specific methods they used (e.g., Egger's test, funnel plots), whether any correction procedures were applied (e.g., trim-and-fill), and how the corrections affected the statistical significance of the reported effect sizes. Our final dataset consisted of 48 meta-analytic models, as several of the 28 eligible meta-analytical studies reported multiple distinct meta-analytic models within a single publication.

Estimating "true effects"

We followed the methodological framework of Yang et al. [13,14], which aims to estimate bias-corrected effect sizes by accounting for two common forms of publication bias. Small-study effects refer to the tendency for smaller studies to report larger effects, and the decline effect describes the tendency for earlier studies to report effects of higher magnitudes than more recent ones. Yang et al. [13,14]'s multilevel meta-regression model includes both bias terms and estimates the average effect size that would be expected in the absence of these biases. The resulting intercept is interpreted as the bias-corrected estimate of the true effect. To assess statistical power, Type M error, and Type S error, we first estimated the "true effect" size for each meta-analysis. We define the true effect as the underlying average effect size that would be expected in the absence of sampling error and publication bias. As a proxy, we used the meta-analytic mean effect size. When evidence of publication bias was detected, we applied the full bias-correction model; if only one type of bias (e.g., small-study effects but no decline effect) was detected, we used a reduced version of the model including only the significant bias term. If

no bias was detected, the uncorrected mean was used. These bias-corrected or uncorrected mean values served as the estimated "true effects" for subsequent analyses.

Estimating power and error rates

Using the estimated true effects, we calculated statistical power, and Type M and Type S errors at two levels. At the meta-analysis level, one estimate of each metric was obtained per meta-analysis, based on the corresponding average sampling variance. These estimates were then summarised across all meta-analyses using linear mixed-effects models, effectively constituting a second-order meta-analysis. In parallel, we calculated the same metrics at the primary-study level, using each study's standard error and the corresponding meta-analytic mean effect (bias-corrected when applicable, otherwise uncorrected) as the assumed "true effect." This two-level approach allowed us to compare inferential reliability both across and within meta-analyses. Statistical power, defined as the probability of detecting a statistically significant effect (two-sided, $\alpha = 0.05$), assuming the estimated true effect is correct; higher values indicate greater sensitivity to detect a true effect as statistically significant. Type M error, the expected degree of exaggeration in effect size estimates, conditional on statistical significance; values above 1 indicate overestimation, with larger numbers reflecting severe inflation. Type S error, the probability that a statistically significant result has the incorrect sign (i.e., false direction); high values indicate a greater risk of misinterpreting the effect size direction. At the meta-analysis level, we calculated these metrics using the average sampling variance across all effect sizes within each meta-analysis. At the primary study level, we applied the same assumptions as for the meta-analysis level - namely, that the meta-analytic mean effect (bias-corrected when applicable, otherwise uncorrected) represents the true effect, that effect size estimates are normally distributed, and that statistical significance is evaluated using a two-sided α of 0.05.

We calculated all estimates analytically, assuming normally distributed effect size estimates and using the standard formulas of Gelman and Carlin [16] as implemented by Yang et al. [13,14]. At the meta-analysis level, we summarised estimates of statistical power, Type M error, and Type S error using weighted linear regression (lm function in base R), with the number of effect sizes per meta-analysis as weights. At the primary study level, we first calculated the same metrics for each study. Then we aggregated them across studies using linear mixed-effects models (lmer function, lme4 v1.1.37 [47]), including study identity as a random effect. We implemented meta-analytic models and bias-correction regressions using the metafor package (v4.6.0) [48], and figures were created using ggplot2 v3.5.2 [49]. All analyses were conducted in R v4.4.2 [50]. The data and scripts we used are available at https://github.com/Ayumi-495/statistical_power_AnimCogn.

3. Results

Among the 28 eligible meta-analytical studies, 25 tested for potential publication bias. Ten of these also used bias-correction procedures, such as the trim-and-fill method, to adjust the estimates of overall effect sizes. Four meta-analytical studies identified possible publication bias

(e.g., significant Egger's tests or funnel-plot asymmetry) but did not adjust the effect sizes, often noting that the trim-and-fill method found no missing studies. Two meta-analytical studies explicitly stated that no evidence of publication bias was detected, and therefore no correction was made. The remaining nine meta-analytical studies mentioned assessing publication bias but provided insufficient detail on whether corrections were applied. Of the 48 meta-analytic means from the 28 meta-analytical studies, 31 were initially statistically significant, but 18 (58.1%) lost significance after publication bias correction.

Our analyses reveal that, at the meta-analysis level, average statistical power to detect the uncorrected mean effect was 82.2% (95% CI: 71.7%–94.2%; mean: 114%) but decreased to 44.9% (95% CI: 33.6%–60.0%; mean: 62.1%) after publication bias correction (Figure 2b, Table S1). Type M error increased from 1.11 (95% CI: 1.03–1.20; mean: 1.30) to 2.03 (95% CI: 1.37–3.02; mean: 5.84) after correcting for publication bias, and some error values were over 20 (Figure 2b, Table S1). Type S error was near zero before publication bias correction (0.06%, 95% CI: 0–0.19%; mean: 0.43%) but increased to 1.21% (95% CI: 0.34–2.35%; mean: 3.70%) after correction (Figure 2c, Table S1).

At the primary study level, we found that the average power was low 17.20% (95% CI: 16.4%–18.0%; mean: 23.1%) before publication bias correction and dropping further to 9.06% (95% CI: 8.75%–9.37%; mean: 12.2%) when using bias-corrected meta-analytic mean effect estimates (Figure 2a, Table S1). Type M error increased from 2.86 (95% CI: 2.76–2.95; mean: 3.48) to 7.79 (95% CI: 7.13–8.51; mean: 9.49), with some values exceeding 20 (Figure 2b, Table S1). Type S error increased from 2.69% (95% CI: 2.50–2.90%; mean: 4.34%) to 9.85% (95% CI: 9.25–10.5%; mean: 13.4%) after correction (Figure 2c, Table S1).

4. Discussion

How do the low statistical power and high Type M errors we found compare with those in other fields? At the meta-analysis level, statistical power in animal cognition (82.2% before correction, dropping to 44.9% after; Table S1) was similar to that in ecology and evolution (55% to 36%) and global change biology (40–67% before correction, declining thereafter) [13,14].

Type M errors in animal cognition meta-analyses sometimes exceeded a twentyfold overestimate (observed in 3 cases after correction; Figure 2b). This indicates that statistically significant findings in this field can substantially overstate the true effect sizes, creating a misleading impression of the magnitude of the phenomena. By contrast, ecology and evolution typically report Type M values between 1 and 2 after correction [13,14], implying only twofold or more minor exaggeration of true effects.

