A new effect size for meta-analysis of magnitude: lnM

Shinichi Nakagawa[1,2*], Ayumi Mizuno[1], Coralie Williams[2,3], Malgorzata Lagisz[1,2], Rose E. O'Dea[2], Daniel W. A. Noble[3], Alistair M. Senior[4,5,6], Erick Lundgren[1], and Santiago Ortega[1]

[1]Collaboration for Open Science and Synthesis in Ecology and Evolution, Department of Biological Sciences, Faculty of Science, University of Alberta, Edmonton, Canada
[2]Evolution and Ecology Research Centre, School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, New South Wales, Australia
[3]School of Mathematics and Statistics, The University of New South Wales, Sydney, New South Wales, Australia
[4]School of Life and Environmental Sciences, The University of Sydney, Sydney, New South Wales, Australia
[5]Charles Perkins Centre, The University of Sydney, Sydney, New South Wales, Australia
[6]Sydney Precision Data Science Centre, The University of Sydney, Sydney, New South Wales, Australia
[*]snakagaw@ualberta.ca

February 22, 2026

Author Statement

Conceptualization: SN, REO. Data curation: SN, Formal analysis: SN, REO. Funding Acquisition: SN, Investigation: SN. Methodology: SN. Project administration: SN, SO. Supervision: SN. Validation: SO, EL, DWAN. Visualization: AM, SN. Writing - Original Draft: SN. Writing - Review & Editing: All authors.

Conflict of Interest

The authors declare no conflict of interest

Type of article: Method
Word count: 160 (abstract) and 4996 (main text)
Number of references: 41
Number of figures and tables: 5 figures, 1 tables and 0 box

Data Availability Statement

All data, scripts and relevant files used for this study can be found at the GitHub repository (link) and a version of it will be archived at Zenodo (link) once it is accepted for publication.

Abstract

Meta-analyses in ecology and evolution often consider the magnitude of differences between groups rather than their direction. Yet, a common practice is to coerce signed effects (e.g., d and response ratio) into magnitudes by taking absolute values. This transformation induces strong upward bias and non-normal (Gaussian) sampling distributions, violating the assumptions of standard meta-analytic models. Here we introduce lnM, a log-ratio effect size for the magnitude of difference between two groups, defined from standard one-way ANOVA components. Unlike absolute-value approaches, the proposed lnM is asymptotically normal, and can be analysed, using standard multilevel meta-analysis and meta-regression with both categorical and continuous moderators. We combine theory, simulations, and worked examples to compare lnM with absolute-value approaches. We show when the delta-method and parametric single-fit bootstrap estimators for lnM perform well, and how one may assess publication bias. The lnM effect size provides a direction-free, meta-regression-friendly measure of magnitude that is applicable to both ratio- and interval-scale traits, offering a practical solution for synthesising the magnitude of ecological and evolutionary effects and beyond.

Keywords— meta-regression, biological optima, coefficient-of-variation ratio, location-scale models, Bayesian statistics

1 Introduction

Meta-analysis has progressed from a niche statistical exercise to a cornerstone of synthesis across the medical, social, and biological sciences (Arnqvist and Wooster, 1995; Gurevitch et al., 2018). Standardised (unit-free) effect sizes that compare two groups, such as standardised mean differences (SMD; its estimators, known as Cohen's d or Hedges' g) and log response ratios (lnRR), now underpin thousands of quantitative reviews (Hedges, 1981; Hedges et al., 1999). In contrast, newer variance-based metrics, such as the log variability ratio (lnVR) and the log coefficient of variation ratio (lnCVR), have extended this toolkit to address differences in within-group variation (Nakagawa et al., 2015). Despite these advancements, we still lack a widely adopted effect size specifically designed to isolate the magnitude of a difference; that is, how far apart two groups are, irrespective of direction (cf. Morrissey, 2016b; Kulinskaya and Hoaglin, 2023). Magnitude is often more biologically meaningful than the sign of an effect, particularly when researchers investigate deviations from biological optima or population divergence. In such scenarios, the absolute size of the change matters more than the direction (e.g. Kingsolver et al., 2001) (reviewed in Morrissey, 2016b). However, ecologists and evolutionary biologists currently lack a rigorous method to pool absolute effects across studies.

In practice, investigators often circumvent this gap by using the absolute value of signed effects (i.e., $|\text{lnRR}|$ or $|\text{SMD}|$). However, taking absolute values folds sampling error at zero and can bias estimates upward. Specifically, Morrissey (2016b) showed that taking absolute values could inflate pooled estimates because the nonlinear transformation of noisy effect-size estimates is subject to Jensen's inequality: the greater the sampling variance, the larger the upward bias. Consequently, meta-analytic summaries based on these absolute values overestimate average magnitudes, leading to misleading conclusions about, for example, the strength of selection, ecological impacts, or treatment efficacy (Morrissey, 2016b,a). To mitigate this, Morrissey (2016b) advocated an "analyse-then-transform" strategy, which first fits models to the signed effect and then transforms the result. While this method can recover unbiased overall means, it is difficult to extend to flexible meta-regression with multiple moderators and has not been widely adopted.

Recently, Kulinskaya and Hoaglin (2023) developed a principled framework for unsigned mean-based effects. They clarified the distributional basis of absolute mean differences and absolute SMDs ($|d|$), relating $|d|$ to folded-t and d^2 to noncentral F distributions, and provided methods for estimation and interval construction under random-effects meta-analytic models. While this work represents an important advance over naive absolute transformations, a major practical limitation remains for applied ecologists. Because unsigned metrics (e.g., $|d|$) are bounded at zero and have non-Gaussian, skewed sampling distributions (e.g., folded normals), they are challenging to integrate into standard modelling frameworks. This distributional constraint hinders the use of standard linear meta-regressions, which is particularly problematic for continuous moderators.

To resolve the above-mentioned problems, we propose an effect size that is often more biologically meaningful, the "magnitude of separation" between groups relative to within-group noise, irrespective of direction. We introduce the log-magnitude effect size, lnM, which target the magnitude of mean separation relative to within-group variation on a log scale. Importantly, lnM is compatible with standard meta-analytic models (cf. Cinar et al., 2022; Williams et al., 2025)and it supports meta-regression with multiple categorical or continuous moderators. By operating on the log scale, lnM avoids the zero-boundary problem, enabling inference on a (nearly) Gaussian, unbounded scale, that remains interpretable for both ratio and interval data. For cases in which lnM cannot be defined analytically (a key limitation described below), we show that a single-fit parametric bootstrap provides reliable point estimates and sampling variances (Nakagawa et al., 2025b). Further, we describe how to interpret lnM, whose 0 (when between-group separation and with-group variation are the same) is approximately equivalent to $d = 1.4$; this equivalence may feel counterintuitive and surprising, as a d of 1.4 is very large (further discussed in Section 2.6).

The remainder of the paper proceeds as follows. First, we connect historical motivations for comparing magnitudes to Lynch's variance-ratio framework and recent meta-analysis of variation (Lynch, 1990; Nakagawa et al., 2015; Senior et al., 2020). We then derive lnM and its sampling variance, validate its performance via simulation, and describe its interpretation and connection to d (SMD). Finally, we re-analyse two published datasets of ecological examples to illustrate how lnM delivers regression-friendly magnitude estimates, contrasting to absolute-value approaches. An online tutorial illustrates `R` (link) implementation using `metafor` (Viechtbauer, 2010), and helper functions in `orchaRd` (Nakagawa et al., 2021, 2023).

2 Developing a New Effect Size: lnM

2.1 Inspiration

Interest in the distance between two ecological or evolutionary groups, regardless of the sign of the difference, has a long history in evolutionary biology (Haldane, 1949). Notably, Lynch (1990) proposed the ratio of between– to within–population variance as a scale-free yardstick for morphological divergence under a neutral model of evolution. More recently, the log-variability ratio (lnVR) and its derivative, along with the log-coefficient-of-variation ratio (lnCVR), were proposed to meta-analyse differences in variability (sample standard deviations) between two groups (Nakagawa et al., 2015; Senior et al., 2020). Drawing on these foundations, we develop the log-magnitude statistic, lnM. This proposed effect size maintains an intuitive link to Lynch's variance-ratio, which is clearly inspired by classical ANOVA, but with key methodological refinements. First, by working on the log scale, lnM converts the ratio of variance components into an asymptotically normally distributed quantity. Second, unlike Lynch´s ratio (which compares variances), lnM compares standard deviations (the square root of variance), which is more comparable to lnVR, the ratio of two standard deviations. Also, whereas Lynch's metric requires natural-log-transformed (ln) sample means and variances, lnM is calculated using statistics on the original scale (see below).

2.2 Definition

We define lnM as an effect size for the *magnitude* of a two-group difference (how far apart two groups are), irrespective of direction. Let the observed summary statistics be ($\bar{X}_1$, s_1, n_1) for group 1 and ($\bar{X}_2$, s_2, n_2) for group 2; these are the same basic inputs required for SMD and lnRR. For a two-group comparison, the classical one-way ANOVA identities provide convenient building blocks. The between- and within-group mean squares are

$$MS_B = \frac{n_1 n_2}{n_1 + n_2} \, (\bar{X}_1 - \bar{X}_2)^2, \tag{1}$$

$$MS_W = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}. \tag{2}$$

We define the within-group variance as $s_W^2 = MS_W$. To obtain a *between-component* on the original measurement scale, we subtract the within-group mean square and rescale by the harmonic-mean sample size $n_0 = 2n_1n_2/(n_1+n_2)$, defining (Lynch, 1990):

$$s_B^2 \equiv \frac{MS_B - MS_W}{n_0} = \frac{MS_B - s_W^2}{n_0}, \tag{3}$$

where s_B^2 is a function of the squared mean difference (i.e., the dispersion/separation between two point means around the grand mean).

We then define lnM as the log ratio of the between-component (mean-separation) standard deviation and within-group standard deviation:

$$\text{lnM} \equiv \ln\left(\frac{s_B}{s_W}\right). \tag{4}$$

We now apply the first-order Taylor expansion, also known as the delta method (Ver Hoef, 2012), to derive the sampling variance for both independent and dependent cases (full derivations in the Appendix). The sampling variance for two independent groups is as follows:

$$\text{Var}_{\text{IND}}(\text{lnM}) = \frac{1}{4\Delta^2}\left[\left(\tfrac{n_0}{2}\right)^2 \left(2\,s_D^4 + 4\,s_D^2\,\delta^2\right) + \frac{MS_B^2}{MS_W^2}\,\frac{2\,MS_W^2}{n_1 + n_2 - 2}\right], \tag{5}$$

where $\Delta = MS_B - MS_W$ represents the difference between mean squares, $n_0 = 2n_1n_2/(n_1+n_2)$ is the harmonic-mean sample size, $\delta = \bar{X}_1 - \bar{X}_2$ is the observed difference in group means and $s_D^2 = s_1^2/n_1 + s_2^2/n_2$ is the sampling variance of δ. The symbols MS_B and MS_W are the sample between- and within-group mean squares defined in the preceding equations.

Similarly, the sampling variance for paired (dependent) groups is:

$$\mathrm{Var}_{\mathrm{DEP}}(\mathrm{lnM}) = \frac{1}{4\Delta^2}\left[\left(\frac{n}{2}\right)^2\left(\frac{2\,s_D^4}{n^2} + \frac{4\,\delta^2\,s_D^2}{n}\right) + \frac{MS_B^2}{MS_W^2}\,\frac{s_1^4 + s_2^4 + 2\,r^2 s_1^2 s_2^2}{2\,(n-1)}\right], \tag{6}$$

where n is the number of paired observations (i.e., $n_1 = n_2 = n$), r is the within-pair correlation, $s_D^2 = s_1^2 + s_2^2 - 2\,r\,s_1 s_2$ is the variance of the within-pair difference, and $\delta = \bar{X}_1 - \bar{X}_2$ is the corresponding mean difference (this correlation can be obtained from primay studies or often assumped to be $r = 0.5$ or 0.8) (Noble et al., 2017; Pustejovsky and Tipton, 2022). The quantity Δ again denotes $MS_B - MS_W$; for the paired design $MS_B = (n/2)(\bar{X}_1 - \bar{X}_2)^2$ and $MS_W = (s_1^2 + s_2^2)/2$.

By examining these formulas, we identify two practical features of lnM. First, the precision of lnM depends strongly on how far the between-group mean square MS_B sits above the within-group mean square MS_W. As the delta-method expressions show, Var(lnM) scales approximately as $(MS_B - MS_W)^{-2}$: when MS_B is distinct from MS_W, standard errors shrink rapidly; however, as MS_B approaches MS_W ($MS_B \approx MS_W$) the variance becomes large, the consequences of which are discussed later. Second, lnM becomes undefined when $MS_B \leq MS_W$. For example, when $MS_B = MS_W$, the estimate of between-group variance is zero, causing lnM to approach $-\infty$. In practice, such cases are common: whenever group means are similar, a non-trivial fraction of contrasts will fall into this undefined or numerically unstable region. Discarding these cases would systematically remove studies with small separations, a limitation that likely restricted the adoption of Lynch's rate-of-evolution index (Lynch, 1990), which relies on a similar variance-ratio construction. In the next section, we address both issues by introducing a simulation-based approach that stabilises lnM, providing finite estimates and standard errors even when $MS_B \leq MS_W$.

2.3 Extension

While the closed-form variance formulas derived above are elegant and easy to use, they encounter limitations in two critical scenarios. First, when sample sizes are small, the first–order Taylor approximation can be imprecise (e.g., Senior et al., 2020). Second, as noted earlier, when the difference $\Delta = MS_B - MS_W$ is non-positive, $\ln M$ becomes undefined. A practical remedy is the single–fit parametric bootstrap, originally introduced by Mandel (2013) and Fletcher and Jowett (2022). Such a single-fit bootstrap is significantly faster and easier than traditional bootstrap methods (Tibshirani and Efron, 1993; Davison and Hinkley, 1997). More recently, Nakagawa et al. (2025b) integrated this method into SAFE; Single-fit, Accurate, Fast, and Easy bootstrap to obtain bias-corrected point estimates and sampling variances for effect-size statistics. SAFE replaces complex algebra with four vectorised steps that usually run in milliseconds: (1) fit once, (2) draw once, (3) transform once, and (4) summarise twice.

The method leverages standard distributional theory: we assume sample means $(\bar{X}_1, \bar{X}_2)$ are normally distributed, while sample variances follow a χ^2 (chi-square) distribution (s_1^2, s_2^2) (Anderson, 2003). The χ^2 distribution arises whenever we take the sum of squared normal variables; in fact, $(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$ whenever data are independently and normally distributed with true variance σ^2. For multivariate settings, we employ the Wishart distribution, the multivariate analogue of the χ^2 distribution. If we collect p correlated normal variables, their sample covariance matrix follows a Wishart distribution, denoted $W_p(\Sigma, df)$, where the true covariance matrix Σ plays the role of σ^2, and the degrees of freedom df depend on the sample size. These distributional facts guarantee that simulated variance draws are always positive and respect the independence between means and variances under normal sampling. Using SAFE we can then estimate lnM and its sampling variance under these distributional assumptions as follows.

Step 1: Fit once. We begin by taking the observed pair of sample statistics for which we wish to calculate lnM $(\bar{X}_1, \bar{X}_2, s_1^2, s_2^2)$ and sample sizes (n_1, n_2) (and correlation r for paired data). For independent groups, we assume the sample mean and variance for the two groups are distributed as:

$$\bar{X}_g^* \sim \mathcal{N}\!\left(\bar{X}_g,\ \frac{s_g^2}{n_g}\right), \qquad \frac{(n_g - 1)\,S_g^{2*}}{s_g^2} \sim \chi^2_{n_g - 1}, \quad g = 1, 2, \tag{7}$$

where $\bar{X}_g^*$ is independent of S_g^{2*}. Note that drawing from these distributions (Step 2, below) produces four vectors of replicates, which we refer as 'clouds' (boostrapped vectors): $(\bar{X}_1^*, \bar{X}_2^*, s_1^{2*}, s_2^{2*})^\top$ (i.e., quadruplets of clouds).

For paired groups ($n_1 = n_2 = n$), we define the covariance matrix:

$$\Sigma = \begin{bmatrix} s_1^2 & r s_1 s_2 \\ r s_1 s_2 & s_2^2 \end{bmatrix}. \tag{8}$$

The joint distributions for both means and variances are then:

$$(\bar{X}_1^*, \bar{X}_2^*)^\top \sim \mathcal{N}\left((\bar{X}_1, \bar{X}_2)^\top, \frac{\Sigma}{n}\right), \qquad (n-1)\, S^* \sim W_2(\Sigma, n-1), \tag{9}$$

independently, where S^* is a 2×2 covariance matrix. In practice, we only use its diagonal entries S_{11}^* and S_{22}^* to form MS_W.

Step 2: Draw once. We next simulate a large Monte Carlo batch, say $B = 10,000$ clouds, from the normal–χ^2/Wishart distributions, as above. This produces surrogate means and variances that mimic repeated sampling from the underlying populations. In practice, this step can be done for independent cases with the R functions `rnorm` and `rchisq`, while dependent cases require `mvrnorm` from the `MASS` package (Venables and Ripley, 2013), and the base function `rWishart`.

Step 3: Transform once. We convert every surrogate draw into a value of lnM^* by the same algebra that defines Equations (1)-(4). Crucially, we retain only those replicates where $MS_B^* > MS_W^*$.

Step 4: Summarise twice. 'Twice' means we obtain two quantities: a bias-corrected (BC) point estimate and standard error (the square root of sampling variance). Having obtained a bootstrap sample $\{\mathrm{lnM}^{*(i)}\}_{i=1}^B$, these can be then estimated by:

$$\mathrm{SE}_{\mathrm{SAFE}}(\mathrm{lnM}) = \sqrt{\mathrm{Var}_{\mathrm{SAFE}}(\mathrm{lnM})} = \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}\left(\mathrm{lnM}^{*(i)} - \overline{\mathrm{lnM}^*}\right)^2}, \tag{10}$$

$$\mathrm{bias} \;=\; \overline{\mathrm{lnM}^*} \;-\; \mathrm{lnM}_{\mathrm{PI}}, \tag{11}$$

$$\mathrm{lnM}_{\mathrm{BC}} = \mathrm{lnM}_{\mathrm{PI}} - \mathrm{bias} = 2\mathrm{lnM}_{\mathrm{PI}} - \overline{\mathrm{lnM}^*}(= \mathrm{lnM}_{\mathrm{SAFE}}), \tag{12}$$

where $\overline{\mathrm{lnM}^*}$ is the average of bootstrap values (i.e., $\overline{\mathrm{lnM}^*} = \frac{1}{B}\sum_{b=1}^{B} \mathrm{lnM}^{*(b)}$) and $\mathrm{lnM}_{\mathrm{PI}}$ is a plug-in estimate obtained from equation (4).

Bias in the plug-in lnM estimate arises because lnM is a nonlinear function of the sample means and variances, so $\mathbb{E}[\mathrm{lnM}_{\mathrm{PI}}] \neq \mathrm{lnM}_{\mathrm{true}}$ in finite samples; in other words, the plug-in value of lnM is biased in relation to the true value of lnM. The bootstrap cloud $\{\mathrm{lnM}^{*(i)}\}$ directly approximates the true sampling distribution of lnM, and its mean minus the plug-in value provides an empirical estimate of that systematic error. Subtracting this estimated bias yields the bias-corrected point estimate $\mathrm{lnM}_{\mathrm{BC}}$, which restores accuracy even when samples are small or in cases where the delta approximation is poor (see the online supplement for the full implementation of SAFE in `R`). Furthermore, because SAFE samples directly from the finite-sample distribution of the four summary statistics $(\bar{X}_1, \bar{X}_2, s_1^2, s_2^2)$, its standard error and bias correction remain valid when n is small, when the delta approximation is poor, or when Δ is close to zero.

However, the issue we mentioned earlier is that $\mathrm{lnM}_{\mathrm{PI}}$ is not always definable (Equation (4)). In such a case, we define lnM as the mean of the valid bootstrap replicates:

$$\mathrm{lnM}_{\mathrm{SAFE}} = \overline{\mathrm{lnM}^*} \quad \text{if} \quad MS_B \leq MS_W \tag{13}$$

Therefore, even in the problematic scenario where $MS_B \leq MS_W$ (for which the analytic lnM is undefined), SAFE still returns a usable estimate and standard error (SE), provided that at least some bootstrap replicates satisfy $MS^*_B > MS^*_W$. Although the accuracy of $\mathrm{lnM}_{\mathrm{SAFE}}$ as in equation (13) cannot be validated, we expect these values to be smaller than all lnM values that can be estimated without the bootstrap. We now use simulation to compare the performance of the SAFE approach with that of the closed-form formulas, including the behavior of $\mathrm{lnM}_{\mathrm{SAFE}}$ values (i.e., Equation (13)).

2.4 Simulation design

We simulated samples for two groups and estimated the sample mean and standard deviation for group 1 and group 2 and correlation between two groups: $(\bar{X}_1,\ s_1, \bar{X}_2,\ s_2, r)$. In our simulation, we denote true (population) values of these statistics as $(\mu_1, \sigma_1, \mu_2, \sigma_2, \rho)$ respectively. We closely followed the style of a simulation study used by Senior et al. (2020) who studied the performance of the delta-method-based formulas for lnRR, lnVR, and lnCVR (cf. Lajeunesse, 2015); we also followed Williams et al. (2024) for designing and reporting, where posssible.

For every run, we drew two vectors of length n_1 and n_2 (sample size for group 1 and group 2) from a bivariate normal distribution

$$\begin{bmatrix} X_{1i} \\ X_{2i} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho\,\sigma_1\sigma_2 \\ \rho\,\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \right), \qquad i = 1, \ldots, \max(n_1, n_2), \tag{14}$$

where $\rho = 0$ gives independent groups and $\rho = 0.8$ produces paired data for dependent groups. The population ('true') means were set to $\mu_1 = 0$ and $\mu_2 = \theta$, with $\theta \in \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 1.2, 1.4, 1.6, 1.8, 2, 2.5, 3, 4, 5\}$ to span from negligible to large magnitudes. Throughout, the within-group standard deviations were equal ($\sigma_1 = \sigma_2 = 1$), ensuring that changes in lnM stem from the distance between means rather than from differences in dispersion (within-group standard deviation); examining Equation (2) tells us differences in standard deviation between two groups do not have a major impact on lnM as the two sample standard deviations are weighted by their sample sizes. Note that θ^2 is equivalent to s_B^2 and $(\sigma_1 + \sigma_2)/2$ is s_W^2 where $\mathrm{lnM} = \ln(s_B/s_W)$.

For independent groups, balanced and unbalanced designs were examined with (n_1, n_2) chosen from {(5,5), (10,10), (20,20), (100,100), (3,7), (6,14), (12,28), (40,160)}. This was explored as unbalanced designs may impact lnM *via* changes in s_B (cf., Equation (1)). For dependent groups, we only selected balanced designs. Therefore, we have 144 parameter sets (combinations) for independent groups, while 72 sets for dependent groups (in total, 216 combinations).

For every replicate ($K = 10^5$ per parameter set; 216 sets), we computed: 1) the plug-in point estimates, either $\widehat{\mathrm{lnM}}_{\mathrm{PI}}$ and its first–order delta–method variance $\widehat{\mathrm{Var}}_{\mathrm{IND}}(\mathrm{lnM})$ Equation (5) (independent groups) or $\widehat{\mathrm{Var}}_{\mathrm{DEP}}(\mathrm{lnM})$ Equation (6), collectively these are called $\widehat{\mathrm{Var}}_{\mathrm{delta}}(\mathrm{lnM})$ (dependent groups; note that we added 'hat' symbols to these to make it clear that these are estimators) and 2) the SAFE bootstrap cloud $\{\mathrm{lnM}^{*(b)}\}_{b=1}^{B}$ with $B = 10^5$ draws (note that we collected 10^5 bootstrap draws regardless of the rate of discarded draws), together with its bias-corrected mean $\widehat{\mathrm{lnM}}_{\mathrm{BC}}$ and the bootstrap variance $\widehat{\mathrm{Var}}_{\mathrm{SAFE}}(\mathrm{lnM})$. For each estimator $\widehat{\mathrm{lnM}}_l \in \{\widehat{\mathrm{lnM}}_{\mathrm{PI}}, \widehat{\mathrm{lnM}}_{\mathrm{BC}}\}$ we calculated the empirical bias:

$$\mathrm{bias}(\widehat{\mathrm{lnM}}_l) = \overline{\widehat{\mathrm{lnM}}}_l - \mathrm{lnM}_{\mathrm{true}} \tag{15}$$

$$\overline{\widehat{\mathrm{lnM}}}_l = \frac{1}{K} \sum_{k=1}^{K} \widehat{\mathrm{lnM}}_l^{(k)}. \tag{16}$$

where $\mathrm{lnM}_{\mathrm{true}}$ is defined by the population parameter values in Equation (1)-(4) and k denotes the k-th simulation ($k = 1, 2, ...K$; $K = 10^5$). We can interpret this bias as the deviation of the estimator (lnM_l) from the true population value. For each variance estimator $\widehat{\mathrm{Var}}_l \in \{\widehat{\mathrm{Var}}_{\mathrm{delta}},\ \widehat{\mathrm{Var}}_{\mathrm{SAFE}}\}$ we calculated the relative bias as follows:

$$\text{relative bias}(\widehat{\mathrm{Var}}_l) = \frac{\overline{\widehat{\mathrm{Var}}}_l - \mathrm{Var}_{\mathrm{MC}}\,(\widehat{\mathrm{lnM}})}{\mathrm{Var}_{\mathrm{MC}}\,(\widehat{\mathrm{lnM}})} \times 100, \tag{17}$$

$$\overline{\widehat{\mathrm{Var}}_l} = \frac{1}{K}\sum_{k=1}^{K} \widehat{\mathrm{Var}}_l^{(k)}, \tag{18}$$

$$\mathrm{Var}_{\mathrm{MC}}\left(\widehat{\ln\mathrm{M}_l}\right) = \frac{1}{K-1}\sum_{k=1}^{K}\left(\widehat{\ln\mathrm{M}_l}^{(k)} - \overline{\widehat{\ln\mathrm{M}_l}}\right)^2. \tag{19}$$