At the primary-study level, statistical power averaged 17.2% before correction and only 9.1% after correction for publication bias in animal cognition (Table S1), lower than values reported for ecology and evolution (23% before bias correction, 15 % after [14]), psychology (36 % after bias correction [12]), and neuroscience (21% after correction [51]), and comparable to the lowest bias-corrected estimates in medicine (9%: [52]). Effect-size inflation (Type M errors) was also greater (2.86 before correction, 7.79 after correction; some >20; Figure 2c, Table

S1), whereas other fields report lower values (ecology and evolution [14]: 2.5 to 4.0; global change biology [13]: 2 to 6). These results indicate that animal cognition studies have exceptionally low power and large effect-size exaggeration, making the field particularly vulnerable to biased inference.

In contrast to our findings for power and Type M error, we found that Type S error in animal cognition was very low at the meta-analysis level (0.06% uncorrected, 1.21% corrected; Table S1), below the values reported for ecology and evolution [14]. At the primary-study level, however, it rose from 2.69% to 9.85% after bias correction (Table S1), exceeding typical estimates in ecology and evolution (5–8% [14]) and global change biology (<5% [13]). Thus, although low-powered primary studies in animal cognition carry a higher risk of sign errors, meta-analytic synthesis appears to mitigate this, leaving the field relatively less affected than related disciplines.

In primary studies in animal cognition, small sample sizes and high individual variability are common [8]. These constraints, though often unavoidable, make significant findings more likely to arise from random upward fluctuations than from accurate estimation, leading to exaggerated effect sizes (i.e., high Type M error) [53]. The problem may be further exacerbated when subjects that fail to complete cognitive tasks are excluded, a practice that has been noticed in the field [5,54,55]. Although sometimes methodologically justified, such exclusions further reduce sample sizes and can bias estimates by systematically omitting lower-performing individuals or alternative behavioural strategies, inflating effect sizes and reducing generalisability.

Problems at the primary-study level inevitably carry over into meta-analyses. When available studies are few, underpowered, or selectively report results, the meta-analysis rests on weak and biased evidence. Under these conditions, estimates are prone to inaccuracy, overstate true effects, and display unstable statistical power [51,56–58]. Rather than correcting these weaknesses, meta-analyses can unintentionally amplify them, creating a false sense of reliability when significant results are reported without considering underlying study quality or structure [56,59].

Publication bias remains a pervasive methodological challenge in meta-analytical studies, including those in animal cognition. While detection tools such as Egger's regression and funnel plots were commonly used, correction methods appeared in fewer than 10% of studies and relied almost exclusively on the trim-and-fill method. In some cases, publication bias was detected using Egger's regression, but not with the trim-and-fill method. As a result, the authors reported that they could not impute any missing studies and therefore could not adjust the estimate.

Although widely used, the trim-and-fill method has well-known limitations: it struggles to identify or adjust for missing studies under high heterogeneity or when asymmetry stems from causes other than publication bias [60]. It is also difficult to apply in multilevel or multivariate models, limiting its use in modern analyses [60]. Egger's regression can be used to adjust estimates (e.g., by extrapolating to the intercept), but this approach was rarely implemented in the meta-analytical studies we included. Future work should consider regression-based

alternatives that address selective reporting and statistical dependence, such as the robust estimation approach recently proposed by Yang et al. [61], which has performed reliably across hundreds of meta-analyses in ecology and evolution.

To address the statistical, methodological, and interpretive challenges identified in this study, we recommend a multi-pronged approach involving both empirical researchers and meta-analysts (Table 1). At the primary-study level, practices such as larger sample sizes, reporting null or non-significant results, and documenting task failures alongside successes are essential to mitigate selective reporting, reduce publication bias, and improve transparency [6,62–64]. At the meta-analytic level, publication bias should be detected and corrected using methods tailored to the sample size and heterogeneity of the data, with careful consideration of the quality and structure of the contributing studies.

More broadly, at both the primary-study and meta-analytic levels, we recommend two key open-science practices: (1) pre-registration and (2) data and code sharing. Pre-registration requires researchers to specify their aims, questions, predictions, experimental designs, and analytical methods before data collection, reducing the risk of selective reporting and post-hoc modifications [6,62–66]. While initially planned analyses follow pre-specified predictions, additional exploratory analyses conducted after data collection are acceptable if transparently reported and justified [64,67,68]. This distinction helps readers differentiate confirmatory from exploratory findings, strengthening the reliability of the former while retaining the hypothesis-generating value of the latter [65]. Public sharing of data and code, and also detailed methodological descriptions via free and public platforms like OSF or Zenodo, should be adopted and meaningfully evaluated by journals and funding agencies [69–71].

Variation in primary study design is often necessary in animal cognition research, as experiments must be adapted to the perceptual, cognitive, and ecological characteristics of different species. This flexibility is a strength, enabling species-appropriate and ecologically meaningful insights. Still, unfortunately, it can also affect the comparability of studies, make meta-analytic synthesis more difficult, and increase uncertainty and irreproducibility [7,8,38], which has also been widely recognised in other fields using animal subjects [72]. Recent developments in Big Team Science offer promising solutions, including coordinated multi-lab studies, standardised protocols, and inclusive sampling across taxa, including understudied species [73,74]. Such collaborative efforts can enhance reproducibility, increase statistical power, and enable more rigorous, ecologically grounded cross-species comparisons. Incorporating data generated through these initiatives into future meta-analyses could substantially improve the robustness and generalisability of conclusions in animal cognition research.

Our study has two limitations that also present directions for future work. First, the relatively small number of eligible meta-analytical papers constrained the comprehensiveness of our estimation. In many cases, even after contacting authors, the data and code were unavailable, particularly when supplementary materials were insufficient for reanalysis. This highlights the need to promote open-data practices and to establish infrastructure where data sharing is standard

rather than exceptional. Second, our focus on three commonly used effect size metrics (SMD, lnRR, and Zr) resulted in the exclusion of studies employing alternative metrics. Although such cases were relatively rare (30.6%; 15 out of 49 meta-analytical studies), including a broader range of effect size measures in future studies could enhance the generalisability and applicability of meta-analytic findings.

Open-science initiatives can help guide animal cognition research and its related disciplines toward greater transparency, reliability, and cumulative progress. Meanwhile, we acknowledge that not all data can or should be shared openly. Access to behavioural videos, task protocols, or long-term observations may need restriction when they contain sensitive information (e.g., location information of threatened species) or when ethical or contractual obligations are involved. By adopting open science practices where feasible and transparently outlining restrictions, the field can enhance the credibility of individual primary and meta-analytical studies, yielding more robust insights into cognition.

Strengthening the evidential base is particularly important in ecological and evolutionary research, where cognition is increasingly recognised as an adaptive trait shaped by natural selection [4,75]. Rigorous and transparent estimation of cognitive abilities is essential for understanding how behavioural adaptations influence survival, reproduction, and species interactions, especially under rapid environmental change. In the Anthropocene, behavioural and cognitive flexibility may determine which species are most resilient to human-induced pressures [76–78]. A more robust evidence base will also enable cognitive traits to be more effectively integrated into ecological models and conservation planning, helping identify characteristics associated with resilience or vulnerability in a changing world.