$\mathrm{Var}_{\mathrm{MC}}\left(\widehat{\ln\mathrm{M}_l}\right)$ is a variance estimate based on all Monte Carlo simulation runs ($K = 10^5$) for a combination of parameters, and $\widehat{\ln\mathrm{M}_l} \in \{\widehat{\ln\mathrm{M}}_{\mathrm{PI}}, \widehat{\ln\mathrm{M}}_{\mathrm{BC}}\}$. Yet, we only used $\widehat{\ln\mathrm{M}_l} = \widehat{\ln\mathrm{M}}_{\mathrm{BC}}$ as mentioned, $\widehat{\ln\mathrm{M}}_{\mathrm{PI}}$ was often undifiened so not estimable and hence its variance was not reliable. This relative bias was assumed to represent the relative deviation of the estimator $(\widehat{\mathrm{Var}}_l)$ from the true population value (note that 100 means that the estimator is twice the true value). We also obtained the coverage of nominal 95% Wald intervals (i.e., using the 0.975th quantile of the standard normal z distribution). Importantly, our pilot simulations showed that the delta-method variance could blow up when θ and n are small, by as much as $10^5 - 10^{10}$ (see the online supplements). Therefore, we set an upper limit of 20 for the delta-method variance (this was never an issue for the SAFE variance, which could reach around 10, but never exceeded 20; the frequencies of how often it exceeded 20 were recorded).

2.5 Simulation results

Across all scenarios, the point estimators displayed the same qualitative pattern. When the true difference between the group means was trivial ($\theta < 1$) and samples were very small ($n_i \leq 10$) both methods (the plug-in and SAFE) usually overestimated the true $\ln M$; for example, with $n_1 = n_2 = 5$ (indepedent) and $\theta = 0.7$ the upward bias approached one log-unit (Fig. 1) although SAFE underestimated in some ranges of θ. The bias fell rapidly with either increasing θ or increasing n, becoming negligible ($|\text{bias}| < 0.05$) once $\theta \geq 1$ or the minimum sample size contained at least twenty observations. SAFE reduced the bias by around 23.5% relative to the plug-in (PI) estimator, an advantage that was most pronounced in the extreme lower-left corners of each panel in Fig. 1. Paired designs exhibited the same shape, but with markedly smaller absolute bias because the within-pair correlation provides additional information: at $\theta = 0.5$ and $n = 5$, the bias was approximately half that of the corresponding independent case design.

More striking differences emerged for the estimated sampling variances (Fig. 2). For very small samples with $\theta < 1$, the delta–method variance was seriously inflated, exceeding the Monte-Carlo variance by approximately 300% (note this is likely to be a conservative estimate of the inflation as we capped the delta-method variance to be the maximum of 20 for computation reasons in our simulation). This overestimation decayed smoothly and was practically absent for $n_i \geq 20$ or $\theta \geq 1.5$. By contrast, the SAFE variance was less biased over the whole grid with a tendency to under-estimate (up to by $\sim 80\%$) when many bootstrap draws were discarded because $MS_B^* \leq MS_W^*$ (note that we always collected 10^5 bootstrap draws regardless of the rate of discarded draws; also see supplemental information for the results of 95% coverage, root mean squared errors [RMSE], and Monte Carlo standard errors [MCSE]).

The frequency with which the analytic formula became undefined (because $MS_B \leq MS_W$) ranged from 63.5% for the most unbalanced design with $n_1 = 3, n_2 = 7$ to 0% when both groups contained at least one hundred observations (Table S1). SAFE always produced an estimate, even though approximately 40–85% of its bootstrap replicates were discarded in the small-n designs (i.e., $n_1 + n_2 = 10$). Also it is notable that when sample sizes larger than 40 (i.e., $n_1 + n_2 \geq 40$), point estimates and SE are mostly reliable. Given these results, the SAFE procedure is recommended over the combination of the plug-in point estimate and the delta-method-based SE. Therefore, in our worked examples below, we primarily used the SAFE estimators.

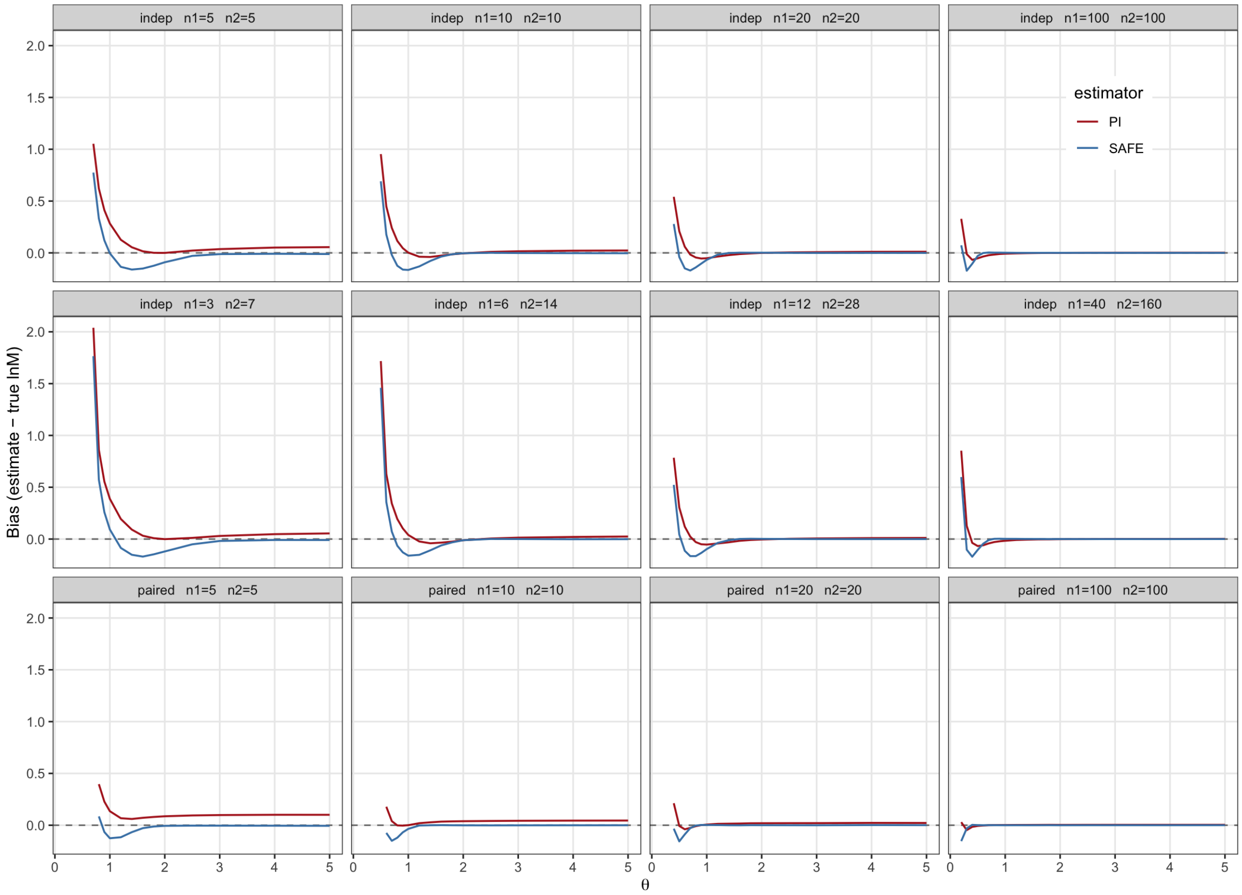

Figure 1: Bias of $\widehat{\ln M}$ (estimate minus true value) across all designs in relation to θ, which is the difference between the mean of group 1 $\mu_1(=0)$ and that of group 2 $\mu_2(=\theta)$. Solid red lines: plug-in (PI) estimator (solid blue lines): SAFE estimator (solid red lines). Horizontal dashed line denotes zero bias; `indep` is an independent-group design while `paired` is a paired-group design with sample sizes of two groups denoted as n_1 and n_2. The y-axis is shown on the natural logarithm (ln) scale. Note some values for small sample size scenarios seem to be truncated, as these values could not be defined.

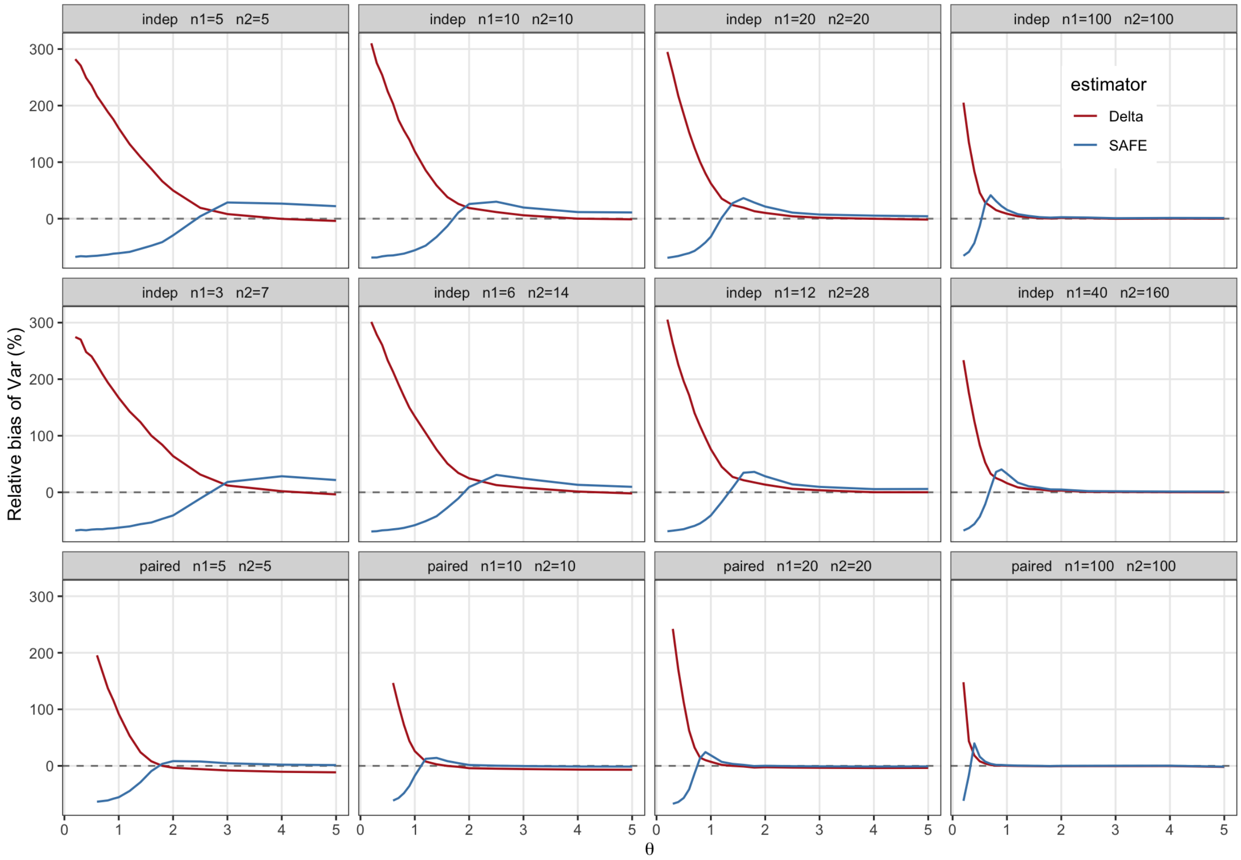

Figure 2: Relative bias of the estimated sampling variance of $\widehat{\ln\mathrm{M}}$ (the delta-method estimator in red and the SAFE estimator in blue) in relation to θ, which is the difference between the mean of group 1 $\mu_1(=0)$ and that of group 2 $\mu_2(=\theta)$, expressed as $100\,[\frac{\widehat{\mathrm{Var}}_l - \mathrm{Var}_{\mathrm{MC}}(\widehat{\ln\mathrm{M}})}{\mathrm{Var}_{\mathrm{MC}}(\widehat{\ln\mathrm{M}})}]$; `indep` is an independent-group design while `paired` is paired-group design with sample sizes of two group denoted as n_1 and n_2. Colours and layout follow Fig. 1. A value of zero (grey dashed line) indicates an unbiased variance estimator; positive values mean over-estimation. Note some values for small sample size scenarios seem to be truncated, as these values could not be defined.