5. Acknowledgements

We sincerely thank Alessandra Berry, Arie Kaffman, Chiara Musillo, Cristina Ottaviani, and Federica Amici for generously sharing their data, which was essential for this research. We also acknowledge that this work was conducted at the University of Alberta, situated on the traditional territories of the Néhiyaw (Cree), Niitsitapi (Blackfoot), Métis, Nakoda (Stoney), Dene, Haudenosaunee (Iroquois), and Anishinaabe (Ojibway/Saulteaux), in the lands of Treaties 6, 7, and 8 and the homeland of the Métis Nation.

6. AI declaration

During the preparation of this work, the author used ChatGPT-5 (OpenAI) to check grammar and spelling. After using this tool, the author carefully reviewed and edited the content as needed and takes full responsibility for the content of the published article.

7. Funding

SN and AM were supported by the Canada Excellence Research Chairs, Government of Canada (CERC-2022-00074) and ML was supported by ARC (Australian Research Council) Discovery Project grants DP210100812 and DP230101248. LG was supported by the Canada Research Chair and NSERC Discovery Grant.

Table 1. Summary of common methodological issues in animal cognition research in primary and meta-analytical studies, their potential impacts, and practical solutions. Entries are grouped by research phase: Study design, Data collection, Statistical analysis, and Reporting. The "Target" indicates whether recommendations apply to primary, meta-analytical studies, or both. The "Reference" lists supporting sources.

Phase	Target	Problem	Consequence	Recommended Solution	Reference
Study design	Primary studies	Low statistical power in primary studies	High false negative rate; increased Type M/S errors; unreliable effect estimates	Increase sample sizes where feasible; consider design strategies that improve power	[7,79,80]
Study design/ Statistical analysis	Both	High heterogeneity in task design and contexts	Inflated uncertainty; poor construct validity; limits cross-study comparability and interpretability in meta-analyses	Primary studies should justify design choices and report species and contextual details in full. Also, try collaborative or team-science approaches (e.g., multi-lab designs) to enhance reproducibility, generalizability, and equity. Meta-analyses should quantify and account for heterogeneity (e.g., via subgroup analysis or meta-regression), and can also benefit from prospective and collaborative frameworks to improve robustness and inclusivity	[7,81–83]
Statistical analysis	Meta-analytical studies	Inadequate detection or correction of publication bias	Overrepresentation of significant results; inflated effect sizes; false confidence	Apply bias correction methods appropriate to data structure and heterogeneity (e.g., robust estimation approaches)	[61]
Reporting	Primary studies	Selective reporting of individuals (e.g., excluding task failures) and outcomes (e.g., omitting non-significant results)	Inflated estimates of cognitive ability; reduced generalisability; lower statistical power; biased scientific record; increased publication bias	Study registration - report all tested individuals, including those who failed tasks; transparently report all outcomes, including non-significant results	[84]

Reporting	Primary studies	Lack of standardised reporting practices	Reduced reproducibility; difficult study quality assessment; hindered meta-analytic synthesis	Follow structured reporting guidelines (e.g., ARRIVE); clearly document procedures, task designs, subject characteristics, and analyses	[85–87]
Reporting	Both	Incomplete or inaccessible data/code	Limits reproducibility and secondary use (e.g., in meta-analyses)	Share data and code in open repositories (e.g., OSF, Zenodo); ensure materials are sufficient for reproducibility; External validation (peer-review)	[69–71,88]