2.6 Interpretation

Before we apply lnM to example datasets, we provide some guidance on the interpretation of the new effect size. A convenient way to read lnM is through the ratio of the between-component (mean-separation) SD to the within-group SD. By definition,

$$\frac{s_B}{s_W} = e^{\mathrm{lnM}}. \tag{20}$$

Thus $\mathrm{lnM} = 0$ corresponds to $s_B = s_W$, values $\mathrm{lnM} > 0$ indicate that the between-group spread exceeds the typical within-group SD (e.g. $\mathrm{lnM} = 0.69$ implies $s_B \approx 2s_W$), and $\mathrm{lnM} < 0$ indicates that the observed separation is smaller than the within-group SD. Importantly, testing $\mathrm{lnM} = 0$ does not ask "are the group means equal?", but rather "is the between-group dispersion at least as large as the within-group SD?". A significantly negative lnM, therefore, indicates that, even if the mean difference is non-zero in a strict sense, it is small compared to within-group noise.

For interpretability, we believe that it is helpful to map lnM onto another familiar effect size that is also expressed in units of within-group SD, standardised mean difference, SMD, or d (Hedges, 1981). Conveniently, we can define SMD (termed, here, d_{eq}, indicating d-equivalent) in terms of lnM (see Appendix B for derivation):

$$d_{\mathrm{eq}} = \sqrt{2e^{2\,\mathrm{lnM}} + \frac{2}{n_0}}. \tag{21}$$

A simple large-sample approximation (when $2/n_0$ is small) is:

$$d_{\mathrm{eq}} \approx \sqrt{2}\,e^{\mathrm{lnM}} = \sqrt{2}\,\frac{s_B}{s_W}. \tag{22}$$

We treat this mapping as an interpretive aid rather than a primary analysis scale; Table 1 gives a simple rule-of-thumb correspondence between lnM, s_B/s_W, and the approximate d_{eq}. As mentioned earlier, $\mathrm{lnM} = 0$ is surprisingly large in terms of d ($d = 1.41$). Overall average of $|d|$ should, however, be almost always larger than d for the same dataset, which include both postive and negative d values, so that thinking lnM in terms of d might be misleading if not interpreted carefully (i.e., a relatively large value of d_{eq} is expected for overall means for a meta-analysis of magnitude). Notably, it is possible to centre lnM at a specific SMD value (e.g., $d = 0.2$ as 0 for adjusted lnM; see Appendix C for a full exploration). Yet we will not use the adjusted version of lnM because it is more informative to compare differences in lnM between groups (i.e. contrasts) and slopes of lnM rather than interpreting an overall lnM of a dataset, as we show in our examples below. We also note that, given this relationship between lnM and SMD, one might wonder why we do not use $\ln|d|$ as an effect size. In Appendix D, we explain the limitations of $\ln|d|$ and over preference in the use of lnM.

Table 1: Rule-of-thumb mapping from lnM to the ratio $s_B/s_W = e^{\mathrm{lnM}}$ and the approximate absolute standardized mean difference $d_{\mathrm{eq}} \approx \sqrt{2}\,e^{\mathrm{lnM}}$ (large-sample approximation); these conversions are useful for meta-analytic means, which are based on large sample sizes.

lnM	$s_B/s_W = e^{\mathrm{lnM}}$	$d_{\mathrm{eq}} \approx \sqrt{2}\,e^{\mathrm{lnM}}$
-2.0	≈ 0.14	≈ 0.19
-1.5	≈ 0.22	≈ 0.32
-1.0	≈ 0.37	≈ 0.52
-0.7	≈ 0.50	≈ 0.70
-0.4	≈ 0.67	≈ 0.95
0.0	1.00	≈ 1.41

3 Worked Examples

3.1 Meta-analytic models considered

In the two worked examples below we reanalyse published datasets whose original studies used $|d|$. In both cases we use multilevel meta-analysis and meta-regression models to account for dependence among effect sizes within studies (e.g., multiple effect-size estimates per study). Let y_{ij} denote the ith effect size from study j, with corresponding known sampling variance v_{ij}. We first fit an intercept-only three-level meta-analytic model (Nakagawa and Santos, 2012; Cheung, 2014):

$$y_{ij} = \beta_0 + u_{0j} + w_{ij} + \varepsilon_{ij}, \tag{23}$$

where β_0 is the overall mean effect, $u_{0j} \sim \mathcal{N}(0, \sigma^2_{\text{study}})$ is the random intercept at the study level, $w_{ij} \sim \mathcal{N}(0, \sigma^2_{\text{effect}})$ is an additional random effect at within-study level (i.e. effect-size level), and $\varepsilon_{ij} \sim \mathcal{N}(0, v_{ij})$ is the sampling error (although we have two levels: between- and witnin-study level, the model is referred to "three level" in the literature, considering the sampling error random effect as another level). Between-study variance is denoted by σ^2_{study} and within-study (among-effect) variance by σ^2_{effect} along with effect-size specific sampling variance v_{ij}. Notably, this type of meta-analysis is called the inverse-variance method because effect sizes are weighted by the inverse of their sampling variances (Borenstein et al., 2021). Given the simulation results, we should also run the same model only with effect sizes based on sample sizes larger than 40 (i.e., $n_1 + n_2 \geq 40$). Such a model might prevent overestimating the meta-analytic mean (cf. Ioannidis et al., 2017).

To examine how effect sizes vary with moderators, we extend Equation (23) to a three-level meta-regression model:

$$y_{ij} = \beta_0 + \sum_{p=1}^{P} \beta_p x_{p,ij} + u_{0j} + w_{ij} + \varepsilon_{ij}, \tag{24}$$

where $x_{p,ij}$ is the value of the pth moderator ($p = 1, \ldots, P$) for effect size i in study j, and β_p is the corresponding meta-regression coefficient.

To assess small-study effects (the tendency for smaller studies to report larger effects), we use conversions of the harmonic-mean sample size ($1/\sqrt{\tilde{n}_0}$ and $1/\tilde{n}_0$ when $\tilde{n}_0 = 1/2n_0$ with the harmonic-mean sample size defined as above) as a moderator, following the approach proposed by Nakagawa et al. (2022), a modified multilevel version of the original Egger's regression test (Egger et al., 1997). We then fit a small-study effect model of the form:

$$y_{ij} = \beta_0 + \beta_1 \frac{1}{\sqrt{\tilde{n}_{0ij}}} + u_{0j} + w_{ij} + \varepsilon_{ij}, \tag{25}$$

where β_1 quantifies the association between lnM and the half harmonic-mean sample size. A significantly negative or positive β_1 would indicate that smaller studies (lower n_{0ij}) tend to yield larger effect size values, consistent with small-study effects (Jennions and Møller, 2002; Sterne and Egger, 2001). Also, once we find $\beta_1 \neq 0$, we fit the model with $1/\tilde{n}_{0ij}$ instead of $\sqrt{1/\tilde{n}_{0ij}}$ and its intercept β_0 as a bias-corrected overall effect (following Nakagawa et al., 2022) (cf. Stanley and Doucouliagos, 2014; Irsova et al., 2025). Although Nakagawa et al. (2022) has recommended including all important moderators for this model to account for heterogeneity due to such moderators, we only run this uni-moderator small-study effect test in the examples below. Importantly, we should never use sampling SE in place of $\sqrt{1/\tilde{n}_{0ij}}$, because the point estimate and sampling SE are strongly correlated in lnM as discussed above.

Finally, it is important to emphasise that the example datasets are used here primarily to illustrate the use of lnM, rather than to reproduce the original authors' inferential goals. Accordingly, our model structures and choice of moderators differ from those in the original analyses (e.g., we do not include phylogenetic random effects) (cf. Cinar et al., 2022; Williams et al., 2025); therefore, the numerical results are not directly comparable with those in the original studies. Yet, we report results for both lnM and conventional effect-size measures (i.e., $|d|$) so that readers can compare lnM-based inferences with familiar metrics in each example. We used the `R` package `metafor` (Viechtbauer, 2010) for model implementation with a helper function to estimate lnM from `orchaRd` (Nakagawa et al., 2021, 2023). (the code and complete results are presented in [link]).

3.2 Example 1: the effect of poor early-life nutrition

We re-analysed 683 control-treatment pairs from Almeida et al. (2021) that investigated early nutritional impacts on organisms, computing $|d|$ and lnM for each comparison. The delta-method estimator for lnM failed for 270 entries, whereas the SAFE procedure returned finite point estimates and variances for all 683 entries, rescuing almost 40% of the dataset for variance-shift analysis. A multilevel meta-analysis of the dataset, whose outcome was on "Condition" with 336 effect sizes (Fig. 3A,B) gave a large overall mean $|d|$ ($\beta_0^{|d|} = 1.35$, 95% CI: 0.98 to 1.71 with $I^2_{total} = 95.45\%$). The overall effect lnM_{SAFE} was negative and non-significant ($\beta_0^{\text{lnM}} = -0.14$, 95% CI: -0.37 to 0.08 with $I^2_{total} = 89.43\%$). Yet, the conversion d_{eq} showed a similar result with $|d|$ ($\beta_0^{d_{\text{eq}}} = 1.22$, 95% CI: 0.98 to 1.54), as expected.

Next, we ran a meta-regression with three continuous moderators (with 265 effect sizes): 1) severity of nutritional deprivation, 2) duration of deprivation and 3) recovery period. One notable result is that for $|d|$, the slope for the recovery period was negative but not significantly so ($\beta^{|d|}_{\text{recovery}} = -0.70$, 95% CI: -1.69 to 0.28), whereas for lnM, the slope was significantly negative ($\beta^{\text{lnM}}_{\text{recovery}} = -0.84$, 95% CI: -1.61 to -0.06), implying that as the recovery period increases, the two groups become more similar; the nutritionally deprived group was able to recover (Fig. 3C,D; the other variables were statistically significant for both $|d|$ and lnM). This difference in inference is mainly due to the boundary at zero for $|d|$.

Egger-type regressions (small-study-effect text) (Fig. 3E,F) suggested a weak and non-significant small-study pattern for $|d|$ ($\beta^{|d|}_{\text{nSE}} = 1.09$, 95% CI: -0.20 to 2.37) but a stronger and significant pattern for lnM ($\beta^{\ln M}_{\text{nSE}} = 1.61$, 95% CI: 0.79 to 2.44). Therefore, we estimated a bias-corrected overall effect ($\beta_0^{\text{lnM}_{\text{SAFE}}} = -0.59$, 95% CI: -0.91 to 0.27; $\beta_0^{d_{\text{eq}}} = 0.78$, 95% CI: 0.57 to 1.08); this represents around 65% reduction in the overall mean estimation (in terms of d_{eq}). Similarly, our sensitivity analysis by restricting effect sizes based on $n_1 + n_2 \geq 40$ showed a smaller overall effect than the original overall estimate ($\beta_0^{\text{lnM}_{\text{SAFE}}} = -0.41$, 95% CI: -0.95 to 0.12; $\beta_0^{d_{\text{eq}}} = 0.94$, 95% CI: 0.55 to 1.60). Therefore, both of the above-mentioned overall means are likely to be severely overestimated.

3.3 Example 2: the impacts of anthropogenic noise

For the dataset of Kunc and Schmidt (2019), which explored the impact of anthropogenic noise on animals, with 486 entries (effect sizes); the delta-method lnM failed for 177 of these, whereas SAFE produced lnM estimates and variances for all contrasts, rescuing 177 comparisons. After restricting the dataset to birds, mammals, amphibians and fishes (379 effect sizes from 88 studies), a multilevel meta-analysis (Fig. 4A,B) showed a moderate-to-large mean shift ($\beta_0^{|d|} = 0.87$, 95% CI: 0.71 to 1.02 with $I^2_{total} = 84.99\%$); lnM showed a negative overall effect in relation to a clear separation (i.e. lnM > 0) ($\beta_0^{\text{lnM}} = -0.46$, 95% CI: -0.65 to -0.27 with $I^2_{total} = 83.57\%$). Yet, both results are consistent with each other when comparing $|d|$ and d_{eq} ($\beta_0^{d_{\text{eq}}} = 0.89$, 95% CI: 0.74 to 1.07).

Using taxon as a moderator (a categorical moderator with 4 levels; Fig. 4C,D), $|d|$-based meta-regression suggested all four taxa respond positively to noise, with mammals showing the largest mean shift ($\beta^{|d|}_{\text{mammalia}} = 1.48$, 95% CI: 1.10 to 1.87). The corresponding lnM_{SAFE} model gave negative coefficients for amphibians, birds and fishes (e.g. $\beta^{\text{lnM}}_{\text{amphibia}} = -0.74$, 95% CI: -1.20 to -0.29), but a near-zero coefficient for mammals ($\beta^{\text{lnM}}_{\text{mammalia}} = 0.29$, 95% CI: -0.14 to 0.73). Yet, these results are consistent in terms of contrasts among different taxa; for both $|d|$ and lnM, only mammals are significantly positively different from the other taxa.

Egger-type regressions again produced a weakly positive yet non-significant pattern for $|d|$ ($\beta_1^{|d|} = 0.77$, 95% CI: -0.07 to 1.61) but a strong and significant association for lnM ($\beta_1^{\text{lnM}} = 2.60$, 95% CI: 1.40 to 3.82; Fig. 4E,F). Our bias-corrected estimate suggests the "true" overall mean might be approximately 38% smaller than the original estimate ($\beta_0^{\text{lnM}} = -0.93$, 95% CI: -1.22 to -0.64; $\beta_0^{d_{\text{eq}}} = 0.56$, 95% CI: 0.42 to 0.74). Also, our sensitivity analysis for the overall effect revealed a smaller overall effect than the original overall estimate, but more similar to the bias-corrected estimate ($\beta_0^{\text{lnM}} = -0.76$, 95% CI: -1.06 to -0.46; $\beta_0^{d_{\text{eq}}} = 0.66$, 95% CI: 0.49 to 0.89). Therefore, all the meta-analytic results above are likely inflated, although contrasts (differences between taxonomic groups) would remain valid. Fortunately, using lnM allows adjustment for small-study effects (see the online supplements; see the Discussion section for more details).

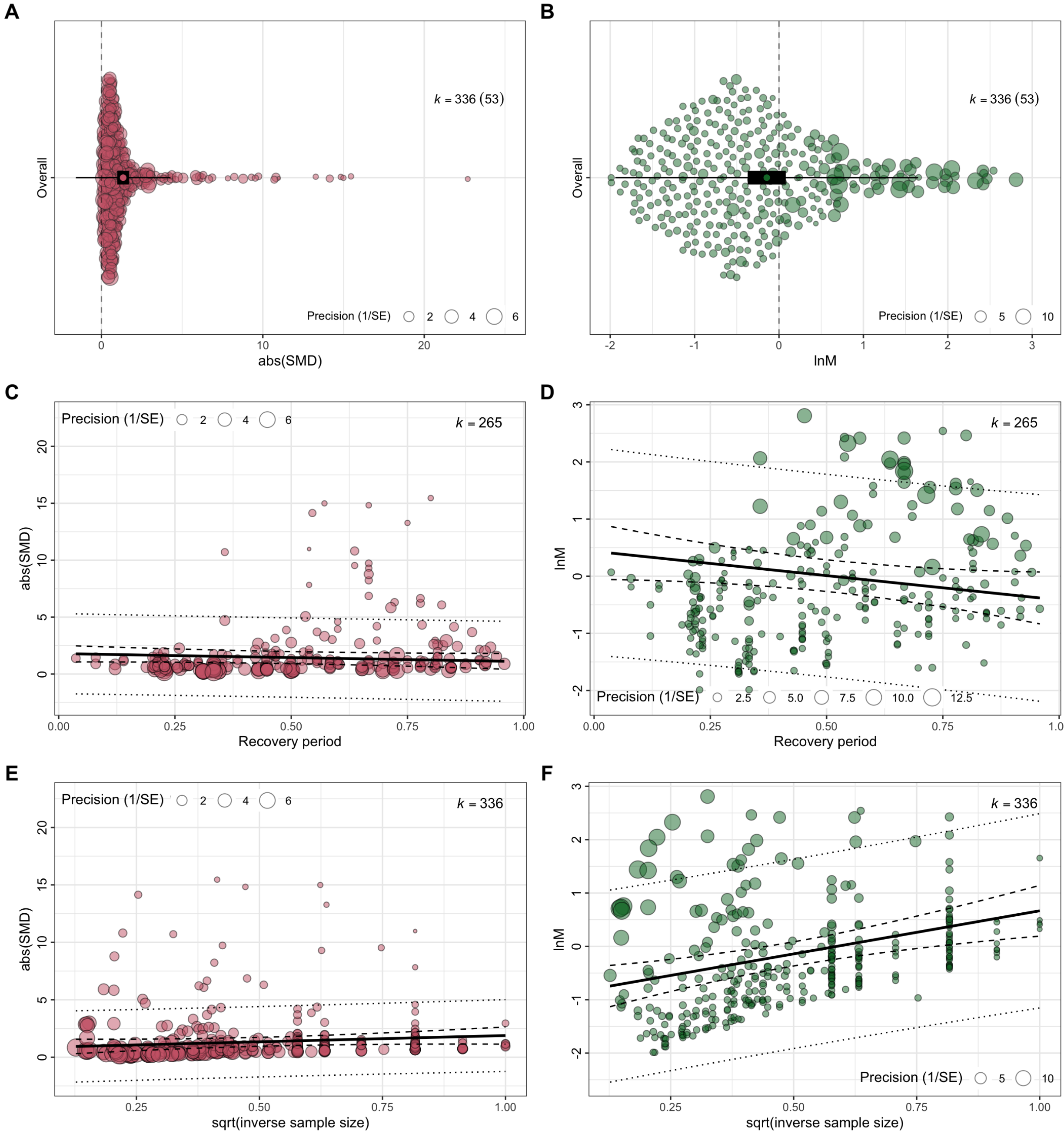

Figure 3: Example 1: the effects of poor early-life nutrition. (A,B) Orchard plots for multilevel meta-analyses of $|d|$ (= `abs(SMD)`) and $\ln M_{\mathrm{SAFE}}$ for the 336 "Condition" effect sizes. Coloured points are study-level effects, with circle area proportional to precision (1/SE); black circles give meta-analytic means, thick horizontal lines show 95% confidence intervals, and thin dotted lines show 95% prediction intervals. (C,D) Bubble plots for meta-regressions on Recovery days: point size is proportional to precision, solid lines show fitted slopes, shaded bands show 95% confidence intervals, and outer dotted lines show 95% prediction intervals. (E,F) Egger-type regressions of $|d|$ and lnM on $\sqrt{1/\tilde{n}_{0ij}}$ (= `sqrt(inverse sample size)`), using the same plotting conventions for regression line, confidence intervals and prediction intervals; k is the number of effect sizes while the number in parentheses are the number of studies.

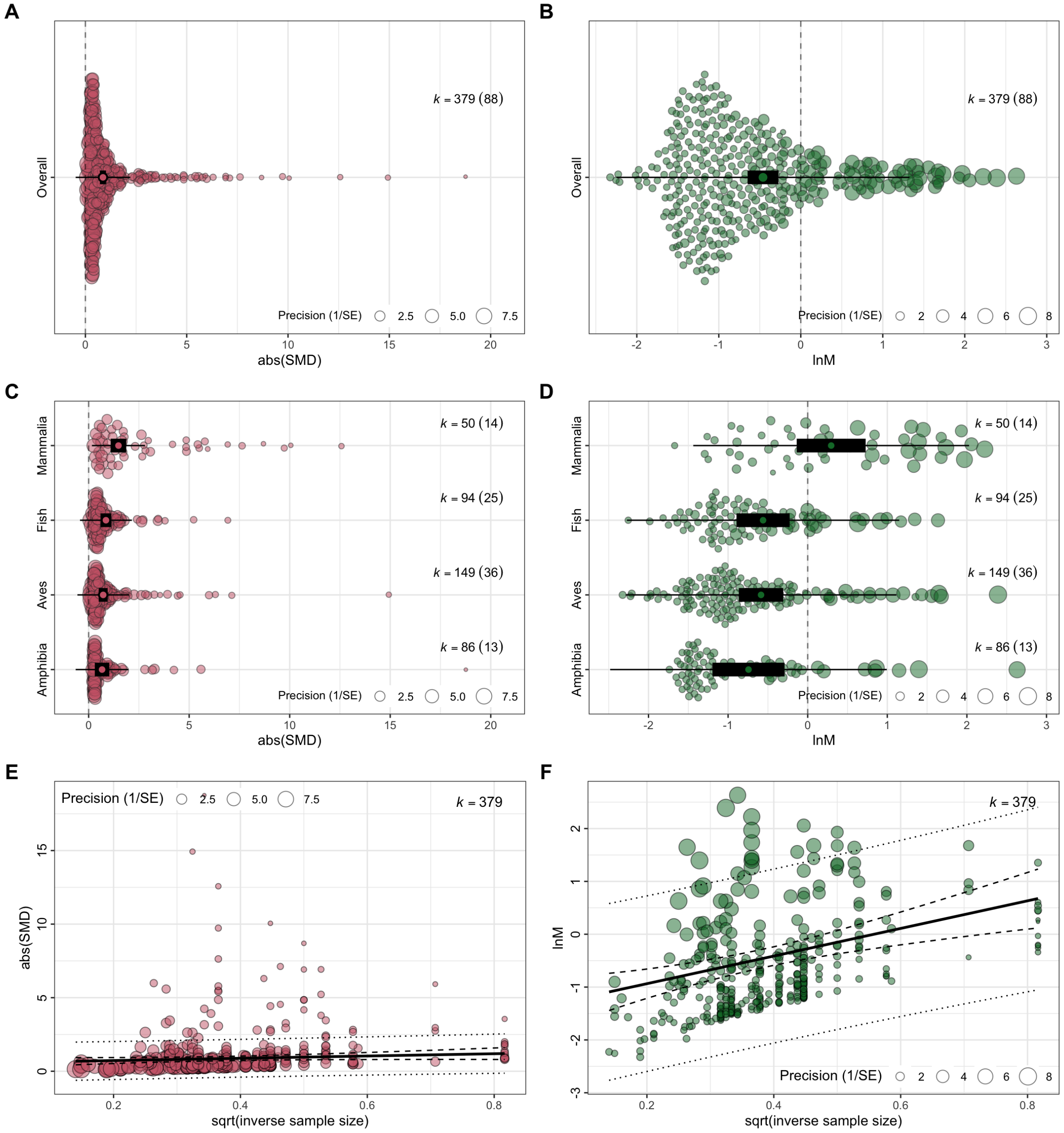

Figure 4: Example 2: the impacts of anthropogenic noise (A,B) Orchard plots for multilevel meta-analyses of $|d|$ (= `abs(SMD)`) and $\ln M_{\text{SAFE}}$ across 379 effect sizes from 88 studies; symbols and lines are as in Fig. 3A,B, with black circles giving meta-analytic means, thick lines 95% confidence intervals, and thin dotted lines 95% prediction intervals. (C,D) Orchard plots from taxon-specific meta-regressions, showing estimated $|d|$ and lnM for amphibians, birds, fishes and mammals with their 95% confidence and prediction intervals; circle size again reflects precision. (E,F) Egger-type regressions of $|d|$ and lnM on $\sqrt{1/\tilde{n}_{0ij}}$ (= `sqrt(inverse sample size)`), with solid lines for fitted slopes, shaded 95% confidence bands and outer dotted lines for 95% prediction intervals; k is the number of effect sizes while the number in parentheses are the number of studies.

4 Discussion

We have introduced lnM, which quantifies the magnitude of separation between two groups as the ratio of between-group to within-group standard deviations, derived directly from the components of one-way ANOVA. By targeting "how far apart" groups are, irrespective of direction, and working on the log scale, lnM avoids the systematic bias that arises when absolute-valued effect sizes (e.g. |lnRR| or |SMD|) are analysed directly (Morrissey, 2016b,a), while remaining compatible with standard random-effects (multilevel) meta-analysis and meta-regression (Hedges, 1981; Hedges et al., 1999; Viechtbauer, 2010; Cheung, 2014). Conceptually, lnM complements existing log-ratio metrics such as lnRR, lnVR and lnCVR (Nakagawa et al., 2015; Senior et al., 2020), but targets a different quantity: the magnitude of mean separation relative to within-group noise. The zero point is interpretable: lnM = 0 corresponds to between-group dispersion matching the within-group SD, so inference is framed in terms of whether between-group spread meets or exceeds within-group variability, not whether group means are identical. Because it is constructed from means and standard deviations rather than raw ratios of means, lnM can be applied to both ratio- and interval-scale traits, including outcomes where lnRR is undefined or unstable (e.g. means near zero or with mixed signs). In this sense, lnM offers a pragmatic route to unsigned magnitude synthesis that complements the treatments of $|d|$ and d^2 based on folded-t and non-central-F distributions, proposed by Kulinskaya and Hoaglin (2023).