337

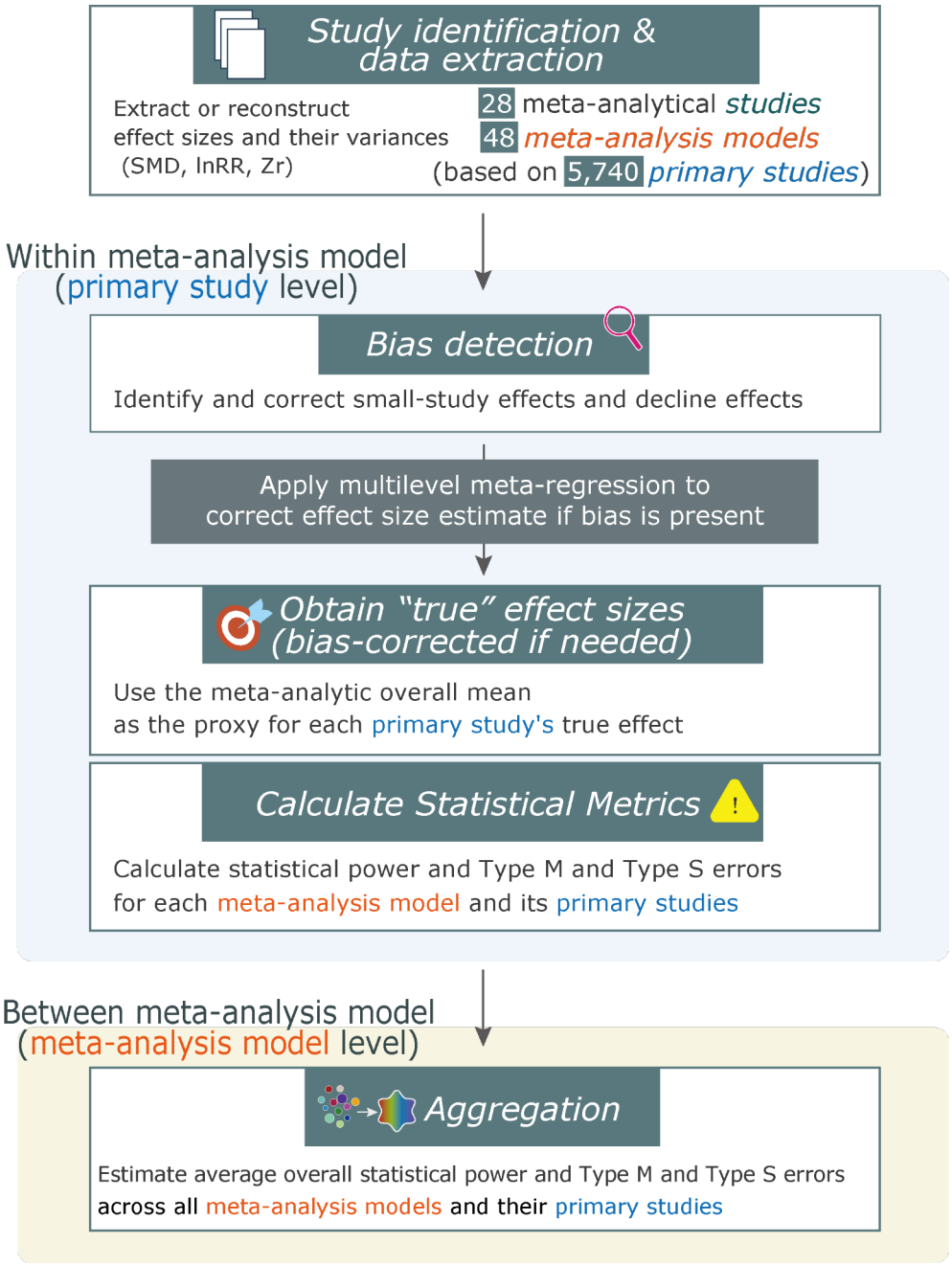

Figure 1. Conceptual workflow diagram. Effect sizes and their variances were extracted or reconstructed from 48 meta-analyses encompassing 5,740 primary studies. Publication bias was assessed within each meta-analysis by detecting small-study and decline effects and, when necessary, correcting them using multilevel meta-regression. The resulting meta-analytic means served as proxies for the "true" effect sizes at both the primary-study and meta-analysis levels. Statistical power (probability of detecting a true effect), Type M error (effect-size exaggeration), and Type S error (probability of obtaining the wrong sign) were then calculated at both levels and aggregated across meta-analyses to yield overall estimates of inferential reliability.

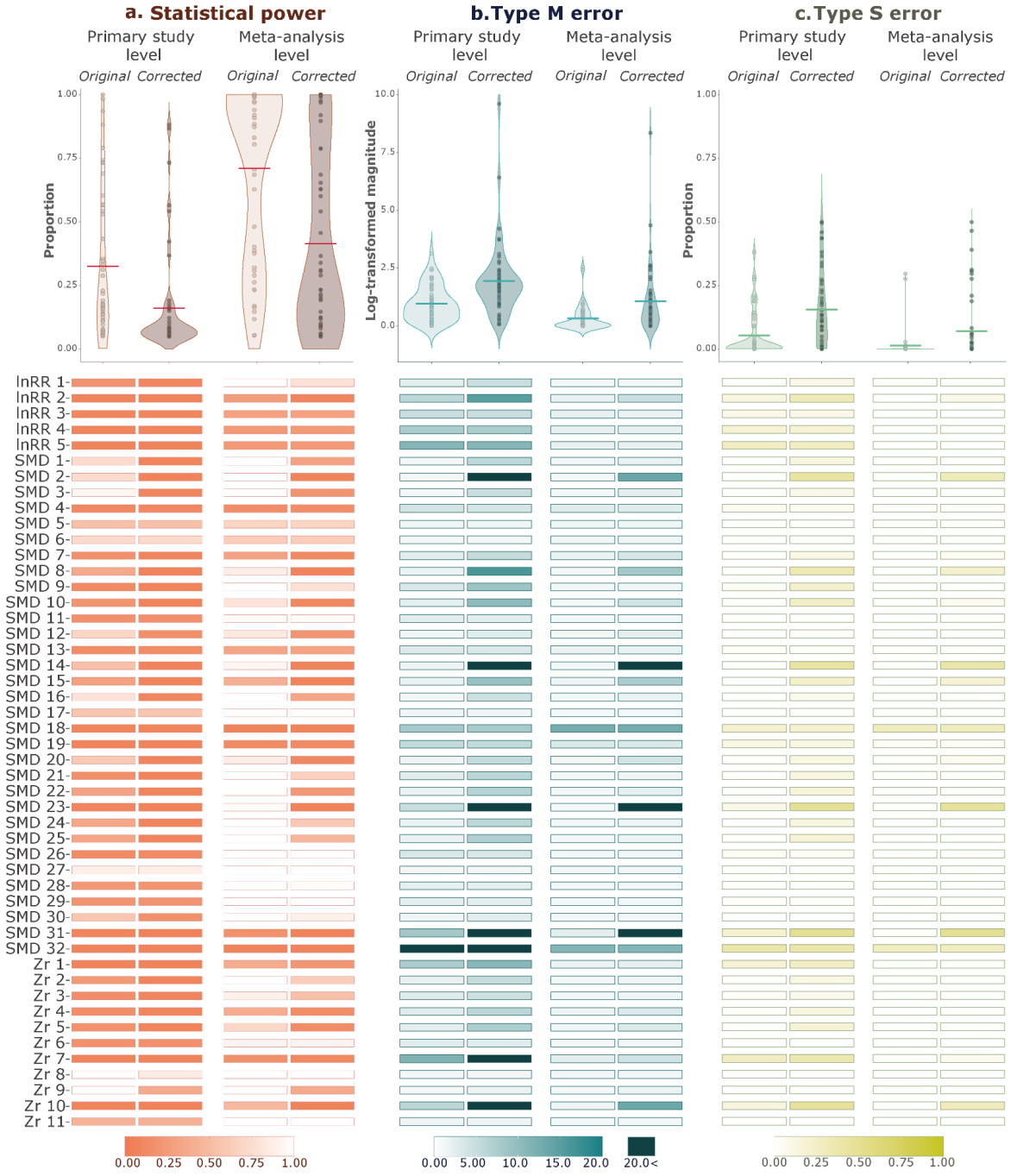

Figure 2. Statistical performance of animal cognition meta-analyses at the primary-study and meta-analysis levels. Each row represents a distinct meta-analysis model, grouped by effect-size metric: lnRR, SMD, or Zr. For each meta-analysis model, metrics are shown for the original (uncorrected) and bias-corrected effect-size estimates. (a) Statistical power: Median power to detect the (true or bias-corrected) meta-analytic effect size; darker red bars indicate lower power. (b) Type M error: Median exaggeration ratio between estimated and true effect sizes; darker blue indicates stronger overestimation (values > 20 shown as "20+"). (c) Type S error: Median

probability of obtaining a significant effect in the wrong direction; darker yellow-green indicates higher Type S error. Violin plots above each panel summarise distributions across meta-analyses. Each point represents one meta-analysis (matching a row in the bar plots), and the horizontal line denotes the overall mean.

References

1. Mizuno A, Lagisz M, Pollo P, Guillette L, Soma M, Nakagawa S. 2025 Systematic mapping and bibliometric analysis of meta-analyses on animal cognition. *Neurosci. Biobehav. Rev.* **177**, 106342.

2. Shettleworth SJ. 2010 *Cognition, Evolution, and Behavior*. Oxford University Press.

3. MacLean EL *et al.* 2012 How does cognition evolve? Phylogenetic comparative psychology. *Anim. Cogn.* **15**, 223–238.

4. Healy SD, Bacon IE, Haggis O, Harris AP, Kelley LA. 2009 Explanations for variation in cognitive ability: Behavioural ecology meets comparative cognition. *Behav. Processes* **80**, 288–294.

5. Farrar BG, Ostojić L, Clayton NS. 2021 The hidden side of animal cognition research: Scientists' attitudes toward bias, replicability and scientific practice. *PLoS One* **16**, e0256607.

6. Howard SR, Barron AB. 2024 Understanding the limits to animal cognition. *Curr. Biol.* **34**, R294–R300.

7. Farrar BG, Boeckle M, Clayton NS. 2020 Replications in comparative cognition: What should we expect and how can we improve? *Anim. Behav. Cogn.* **7**, 1–22.

8. Farrar BG, Voudouris K, Clayton NS. 2021 Replications, Comparisons, Sampling and the Problem of Representativeness in Animal Cognition Research. *Animal Behavior and Cognition* **8**, 273–295.

9. Gurevitch J, Koricheva J, Nakagawa S, Stewart G. 2018 Meta-analysis and the science of research synthesis. *Nature* **555**, 175–182.

10. Stanley TD, Carter EC, Doucouliagos H. 2018 What meta-analyses reveal about the replicability of psychological research. *Psychol. Bull.* **144**, 1325–1346.

11. Bartoš F, Maier M, Shanks DR, Stanley TD, Sladekova M, Wagenmakers E-J. 2023 Meta-analyses in psychology often overestimate evidence for and size of effects. *R. Soc. Open Sci.* **10**, 230224.

12. Lemoine NP, Hoffman A, Felton AJ, Baur L, Chaves F, Gray J, Yu Q, Smith MD. 2016 Underappreciated problems of low replication in ecological field studies. *Ecology* **97**, 2554–2561.

13. Yang Y, Hillebrand H, Lagisz M, Cleasby I, Nakagawa S. 2022 Low statistical power and overestimated anthropogenic impacts, exacerbated by publication bias, dominate field studies in global change biology. *Glob. Chang. Biol.* **28**, 969–989.

14. Yang Y, Sánchez-Tójar A, O'Dea RE, Noble DWA, Koricheva J, Jennions MD, Parker TH, Lagisz M, Nakagawa S. 2023 Publication bias impacts on effect size, statistical power, and magnitude (Type M) and sign (Type S) errors in ecology and evolutionary biology. *BMC Biol.* **21**, 71.

15. Gelman A, Tuerlinckx F. 2000 Type S error rates for classical and Bayesian single and multiple comparison procedures. *Comput. Stat.* **15**, 373–390.

16. Gelman A, Carlin J. 2014 Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspect. Psychol. Sci.* **9**, 641–651.

17. Mizuno A, Lagisz M, Pollo P, Guillette L, Soma M, Nakagawa S. 2024 Synthesis of meta-analyses on animal cognition. See https://osf.io/8hnkb.

18. Nakagawa S, Cuthill IC. 2007 Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biol. Rev. Camb. Philos. Soc.* **82**, 591–605.

19. Lagisz M, Zidar J, Nakagawa S, Neville V, Sorato E, Paul ES, Bateson M, Mendl M, Løvlie H. 2020 Optimism, pessimism and judgement bias in animals: A systematic review and meta-analysis. *Neurosci. Biobehav. Rev.* **118**, 3–17.

20. Abbott KN, Arnott CK, Westbrook RF, Tran DMD. 2019 The effect of high fat, high sugar, and combined high fat-high sugar diets on spatial learning and memory in rodents: A meta-analysis. *Neurosci. Biobehav. Rev.* **107**, 399–421.

21. Jonasson Z. 2005 Meta-analysis of sex differences in rodent models of learning and memory: a review of behavioral and biological data. *Neurosci. Biobehav. Rev.* **28**, 811–825.

22. Kredlow MA, Unger LD, Otto MW. 2016 Harnessing reconsolidation to weaken fear and appetitive memories: A meta-analysis of post-retrieval extinction effects. *Psychol. Bull.* **142**, 314–336.

23. Kavaliotis E, Boardman JM, Clark JW, Ogeil RP, Verdejo-García A, Drummond SPA. 2023 The relationship between sleep and appetitive conditioning: A systematic review and meta-analysis. *Neurosci. Biobehav. Rev.* **144**, 105001.

24. Diao H, Li Y, Sun W, Zhang J, Wang M, Chen Y, Zhou F, Li X. 2023 REM sleep deprivation induced by the modified multi-platform method has detrimental effects on memory: A systematic review and meta-analysis. *Behav. Brain Res.* **454**, 114652.

25. Schettino M, Tarmati V, Castellano P, Gigli V, Carnevali L, Cabib S, Ottaviani C, Orsini C. 2024 Effects of acute stress on reward processing: A comprehensive meta-analysis of rodent and human studies. *Neurobiol. Stress* **31**, 100647.

26. Lambert CT, Guillette LM. 2021 The impact of environmental and social factors on learning abilities: a meta-analysis. *Biol. Rev. Camb. Philos. Soc.* **96**, 2871–2889.

27. Camacho-Alpízar A, Guillette LM. 2023 From whom do animals learn? A meta-analysis on model-based social learning. *Psychon. Bull. Rev.* **30**, 863–881.

28. Moreira PS, Almeida PR, Leite-Almeida H, Sousa N, Costa P. 2016 Impact of chronic stress protocols in learning and memory in rodents: Systematic review and meta-analysis. *PLoS One* **11**, e0163245.

29. Bonapersona V, Kentrop J, Van Lissa CJ, van der Veen R, Joëls M, Sarabdjitsingh RA. 2019 The behavioral phenotype of early life adversity: A 3-level meta-analysis of rodent studies. *Neurosci. Biobehav. Rev.* **102**, 299–307.

30. Bustamante J, Soto M, Miguez G, Quezada-Scholz VE, Angulo R, Laborda MA. 2023 Extinction in multiple contexts reduces the return of extinguished responses: A multilevel meta-analysis. *Learn. Behav.* , 1–15.

31. Ou-Yang B, Hu Y, Fei X-Y, Cheng S-T, Hang Y, Yang C, Cheng L. 2022 A meta-analytic study of the effects of early maternal separation on cognitive flexibility in rodent offspring. *Dev. Cogn. Neurosci.* **56**, 101126.

32. Rocha M, Wang D, Avila-Quintero V, Bloch MH, Kaffman A. 2021 Deficits in hippocampal-dependent memory across different rodent models of early life stress: systematic review and meta-

analysis. *Transl. Psychiatry* **11**, 231.

33. Musillo C, Borgi M, Saul N, Möller S, Luyten W, Berry A, Cirulli F. 2021 Natural products improve healthspan in aged mice and rats: A systematic review and meta-analysis. *Neurosci. Biobehav. Rev.* **121**, 89–105.

34. Moran NP, Sánchez-Tójar A, Schielzeth H, Reinhold K. 2021 Poor nutritional condition promotes high-risk behaviours: a systematic review and meta-analysis. *Biol. Rev. Camb. Philos. Soc.* **96**, 269–288.

35. Szabo B, Whiting MJ, Noble DWA. 2019 Sex-dependent discrimination learning in lizards: A meta-analysis. *Behav. Processes* **164**, 10–16.