Our simulation study clarifies where lnM is reliable and where it becomes fragile. When the true separation is trivial and sample sizes are very small, plug-in lnM tends to be upwardly biased and the first-order delta-method sampling variance can be severely inflated, reflecting the algebraic sensitivity of lnM as the between-group mean square MS_B approaches the within-group mean square MS_W. This inflation leads to very large standard errors, which are simply not usable in meta-analysis. As mean differences or sample sizes increase, both bias and variance distortion decay rapidly, and plug-in lnM with delta-method variances performs well, consistent with earlier work on delta-method effect sizes such as lnRR, lnVR and lnCVR (Lajeunesse, 2011, 2015; Senior et al., 2020). The SAFE bootstrap (Mandel, 2013; Fletcher and Jowett, 2022; Nakagawa et al., 2025b) reduces point-estimate bias across the parameter grid and stabilises variance estimates, especially in the small-n, small-separation region, and crucially continues to return finite estimates when $MS_B \leq MS_W$, where the analytic definition of lnM fails. In practice, this means that SAFE allows analysts to retain all entries with small separations that would otherwise be discarded, while providing more realistic standard errors (although they are downwardly biased for small separations and sample sizes; Fig. 2) for routine meta-analysis and meta-regression.

The worked examples demonstrate how lnM behaves for real datasets and how it compares with the more familiar $|d|$. In the early-life nutrition data (Almeida et al., 2021), SAFE-based lnM estimates allowed us to retain all entries where $MS_B \leq MS_W$ and revealed moderator patterns, such as the decline in separation with increasing recovery time, that were less apparent on the $|d|$ scale. In the anthropogenic noise dataset (Kunc and Schmidt, 2019), lnM and $|d|$ agreed qualitatively on overall magnitudes and taxon-level differences, yet lnM made small-study (publication-bias) patterns clearer with Egger-type regressions using $\tilde{n}_0$-based moderators (Egger et al., 1997; Nakagawa et al., 2022). This showed more pronounced small-study effects for lnM, and bias-adjusted lnM means translated into noticeably smaller d-equivalents than naive summaries. In both examples, lnM could be analysed with standard multilevel models, and the back-transforms to d_{eq} provided familiar interpretations. At the same time, the SAFE procedure ensured that effect sizes were estimable even with $MS_B \leq MS_W$ rather than being discarded.

The proposed lnM also has limitations that warrant caution in certain circumstances. When separations and sample sizes are both very small, all estimators of lnM are noisy and biased to some degree (Fig. 1), and SAFE can only mitigate, not remove, this problem. In such cases, it may be more honest to conclude that dispersion differences are small and imprecisely estimated (e.g., if overall estimates are small, they are likely overestimates), although our proposed sensitivity analysis, with effect sizes satisfying $n_1 + n_2 \geq 40$, should largely eliminate this issue. In addition, the structural dependence of lnM's variance on $(MS_B - MS_W)$ means that larger lnM values tend to receive higher precision and greater weight, which makes this effect size highly sensitive to publication bias (i.e., a small-study effect could bias the overall estimate more than other effect size statistics; Egger et al., 1997; Sterne and Egger, 2001; note that this relationship itself does not bias estimates in meta-analysis and the issue only arises with publication bias). Yet, this tendency due to publication bias (small-study effect) could be substantially mitigated by utilising the harmonic mean of sample sizes (i.e., $\tilde{n}_0$).

Relatedly, with $\tilde{n}_0$, lnM allows us to not only test for but also potentially correct publication bias due to the small-study effect (for extensions of publication-bias tests, see Nakagawa et al., 2025a). Fig. 5 depicts how lnM could identify a small-study effect (Fig. 5C) even when it is not feasible to detect the small study effects using signed effect sizes such as d and lnRR (Fig. 5A,B; see also the online tutorial for how this ability of lnM could be used to correct for estimation bias in meta-analysis and meta-regression models). Notably, Egger-type regression cannot be applied to absolute effect sizes such as $|d|$ and $|\mathrm{lnRR}|$ due to the zero boundary (Fig. 5C). This could mean that, as we saw in our worked examples, past meta-analyses of magnitudes using $|d|$ and $|\mathrm{lnRR}|$ are likely to have produced severely overinflated results due to publication bias, which has never been discussed in the literature as far as we know (cf. Morrissey, 2016b,a). Then, all the past meta-analyses of magnitudes should be reanalysed with lnM to check and correct for publication bias, which could result in different conclusions from the original analysis. Therefore, the above-mentioned limitations of lnM, we believe, appear to be outweighed by the benefits of lnM over absolute-valued effect sizes. We also hope that future methodological research will discover improved estimators of lnM and its sample variance to address the current limitations, as we so far could not identify such estimators.

Taken together, despite its limitations, lnM offers a simple, meta-regression-friendly way to quantify and synthesise magnitudes of separation between groups, avoiding the pathologies of absolute-value effect sizes while remaining compatible with flexible multilevel and meta-regression models (Nakagawa and Santos, 2012; Cheung, 2014). Combined with the SAFE bootstrap, routine sensitivity analyses and the use of $\tilde{n}_0$-based moderators, lnM provides a practical tool for asking how far populations, treatments or strategies sit from each other, in the spirit of Lynch's variance-ratio framework (Lynch, 1990) and more recent work on variation-focused effect sizes (Nakagawa et al., 2015; Senior et al., 2020). We hope lnM will encourage more ecologists, evolutionary biologists, and other scientists to treat magnitude as a main target of synthesis, alongside mean shifts and variance ratios, thereby sharpening inference about the strength and structure of ecological and evolutionary processes and beyond.

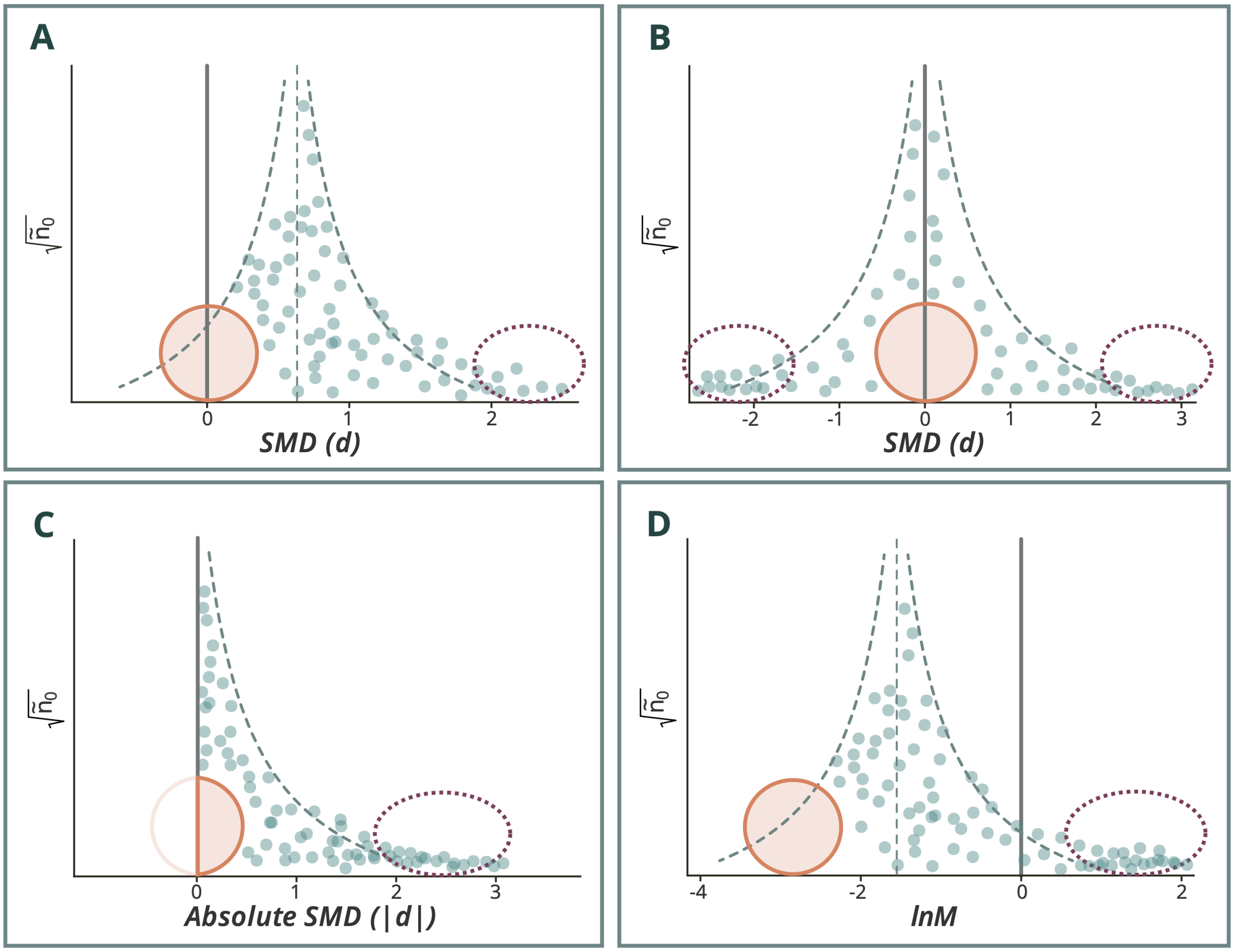

Figure 5: Funnel plots illustrate how magnitude transformations can hide or reveal small-study effects. The vertical axis represents effect-size precision, shown as $\sqrt{\tilde{n}_0}$, where $\tilde{n}_0 = n_0/2$ and $n_0 = 2n_1n_2/(n_1+n_2)$ is the harmonic-mean sample size for a two-group comparison. Teal points are individual effect-size estimates; dashed curves are pseudo-funnel limits (illustrative 95% bounds) that narrow with increasing precision (sample-size); note true funnel limits can be only written with precision (1/SE) not with $\sqrt{\tilde{n}_0}$. Solid vertical lines indicate the null value (0 on each effect-size scale). The orange circle highlights the region of small, imprecise studies (effect sizes) with near-null estimates that would be expected under no selection; the dotted purple ellipses highlight an excess of extreme estimates among low-precision (low-n) studies. (A) A conventional funnel plot for a signed standardised mean difference, SMD (d), showing the familiar funnel asymmetry associated with small-study effects (often interpreted as publication bias or selective reporting that favours larger effects in smaller studies). (B–D) The same underlying dataset (note that that dataset for A is different from that from B-D), shown on three analytically related scales: (B) signed SMD (d) centred at 0, where two-sided selection for statistical significance can yield a symmetric "missing-centre" pattern (few small studies near $d \approx 0$) that does not produce the classic left–right funnel asymmetry targeted by Egger-type regressions; C) absolute standardised mean differences, $|d|$, which folds (B) at 0 and induces a hard boundary and strong skew, violating the linear/Gaussian assumptions underpinning Egger-type regressions and making a funnel plot inherently non-symmetric; (D) the ln-magnitude effect size, lnM, an unbounded and near-Gaussian transformation of magnitude that re-expresses small separations as negative values and large separations as positive values, thereby reinstating a funnel-asymmetry pattern that can be assessed using Egger-type regression with a sample-size-based term (e.g. $1/\sqrt{\tilde{n}_0}$) in the same spirit as panel (A).

Acknowledgments

SN thank Wolfgang Viechtbauer for our initial discussion on lnM and inspiration and are also grateful for discussions with Yefeng Yang. SN, AM, EL and SO were supported by the Canada Excellence Research Chair (CERC-2022-00074) and NSERC Discovery Grant (RGPIN-2025-04813). Also, YY, SN, and ML were supported by the Australian Research Council Discovery Grant (DP230101248). DWAN and AMS were supported by an ARC Future Fellowship (FT220100276; FT230100240).

References

L. Z. Almeida, S. M. Hovick, S. A. Ludsin, and E. A. Marschall. Which factors determine the long-term effect of poor early-life nutrition? a meta-analytic review. *Ecosphere*, 12(8):e03694, 2021.

T. W. Anderson. *An introduction to multivariate statistical analysis*, volume 2. Wiley New York, 2003.

G. Arnqvist and D. Wooster. Meta-analysis: synthesizing research findings in ecology and evolution. *Trends in Ecology & Evolution*, 10(6):236–240, 1995.

M. Borenstein, L. V. Hedges, J. P. Higgins, and H. R. Rothstein. *Introduction to meta-analysis*. John wiley & sons, 2021.

M. W.-L. Cheung. Modeling dependent effect sizes with three-level meta-analyses: a structural equation modeling approach. *Psychological methods*, 19(2):211, 2014.

O. Cinar, S. Nakagawa, and W. Viechtbauer. Phylogenetic multilevel meta-analysis: a simulation study on the importance of modelling the phylogeny. *Methods in Ecology and Evolution*, 13(2):383–395, 2022.

A. C. Davison and D. V. Hinkley. Bootstrap methods and their application. *Cambridge Series in Statistical and Probabilistic Mathematics*, 1:1–592, 1997.

M. Egger, G. D. Smith, M. Schneider, and C. Minder. Bias in meta-analysis detected by a simple, graphical test. *bmj*, 315(7109):629–634, 1997.