36. Macartney EL, Lagisz M, Nakagawa S. 2022 The relative benefits of environmental enrichment on learning and memory are greater when stressed: A meta-analysis of interactions in rodents. *Neurosci. Biobehav. Rev.* **135**, 104554.

37. Poirier M-A, Kozlovsky DY, Morand-Ferron J, Careau V. 2020 How general is cognitive ability in non-human animals? A meta-analytical and multi-level reanalysis approach. *Proc. Biol. Sci.* **287**, 20201853.

38. Keeble L, Wallenberg JC, Price EE. 2022 The evolution of coordination: a phylogenetic meta-analysis and systematic review. *R Soc Open Sci* **9**, 201728.

39. Amici F, Widdig A, Lehmann J, Majolo B. 2019 A meta-analysis of interindividual differences in innovation. *Anim. Behav.* **155**, 257–268.

40. Dougherty LR, Guillette LM. 2018 Linking personality and cognition: a meta-analysis. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **373**. (doi:10.1098/rstb.2017.0282)

41. Roquet RF, Monfils M-H. 2018 Does exercise augment operant and Pavlovian extinction: A meta-analysis. *J. Psychiatr. Res.* **96**, 73–93.

42. Song Y, Du Y, An Y, Zheng J, Lu Y. 2022 A systematic review and meta-analysis of cognitive and behavioral tests in rodents treated with different doses of D-ribose. *Front. Aging Neurosci.* **14**, 1036315.

43. Menting MD *et al.* 2019 The link between maternal obesity and offspring neurobehavior: A systematic review of animal experiments. *Neurosci. Biobehav. Rev.* **98**, 107–121.

44. Speechley EM, Ashton BJ, Foo YZ, Simmons LW, Ridley AR. 2024 Meta-analyses reveal support for the Social Intelligence Hypothesis. *Biol. Rev. Camb. Philos. Soc.* **99**, 1889–1908.

45. Michael A. Woodley of Menie, Peñaherrera-Aguirre M, Sarraf MA. 2022 Signs of a Flynn effect in rodents? Secular differentiation of the manifold of general cognitive ability in laboratory mice (Mus musculus) and Norwegian rats (Rattus norvegicus) over a century—Results from two cross-temporal meta-analyses. *Intelligence* **95**, 101700.

46. Siviter H, Koricheva J, Brown MJF, Leadbeater E. 2018 Quantifying the impact of pesticides on learning and memory in bees. *J. Appl. Ecol.* **55**, 2812–2821.

47. Bates D, Mächler M, Bolker B, Walker S. 2015 Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Softw.* **67**, 1–48.

48. Viechtbauer W. 2010 Conducting Meta-Analyses in R with the metafor Package. *J. Stat. Softw.* **36**, 1–48.

49. Wickham H. 2011 Ggplot2: Ggplot2. *Wiley Interdiscip. Rev. Comput. Stat.* **3**, 180–185.

50. R Core Team. 2024 R: A language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria.

51. Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, Munafò MR. 2013 Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14**, 365–376.

52. Lamberink HJ, Otte WM, Sinke MRT, Lakens D, Glasziou PP, Tijdink JK, Vinkers CH. 2018 Statistical power of clinical trials increased while effect size remained stable: an empirical analysis of 136,212 clinical trials between 1975 and 2014. *J. Clin. Epidemiol.* **102**, 123–128.

53. Forstmeier W, Wagenmakers E-J, Parker TH. 2017 Detecting and avoiding likely false-positive findings - a practical guide: Avoiding false-positive findings. *Biol. Rev. Camb. Philos. Soc.* **92**, 1941–1968.

54. Farrar BG, Altschul DM, Fischer J, van der Mescht J, Placì S, Troisi CA, Vernouillet A, Clayton NS, Ostojić L. 2020 Trialling Meta-Research in Comparative Cognition: Claims and Statistical Inference in Animal Physical Cognition. *Anim Behav Cogn* **7**, 419–444.

55. Schubiger MN, Fichtel C, Burkart JM. 2020 Validity of cognitive tests for non-human animals: Pitfalls and prospects. *Front. Psychol.* **11**, 1835.

56. Nuijten MB, van Assen MALM, Veldkamp CLS, Wicherts JM. 2015 The replication paradox: Combining studies can decrease accuracy of effect size estimates. *Rev. Gen. Psychol.* **19**, 172–182.

57. Halpern SD, Karlawish JHT, Berlin JA. 2002 The continuing unethical conduct of underpowered clinical trials. *JAMA* **288**, 358–362.

58. Ioannidis JPA. 2005 Why most published research findings are false. *PLoS Med.* **2**, e124.

59. Sena ES, van der Worp HB, Bath PMW, Howells DW, Macleod MR. 2010 Publication bias in reports of animal stroke studies leads to major overstatement of efficacy. *PLoS Biol.* **8**, e1000344.

60. Nakagawa S, Lagisz M, Jennions MD, Koricheva J, Noble DWA, Parker TH, Sánchez-Tójar A, Yang Y, O'Dea RE. 2022 Methods for testing publication bias in ecological and evolutionary meta-analyses. *Methods Ecol. Evol.* **13**, 4–21.

61. Yang Y, Lagisz M, Williams C, Noble DWA, Pan J, Nakagawa S. 2024 Robust point and variance estimation for meta-analyses with selective reporting and dependent effect sizes. *Methods Ecol. Evol.* **n/a**. (doi:10.1111/2041-210X.14377)

62. Hardwicke TE, Wagenmakers E-J. 2023 Reducing bias, increasing transparency and calibrating confidence with preregistration. *Nat. Hum. Behav.* **7**, 15–26.

63. Beran MJ. 2018 Replication and pre-registration in comparative psychology. *Int. J. Comp. Psychol.* **31**. (doi:10.46867/ijcp.2018.31.01.09)

64. Simmons J, Nelson L, Simonsohn U. 2021 Pre-registration: Why and how. *J. Consum. Psychol.* **31**,

151–162.

65. Nosek BA, Ebersole CR, DeHaven AC, Mellor DT. 2018 The preregistration revolution. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 2600–2606.

66. Yamada Y. 2018 How to crack pre-registration: Toward transparent and open science. *Front. Psychol.* **9**, 1831.

67. Willroth EC, Atherton OE. 2024 Best laid plans: A guide to reporting preregistration deviations. *Adv. Methods Pract. Psychol. Sci.* **7**. (doi:10.1177/25152459231213802)

68. Lakens D. 2024 When and how to deviate from a preregistration. *Collabra Psychol.* **10**, 117094.

69. Gomes DGE *et al.* 2022 Why don't we share data and code? Perceived barriers and benefits to public archiving practices. *Proc. Biol. Sci.* **289**, 20221113.

70. Culina A, van den Berg I, Evans S, Sánchez-Tójar A. 2020 Low availability of code in ecology: A call for urgent action. *PLoS Biol.* **18**, e3000763.

71. Klein O *et al.* 2018 A practical guide for transparency in psychological science. *Collabra Psychol.* **4**, 20.

72. Vesterinen HM, Sena ES, Egan KJ, Hirst TC, Churolov L, Currie GL, Antonic A, Howells DW, Macleod MR. 2014 Meta-analysis of data from animal studies: a practical guide. *J. Neurosci. Methods* **221**, 92–102.

73. Alessandroni N *et al.* 2025 Challenges and promises of big team comparative cognition. *Nat. Hum. Behav.* **9**, 240–242.

74. Alessandroni N *et al.* 2024 Comparative Cognition Needs Big Team Science: How Large-Scale Collaborations Will Unlock the Future of the Field. *Comparative Cognition & Behavior Reviews* **19**, 67–72.

75. Henke-von der Malsburg J, Kappeler PM, Fichtel C. 2020 Linking ecology and cognition: does ecological specialisation predict cognitive test performance? *Behav. Ecol. Sociobiol.* **74**, 1–26.

76. Mazza V, Šlipogor V. 2024 Behavioral flexibility and novel environments: integrating current perspectives for future directions. *Curr. Zool.* **70**, 304–309.

77. Ducatez S, Sol D, Sayol F, Lefebvre L. 2020 Behavioural plasticity is associated with reduced extinction risk in birds. *Nat. Ecol. Evol.* **4**, 788–793.

78. Hockings KJ *et al.* 2015 Apes in the Anthropocene: flexibility and survival. *Trends Ecol. Evol.* **30**, 215–222.

79. Stevens JR. 2017 Replicability and reproducibility in comparative psychology. *Front. Psychol.* **8**, 862.

80. Farrar BG, Krupenye C, Motes-Rodrigo A, Tennie C, Fischer J, Altschul DM, Ostojić L. 2022 Replication and Reproducibility in Primate Cognition Research. In *Primate Cognitive Studies*, pp. 532–550. Cambridge University Press.

81. Nakagawa S, Lagisz M, Yang Y, Drobniak SM. 2024 Finding the right power balance: Better study

545 design and collaboration can reduce dependence on statistical power. *PLoS Biol.* **22**, e3002423.

546 82. Yang Y, Macleod M, Pan J, Lagisz M, Nakagawa S. 2023 Advanced methods and implementations
547 for the meta-analyses of animal models: Current practices and future recommendations. *Neurosci.*
548 *Biobehav. Rev.* **146**, 105016.

549 83. Nakagawa S, Yang Y, Macartney EL, Spake R, Lagisz M. 2023 Quantitative evidence synthesis: a
550 practical guide on meta-analysis, meta-regression, and publication bias tests for environmental
551 sciences. *Environmental Evidence* **12**, 1–19.

552 84. Farrar BG *et al.* 2023 Reporting and interpreting non-significant results in animal cognition research.
553 *PeerJ* **11**, e14963.

554 85. Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG. 2010 Improving bioscience research
555 reporting: The ARRIVE guidelines for reporting animal research. *J. Pharmacol. Pharmacother.* **1**,
556 94–99.

557 86. Percie du Sert N *et al.* 2020 Reporting animal research: Explanation and elaboration for the ARRIVE
558 guidelines 2.0. *PLoS Biol.* **18**, e3000411.

559 87. Andersen MS, Kofoed MS, Paludan-Müller AS, Pedersen CB, Mathiesen T, Mawrin C, Olsen BB,
560 Halle B, Poulsen FR. 2024 CRIME-Q-a unifying tool for critical appraisal of methodological
561 (technical) quality, quality of reporting and risk of bias in animal research. *BMC Med. Res.*
562 *Methodol.* **24**, 306.

563 88. Grant S *et al.* 2025 TOP 2025: An update to the Transparency and Openness Promotion Guidelines.
564 *BITSS*. (doi:10.31222/osf.io/nmfs6_v2)