D. Fletcher and T. Jowett. Single-fit bootstrapping: A simple alternative to the delta method. *Methods in Ecology and Evolution*, 13(4):835–843, 2022.

J. Gurevitch, J. Koricheva, S. Nakagawa, and G. Stewart. Meta-analysis and the science of research synthesis. *Nature*, 555(7695):175–182, 2018.

J. B. S. Haldane. Suggestions as to quantitative measurement of rates of evolution. *Evolution*, 3(1):51–56, 1949. doi: https://doi.org/10.1111/j.1558-5646.1949.tb00004.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1558-5646.1949.tb00004.x`.

L. V. Hedges. Distribution theory for glass's estimator of effect size and related estimators. *journal of Educational Statistics*, 6(2):107–128, 1981.

L. V. Hedges, J. Gurevitch, and P. S. Curtis. The meta-analysis of response ratios in experimental ecology. *Ecology*, 80(4):1150–1156, 1999.

J. P. A. Ioannidis, T. D. Stanley, and H. Doucouliagos. The power of bias in economics research. *The Economic Journal*, 127(605):F236–F265, 2017. doi: 10.1111/ecoj.12461.

Z. Irsova, P. R. Bom, T. Havranek, and H. Rachinger. Spurious precision in meta-analysis of observational research. *Nature Communications*, 16(1):8454, 2025.

M. D. Jennions and A. P. Møller. Publication bias in ecology and evolution: an empirical assessment using the trim and fill method. *Biological Reviews*, 77(2):211–222, 2002. doi: 10.1017/S1464793101005875.

J. G. Kingsolver, H. E. Hoekstra, J. M. Hoekstra, D. Berrigan, S. N. Vignieri, C. Hill, A. Hoang, P. Gibert, and P. Beerli. The strength of phenotypic selection in natural populations. *The American Naturalist*, 157(3):245–261, 2001.

E. Kulinskaya and D. C. Hoaglin. Simulations for meta-analysis of magnitude measures. *arXiv preprint arXiv:2310.00126*, 2023.

H. P. Kunc and R. Schmidt. The effects of anthropogenic noise on animals: a meta-analysis. *Biology Letters*, 15(11): 20190649, 2019.

M. J. Lajeunesse. On the meta-analysis of response ratios for studies with correlated and multi-group designs. *Ecology*, 92(11):2049–2055, 2011.

M. J. Lajeunesse. Bias and correction for the log response ratio in ecological meta-analysis. *Ecology*, 96(8):2056–2063, 2015.

M. Lynch. The rate of morphological evolution in mammals from the standpoint of the neutral expectation. *The American Naturalist*, 136(6):727–741, 1990.

M. Mandel. Simultaneous confidence intervals based on the bootstrap percentile method. *Ecological Applications*, 23 (3):690–695, 2013.

M. B. Morrissey. Rejoinder: Further considerations for meta-analysis of transformed quantities such as absolute values. *Journal of Evolutionary Biology*, 29(10):1922–1931, 2016a. doi: 10.1111/jeb.12951.

M. B. Morrissey. Meta-analysis of magnitudes, differences and variation in evolutionary parameters. *Journal of Evolutionary Biology*, 29(10):1882–1904, 2016b. doi: 10.1111/jeb.12950.

S. Nakagawa and E. S. Santos. Methodological issues and advances in biological meta-analysis. *Evolutionary Ecology*, 26(5):1253–1274, 2012.

S. Nakagawa, R. Poulin, K. Mengersen, K. Reinhold, L. Engqvist, M. Lagisz, and A. M. Senior. Meta-analysis of variation: ecological and evolutionary applications and beyond. *Methods in Ecology and Evolution*, 6(2):143–152, 2015. doi: 10.1111/2041-210X.12309.

S. Nakagawa, M. Lagisz, R. E. O'Dea, J. Rutkowska, Y. Yang, D. W. Noble, and A. M. Senior. The orchard plot: cultivating a forest plot for use in ecology, evolution, and beyond. *Research Synthesis Methods*, 12(1):4–12, 2021.

S. Nakagawa, M. Lagisz, M. D. Jennions, J. Koricheva, D. W. Noble, T. H. Parker, A. Sánchez-Tójar, Y. Yang, and R. E. O'Dea. Methods for testing publication bias in ecological and evolutionary meta-analyses. *Methods in Ecology and Evolution*, 13(1):4–21, 2022.

S. Nakagawa, M. Lagisz, R. E. O'Dea, P. Pottier, J. Rutkowska, A. M. Senior, Y. Yang, and D. W. A. Noble. orchard 2.0: An r package for visualising meta-analyses with orchard plots. *Methods in Ecology and Evolution*, 14(8): 2003–2010, 2023. doi: https://doi.org/10.1111/2041-210X.14152. URL `https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.14152`.

S. Nakagawa, A. Mizuno, K. Morrison, L. Ricolfi, C. Williams, S. M. Drobniak, M. Lagisz, and Y. Yang. Location-scale meta-analysis and meta-regression as a tool to capture large-scale changes in biological and methodological heterogeneity: A spotlight on heteroscedasticity. *Global Change Biology*, 31(5):e70204, 2025a.

S. Nakagawa, A. Mizuno, C. Williams, S. Ortega, S. M. Drobniak, M. Lagisz, Y. Yang, A. M. Senior, D. W. Noble, and E. Lundgren. Mastering an accurate and generalizable simulation-based method to obtain bias-corrected point estimates and sampling variance for any effect sizes. *arXiv preprint arXiv:2510.27519*, 2025b.

D. W. Noble, M. Lagisz, R. E. O'dea, and S. Nakagawa. Nonindependence and sensitivity analyses in ecological and evolutionary meta-analyses, 2017.

J. E. Pustejovsky and E. Tipton. Meta-analysis with robust variance estimation: Expanding the range of working models. *Prevention Science*, 23(3):425–438, 2022.

A. M. Senior, W. Viechtbauer, and S. Nakagawa. Revisiting and expanding the meta-analysis of variation: The log coefficient of variation ratio. *Research Synthesis Methods*, 11(4):553–567, 2020. doi: 10.1002/jrsm.1431.

T. D. Stanley and H. Doucouliagos. Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, 5(1):60–78, 2014.

J. A. Sterne and M. Egger. Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. *Journal of clinical epidemiology*, 54(10):1046–1055, 2001.

R. J. Tibshirani and B. Efron. An introduction to the bootstrap. *Monographs on statistics and applied probability*, 57(1):1–436, 1993.

W. N. Venables and B. D. Ripley. *Modern applied statistics with S-PLUS*. Springer Science & Business Media, 2013.

J. M. Ver Hoef. Who invented the delta method? *The American Statistician*, 66(2):124–127, 2012.

W. Viechtbauer. Conducting meta-analyses in r with the metafor package. *Journal of statistical software*, 36(3):1–48, 2010.

C. Williams, Y. Yang, M. Lagisz, K. Morrison, L. Ricolfi, D. I. Warton, and S. Nakagawa. Transparent reporting items for simulation studies evaluating statistical methods: Foundations for reproducibility and reliability. *Methods in Ecology and Evolution*, 15(11):1926–1939, 2024.

C. Williams, Y. Yang, D. I. Warton, and S. Nakagawa. Modelling approaches for meta-analyses with dependent effect sizes in ecology and evolution: A simulation study. *Methods in Ecology and Evolution*, 16(10):2362–2379, 2025.

Appendix

Appendix A: Delta-method derivation of the sampling variance of lnM

In this appendix, we outline how the sampling variances in Equations (5) and (6) follow from a first-order Taylor expansion (delta method) applied to lnM. We treat the two-group case with independent samples first and then the paired (dependent) design.

A.1 Independent groups

Recall that for two independent groups with sample sizes n_1, n_2, means $\bar{X}_1, \bar{X}_2$, variances s_1^2, s_2^2, and harmonic-mean sample size $n_0 = 2n_1n_2/(n_1+n_2)$, the ANOVA mean squares are:

$$MS_B = \frac{n_1 n_2}{n_1+n_2}\,(\bar{X}_1-\bar{X}_2)^2 = \frac{n_1n_2}{n_1+n_2}\,\delta^2, \tag{A1}$$

$$MS_W = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}, \tag{A2}$$

with $\delta = \bar{X}_1 - \bar{X}_2$. We also define the difference:

$$\Delta = MS_B - MS_W, \tag{A3}$$

and note that the pooled within-group variance is $s_W^2 = MS_W$. The log-magnitude effect size is:

$$\mathrm{lnM} = \frac{1}{2}\left[\ln(\Delta) - \ln(n_0) - \ln(MS_W)\right]. \tag{A4}$$

For the independent design, the random quantities entering: (A4) are $(\bar{X}_1, \bar{X}_2, s_1^2, s_2^2)$, with

$$\mathrm{Var}(\bar{X}_1) = \frac{s_1^2}{n_1}, \qquad \mathrm{Var}(\bar{X}_2) = \frac{s_2^2}{n_2}, \qquad \mathrm{Cov}(\bar{X}_1,\bar{X}_2) = 0, \tag{A5}$$

$$\mathrm{Var}(s_1^2) = \frac{2s_1^4}{n_1-1}, \qquad \mathrm{Var}(s_2^2) = \frac{2s_2^4}{n_2-1}, \qquad \mathrm{Cov}(s_1^2, s_2^2) = 0, \tag{A6}$$

and, under normal-theory assumptions,

$$\mathrm{Cov}(\bar{X}_j, s_k^2) = 0 \quad \text{for all } j,k \in \{1,2\}. \tag{A7}$$

We now treat lnM as a smooth function of the vector:

$$\boldsymbol{\zeta} = (\bar{X}_1, \bar{X}_2, s_1^2, s_2^2)^\top,$$

and apply the delta method:

$$\mathrm{Var}_{\mathrm{IND}}(\mathrm{lnM}) \approx \nabla_\zeta \mathrm{lnM}^\top\, \Sigma_\zeta\, \nabla_\zeta \mathrm{lnM}, \tag{A8}$$

where Σ_ζ is the covariance matrix of $\boldsymbol{\zeta}$ and $\nabla_\zeta\mathrm{lnM}$ is the gradient of (A4) with respect to $(\bar{X}_1, \bar{X}_2, s_1^2, s_2^2)$.

To simplify notation, note that:

$$\delta = \bar{X}_1 - \bar{X}_2, \qquad s_D^2 = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}, \tag{A9}$$

so that s_D^2 is the sampling variance of δ under independence. Differentiating equation (A4) and using the chain rule for $\Delta = MS_B - MS_W$ yields:

$$\frac{\partial\,\mathrm{lnM}}{\partial \bar{X}_1} = \frac{1}{2}\left(\frac{1}{\Delta}\,\frac{\partial \Delta}{\partial \bar{X}_1}\right) = \frac{1}{2}\left(\frac{1}{\Delta}\,\frac{2n_1n_2}{n_1+n_2}\,\delta\right) = \frac{n_0}{2}\,\frac{\delta}{\Delta}, \tag{A10}$$

$$\frac{\partial\,\mathrm{lnM}}{\partial \bar{X}_2} = -\,\frac{n_0}{2}\,\frac{\delta}{\Delta}. \tag{A11}$$

Similarly, taking derivatives with respect to s_1^2 and s_2^2 and collecting terms shows that all contributions from the mean squares can be written in terms of Δ, MS_W and MS_B. After substituting the variances and covariances of $(\bar{X}_1, \bar{X}_2, s_1^2, s_2^2)$ and simplifying, we obtain:

$$\mathrm{Var}_{\mathrm{IND}}(\mathrm{lnM}) = \frac{1}{4\Delta^2}\left[\left(\tfrac{n_0}{2}\right)^2 \left(2\, s_D^4 + 4\, s_D^2\, \delta^2\right) + \frac{MS_B^2}{MS_W^2}\, \frac{2\, MS_W^2}{n_1 + n_2 - 2}\right], \tag{A12}$$

which is the expression reported in equation (5) in the main text. Here, the first term inside the brackets arises from the uncertainty in the mean difference δ (via s_D^2), and the second term arises from the sampling variability of MS_W.