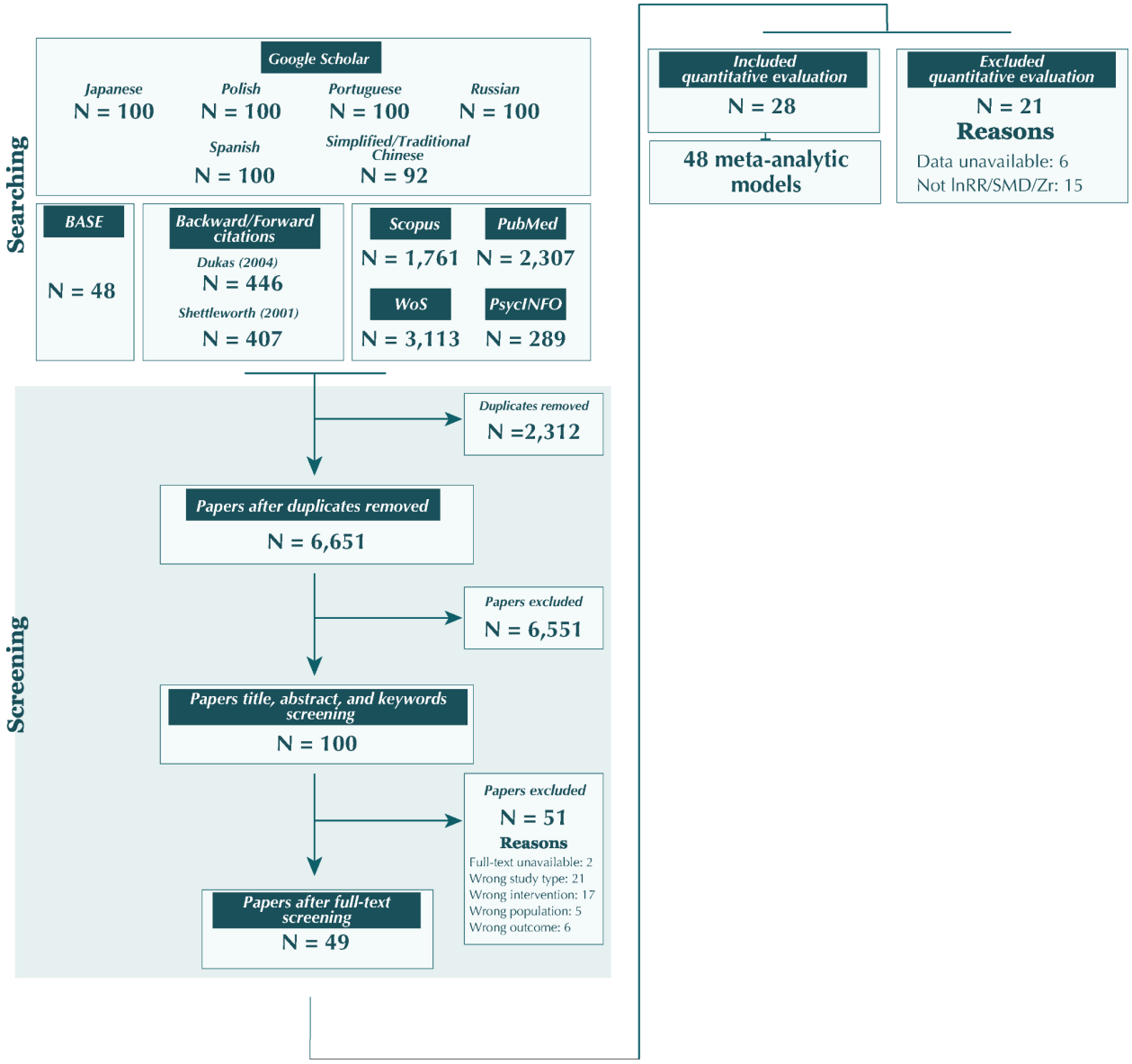

Figure S1. ROSE-like flow chart. This figure follows a format proposed by the Reporting Standards for Systematic Evidence Syntheses (ROSES), providing a structured and transparent framework for reporting systematic reviews and maps in environmental and related disciplines. Although no formal visualisation standard currently exists for systematic mapping in the field of animal cognition, we adapted the ROSES format to improve clarity and reproducibility. The flowchart illustrates the full literature selection process, beginning with an extensive search across multiple sources, including bibliographic databases, grey literature (BASE), backward and forward citation tracking, and non-English records retrieved via Google Scholar. A total of 8,963 records were identified. After removing 2,312 duplicates, 6,651 records remained for title, abstract, and keyword screening. Of these, 100 articles were selected for full-text screening, and 49 were retained for the final quantitative synthesis (Mizuno et al., 2025). Our project focused on

28 out of 49 meta-analytical studies identified in the synthesis, which were eligible for quantitative evaluation. These 28 studies comprised a total of 48 meta-analytic models.

Table S1. The model estimates of (a) Statistical power, (b) Type M error, and (c) Type S error, based on both uncorrected (original - β_0[overall]) and bias-corrected (β_0[bias-corrected]) effect sizes. All estimates are reported at two levels: the primary study level and the meta-analysis level. We used mixed-effects models and weighted regression models to compute the median and 95% confidence intervals. All values were back-transformed from the model estimates. For statistical power and type S error, values below 0 or above 1 were constrained to 0 and 1 (denoted as 0* or 1*). *k* = number of effect sizes; *N* = number of primary studies.

	Level	Effect size types	True effects	Median	95% CI		Mean	*k*	*N*
					lower	upper			
Statistical power	Primary study level	All	β_0[overall]	0.17	0.16	0.18	0.23	5740	1415
			β_0[bias-corrected]	0.09	0.09	0.09	0.12	5740	1415
		lnRR	β_0[overall]	0.19	0.16	0.21	0.29	1602	224
			β_0[bias-corrected]	0.11	0.10	0.12	0.17	1602	224
		SMD	β_0[overall]	0.17	0.16	0.18	0.21	2835	945
			β_0[bias-corrected]	0.08	0.08	0.09	0.11	2835	945
		Zr	β_0[overall]	0.18	0.16	0.20	0.23	1303	246
			β_0[bias-corrected]	0.11	0.10	0.12	0.14	1303	246
	meta-analysis level	All	β_0[overall]	0.82	0.72	0.94	1.14	5740	1415
			β_0[bias-corrected]	0.45	0.34	0.60	0.62	5740	1415
		lnRR	β_0[overall]	0.77	0.36	1*	0.93	1602	224
			β_0[bias-corrected]	0.62	0.30	1*	0.75	1602	224
		SMD	β_0[overall]	0.89	0.76	1*	1.31	2835	945
			β_0[bias-corrected]	0.41	0.27	0.63	0.61	2835	945
		Zr	β_0[overall]	0.75	0.56	1*	0.89	1303	246
			β_0[bias-corrected]	0.36	0.20	0.67	0.43	1303	246
Type M error	Primary study level	All	β_0[overall]	2.86	2.76	2.95	3.48	5740	1415
			β_0[bias-corrected]	7.79	7.13	8.51	9.49	5740	1415
		lnRR	β_0[overall]	3.02	2.69	3.39	4.28	1602	224

			β_0[bias-corrected]	4.99	4.45	5.59	7.07	1602	224
		SMD	β_0[overall]	2.79	2.69	2.89	3.15	2835	945
			β_0[bias-corrected]	9.16	8.13	10.32	10.35	2835	945
		Zr	β_0[overall]	2.87	2.62	3.15	3.40	1303	246
			β_0[bias-corrected]	5.76	5.05	6.56	6.81	1303	246
	meta-analysis level	All	β_0[overall]	1.11	1.03	1.20	1.30	5740	1415
			β_0[bias-corrected]	2.03	1.37	3.02	5.84	5740	1415
		lnRR	β_0[overall]	1.14	0.78	1.66	1.20	1602	224
			β_0[bias-corrected]	1.28	0.87	1.87	1.46	1602	224
		SMD	β_0[overall]	1.07	0.96	1.19	1.32	2835	945
			β_0[bias-corrected]	2.80	1.46	5.40	11.79	2835	945
		Zr	β_0[overall]	1.15	1.00	1.33	1.21	1303	246
			β_0[bias-corrected]	1.80	1.17	2.79	2.47	1303	246
Type S error	Primary study level	All	β_0[overall]	0.0269	0.0250	0.0289	0.0433	5740	1415
			β_0[bias-corrected]	0.0985	0.0925	0.1049	0.1378	5740	1415
		lnRR	β_0[overall]	0.0330	0.0264	0.0404	0.0660	1602	224
			β_0[bias-corrected]	0.0760	0.0640	0.0897	0.1336	1602	224
		SMD	β_0[overall]	0.0251	0.0231	0.0273	0.0368	2835	945
			β_0[bias-corrected]	0.1049	0.0973	0.1129	0.1351	2835	945
		Zr	β_0[overall]	0.0276	0.0223	0.0336	0.0487	1303	246
			β_0[bias-corrected]	0.0912	0.0777	0.1065	0.1378	1303	246
	meta-analysis level	All	β_0[overall]	0.00058	0*	0.00194	0.0043	5740	1415
			β_0[bias-corrected]	0.01215	0.00345	0.02352	0.0370	5740	1415
		lnRR	β_0[overall]	0.00044	0*	0.002	0.00048	1602	224
			β_0[bias-corrected]	0.0011	0*	0.009	0.0066	1602	224
		SMD	β_0[overall]	0.00086	0*	0.00332	0.0065	2835	945

			β_0[bias-corrected]	0.02211	0.00574	0.04719	0.0636	2835	945
		Zr	β_0[overall]	0.00016	0*	0.00078	0.00030	1303	246
			β_0[bias-corrected]	0.00922	0*	0.02731	0.0224	1303	246