A.2 Paired (dependent) groups

For paired data with $n_1 = n_2 = n$, means $\bar{X}_1, \bar{X}_2$, variances s_1^2, s_2^2, and within-pair correlation r, the between- and within-group mean squares are:

$$MS_B = \frac{n}{2}\,(\bar{X}_1 - \bar{X}_2)^2 = \frac{n}{2}\,\delta^2, \tag{A13}$$

$$MS_W = \frac{s_1^2 + s_2^2}{2}, \tag{A14}$$

again with $\delta = \bar{X}_1 - \bar{X}_2$ and $\Delta = MS_B - MS_W$. The lnM definition (A4) remains the same, but the joint sampling distribution of $(\bar{X}_1, \bar{X}_2, s_1^2, s_2^2)$ now includes non-zero covariances:

$$\mathrm{Var}(\bar{X}_1) = \frac{s_1^2}{n}, \qquad \mathrm{Var}(\bar{X}_2) = \frac{s_2^2}{n}, \qquad \mathrm{Cov}(\bar{X}_1, \bar{X}_2) = \frac{r\, s_1 s_2}{n}, \tag{A15}$$

$$\mathrm{Var}(s_1^2) = \frac{2 s_1^4}{n-1}, \qquad \mathrm{Var}(s_2^2) = \frac{2 s_2^4}{n-1}, \qquad \mathrm{Cov}(s_1^2, s_2^2) = \frac{2 r^2 s_1^2 s_2^2}{n-1}, \tag{A16}$$

while the covariances between means and variances remain zero under normality:

$$\mathrm{Cov}(\bar{X}_j, s_k^2) = 0 \quad \text{for all } j, k \in \{1, 2\}. \tag{A17}$$

The variance of the within-pair difference $D = X_1 - X_2$ is

$$s_D^2 = s_1^2 + s_2^2 - 2r\, s_1 s_2. \tag{A18}$$

As in the independent case, we apply the delta method with $\boldsymbol{\zeta} = (\bar{X}_1, \bar{X}_2, s_1^2, s_2^2)^\top$ and equation (A4). Differentiating Δ and MS_W now uses the paired definitions of MS_B and MS_W, but the structure of the gradient is analogous. After substituting the covariance matrix of $\boldsymbol{\zeta}$ for the paired design and simplifying, the variance of lnM is:

$$\mathrm{Var}_{\mathrm{DEP}}(\mathrm{lnM}) = \frac{1}{4\Delta^2}\left[\left(\frac{n}{2}\right)^2 \left(\frac{2\, s_D^4}{n^2} + \frac{4\, \delta^2\, s_D^2}{n}\right) + \frac{MS_B^2}{MS_W^2}\, \frac{s_1^4 + s_2^4 + 2\, r^2 s_1^2 s_2^2}{2\,(n-1)}\right], \tag{A19}$$

which matches equation (6) in the main text. The first bracketed term captures the contribution of uncertainty in the mean difference (through s_D^2), while the second term reflects the sampling variability of MS_W and the correlation-induced coupling between s_1^2 and s_2^2.

In both independent and paired designs, the leading factor $1/(4\Delta^2)$ shows that the sampling variance of lnM grows quickly as MS_B approaches MS_W (i.e. $\Delta \to 0$), and shrinks rapidly once MS_B pulls away from MS_W, as discussed in the main text.

Appendix B: From ANOVA components to the d-equivalent for lnM

We first derive the expression for lnM in terms of the mean difference δ and the within-group mean square MS_W, and then obtain the corresponding absolute standardized mean difference d_{eq}.

For two groups with sample sizes n_1, n_2, means $\bar{X}_1, \bar{X}_2$, and variances s_1^2, s_2^2, the one-way ANOVA mean squares are:

$$MS_B = \frac{n_1 n_2}{n_1 + n_2} (\bar{X}_1 - \bar{X}_2)^2 = \frac{n_1 n_2}{n_1 + n_2} \delta^2, \tag{A20}$$

$$MS_W = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}, \tag{A21}$$

where

$$\delta = \bar{X}_1 - \bar{X}_2.$$

The harmonic-mean sample size is

$$n_0 = \frac{2 n_1 n_2}{n_1 + n_2}. \tag{A22}$$

Following the main text, we define the between-group and within-group variance components as

$$s_B^2 = \frac{MS_B - MS_W}{n_0}, \qquad s_W^2 = MS_W, \tag{A23}$$

so that the log-magnitude effect size is

$$\mathrm{lnM} = \ln\left(\frac{s_B}{s_W}\right) = \tfrac{1}{2}\left[\ln(s_B^2) - \ln(s_W^2)\right]. \tag{A24}$$

Substituting (A23) into (A24) gives:

$$\begin{aligned} \mathrm{lnM} &= \tfrac{1}{2}\left[\ln\left(\frac{MS_B - MS_W}{n_0}\right) - \ln(MS_W)\right] \\ &= \tfrac{1}{2}\left[\ln(MS_B - MS_W) - \ln(n_0) - \ln(MS_W)\right]. \end{aligned} \tag{A25}$$

Now use (A20) and (A22) to write

$$MS_B = \frac{n_1 n_2}{n_1 + n_2} \delta^2 = \frac{n_0}{2} \delta^2.$$

Hence

$$MS_B - MS_W = \frac{n_0}{2} \delta^2 - MS_W, \tag{A26}$$

$$\mathrm{lnM} = \tfrac{1}{2}\left[\ln\left(\tfrac{n_0}{2} \delta^2 - MS_W\right) - \ln(n_0) - \ln(MS_W)\right]. \tag{A27}$$

Dividing the argument of the logarithm by MS_W and using $s_W^2 = MS_W$ yields:

$$\begin{aligned} \mathrm{lnM} &= \tfrac{1}{2}\left[\ln\left(\frac{\tfrac{n_0}{2} \delta^2 - MS_W}{n_0 MS_W}\right)\right] \\ &= \tfrac{1}{2} \ln\left(\frac{\delta^2}{2\, s_W^2} - \frac{1}{n_0}\right). \end{aligned} \tag{A28}$$

Starting from (A28),

$$\mathrm{lnM} = \tfrac{1}{2} \ln\left(\frac{\delta^2}{2\, s_W^2} - \frac{1}{n_0}\right),$$

we exponentiate and rearrange:

$$e^{2\,\mathrm{lnM}} = \frac{\delta^2}{2\, s_W^2} - \frac{1}{n_0}, \tag{A29}$$

$$\frac{\delta^2}{s_W^2} = 2\, e^{2\,\mathrm{lnM}} + \frac{2}{n_0}. \tag{A30}$$

Defining the implied absolute standardized mean difference as:

$$d_{\mathrm{eq}} = \frac{|\delta|}{s_W},$$

we obtain the relationship:

$$d_{\text{eq}} = \sqrt{2e^{2\,\ln\text{M}} + \frac{2}{n_0}}\,, \tag{A31}$$

which matches equation (21).

For typical meta-analytic applications, the term $2/n_0$ is small, yielding the large-sample approximation

$$d_{\text{eq}} \approx \sqrt{2}\,e^{\ln\text{M}} = \sqrt{2}\,\frac{s_B}{s_W}, \tag{A32}$$

which we use for the interpretive values reported in Table 1 and matches equation (22).

Appendix C: Centering lnM to a target absolute SMD

We want to re-anchor $\ln M$ so that zero corresponds to a user-chosen absolute standardized separation $d_0 = |\delta|/s_W$.

From Appendix B, the absolute SMD implied by lnM is:

$$d_{\text{eq}}^2 \;=\; 2\,e^{2\ln\text{M}} + \frac{2}{n_0}, \qquad n_0 = \frac{2n_1n_2}{n_1+n_2}. \tag{A33}$$

Setting $d_{\text{eq}} = d_0$ in (A33) gives:

$$e^{2\ln\text{M}^*} = \frac{d_0^2}{2} - \frac{1}{n_0} = \frac{n_0d_0^2 - 2}{2n_0}. \tag{A34}$$

Taking the logarithm and dividing by 2 on both sides gives:

$$\ln\text{M}^* = \tfrac{1}{2}\ln\!\left(\frac{n_0d_0^2 - 2}{2n_0}\right). \tag{A35}$$

Note that existence requires $(n_0/2)\,d_0^2 > 1$, typically satisfied in meta-analytic applications.

Subtracting the anchor yields:

$$\begin{aligned}\ln\text{M}^{[d_0]} &= \ln\text{M} - \ln\text{M}^* \\ &= \ln\text{M} + \tfrac{1}{2}\ln\!\left(\frac{2n_0}{n_0d_0^2 - 2}\right) = \ln\text{M} + \tfrac{1}{2}\ln\!\left(\frac{n_0}{\frac{n_0}{2}\,d_0^2 - 1}\right).\end{aligned} \tag{A36}$$

Thus $\ln\text{M}^{[d_0]} = 0$ exactly when $d_{\text{eq}} = d_0$.

When n_0 is large, $\ln\text{M}^* \approx \ln(d_0/\sqrt{2})$, so

$$\ln\text{M}^{[d_0]} \;\approx\; \ln\text{M} + \ln\!\left(\frac{\sqrt{2}}{d_0}\right).$$

In particular, for $d_0 = 1$, $\ln\text{M}^{[1]} \approx \ln\text{M} + \ln(\sqrt{2})$.

Importantly, (sampling) variance is invariant under this type of centering. Let $c_i = \frac{1}{2}\ln\left(\frac{n_{0i}}{\frac{n_{0i}}{2}\,d_0^2 - 1}\right)$. Since c_i is a deterministic function of sample sizes, $\ln\text{M}_i^{[d_0]} = \ln\text{M}_i + c_i$ has $\text{Var}(\ln\text{M}_i^{[d_0]}) = \text{Var}(\ln\text{M}_i)$ and identical inverse-variance weights. Confidence intervals translate by c_i. In meta-analysis, therefore, we could fit models on lnM and apply the offset c_i *post hoc* to fitted means and intervals.

Appendix D: Why not use $\ln|d|$ as a magnitude effect size?

Given that the standardized mean difference (SMD) is widely used in meta-analysis, it is natural to ask whether a magnitude analysis could be performed by working on a log scale with the absolute SMD, i.e. $\ln|d|$. At first glance, $\ln|d|$ may appear attractive because it is unbounded above and thus might seem to provide a "Gaussian-friendly" analysis scale. Here we show that, although $\ln|d|$ has a simple delta-method sampling variance, it is undefined at $d = 0$. This boundary issue is a key reason we prefer lnM as our primary magnitude effect size in the main text.

Let d denote the study-level standardized mean difference (Cohen's d), with estimated sampling variance $v_d = \mathrm{Var}(d)$ obtained from standard SMD theory for the relevant design. We define the log-absolute SMD as:

$$y \equiv \ln|d|. \tag{A37}$$

This transform is undefined when $d = 0$ and tends to $-\infty$ as $|d| \to 0$.

We treat $y = \ln|d|$ as a smooth function of d for $d \neq 0$. For $d \neq 0$,

$$\frac{dy}{dd} = \frac{d}{dd}\ln|d| = \frac{1}{d}. \tag{A38}$$

A first-order Taylor expansion (delta method) therefore gives:

$$\mathrm{Var}(\ln|d|) \;\approx\; \left(\frac{1}{d}\right)^2 \mathrm{Var}(d) \;=\; \frac{v_d}{d^2}, \qquad \mathrm{SE}(\ln|d|) \;\approx\; \frac{\sqrt{v_d}}{|d|}. \tag{A39}$$

If one uses an alternative expression for v_d (exact or approximate; independent-groups or paired designs), the delta-method mapping remains Eq. (A39).

For two independent groups of sizes n_1, n_2, a commonly used large-sample approximation for the sampling variance of the SMD is

$$v_d \;\approx\; \frac{n_1 + n_2}{n_1 n_2} \;+\; \frac{d^2}{2(n_1 + n_2 - 2)}. \tag{A40}$$

Substituting into Eq. (A39) yields

$$\mathrm{Var}(\ln|d|) \;\approx\; \frac{n_1 + n_2}{n_1 n_2\, d^2} \;+\; \frac{1}{2(n_1 + n_2 - 2)}. \tag{A41}$$

Other small-sample or design-specific formulas for v_d can be used in the same way.

Equations (A39)–(A41) highlight that $\ln|d|$ behaves awkwardly when the standardized mean difference is close to zero. The first issue is definitional: $\ln|d|$ does not exist at $d = 0$ and it tends to $-\infty$ as $|d| \to 0$. In practice, this means that an analysis based on $\ln|d|$ must adopt an explicit rule for handling $d = 0$ (and very small $|d|$), for example by excluding such cases or by applying a constant such as $\ln\sqrt{d^2 + c}$ for some $c > 0$.

The second issue concerns precision. From Eq. (A39), the delta-method variance is proportional to $1/d^2$, so the sampling variance on the $\ln|d|$ scale increases rapidly as $|d|$ approaches zero. Given that near-null effects are common in empirical syntheses, the chosen handling rule near $d = 0$ can affect the behaviour and interpretation of meta-analytic models.

The motivation for lnM in the main text is to provide a log-scale magnitude effect size without requiring an additional convention specifically to handle $d = 0$. Like $\ln|d|$, lnM has non-trivial behaviour near its own boundary region (notably when $MS_B \leq MS_W$), but our framework makes that boundary explicit and provides a principled operational solution via SAFE. In this sense, Appendix D shows that $\ln|d|$ offers a simple delta-method sampling variance but is undefined at $d = 0$, whereas lnM is designed to target a magnitude parameter while retaining a workable estimation strategy across the problematic boundary region.