esa

ECOSPHERE

Bias in meta-analyses using Hedges' *d*

Elizabeth A. Hamman[1,4]† Paula Pappalardo[1] James R. Bence,[2] Scott D. Peacor,[3] and Craig W. Osenberg[1]

[1]*Odum School of Ecology, University of Georgia, 140 E. Green Street, Athens, Georgia 30602 USA*
[2]*Quantitative Fisheries Center, Department of Fisheries and Wildlife, Michigan State University, 480 Wilson Road #13, East Lansing, Michigan 48824 USA*
[3]*Department of Fisheries and Wildlife, Michigan State University, 375 Wilson Road, 101 UPLA Building, East Lansing, Michigan 48824 USA*

Citation: Hamman, E. A., P. Pappalardo, J. R. Bence, S. D. Peacor, and C. W. Osenberg. 2018. Bias in meta-analyses using Hedges' *d*. Ecosphere 9(9):e02419. 10.1002/ecs2.2419

Abstract. The type of metric and weighting method used in meta-analysis can create bias and alter coverage of confidence intervals when the estimated effect size and its weight are correlated. Here, we investigate bias associated with the common metric, Hedges' *d*, under conditions common in ecological meta-analyses. We simulated data from experiments, computed effect sizes and their variances, and performed meta-analyses applying three weighting schemes (inverse variance, sample size, and unweighted) for varying levels of effect size, within-study replication, number of studies in the meta-analysis, and among-study variance. Unweighted analyses, and those using weights based on sample size, were close to unbiased and yielded coverages close to the nominal level of 0.95. In contrast, the inverse-variance weighting scheme led to bias and low coverage, especially for meta-analyses based on studies with low replication. This bias arose because of a correlation between the estimated effect and its weight when using the inverse-variance method. In many cases, the sample size weighting scheme was most efficient, and, when not, the differences in efficiency among the three methods were relatively minor. Thus, if using Hedges' *d*, we recommend using weights based upon sample size that do not involve individual study estimates of the effect size.

Key words: bias; coverage; effect size; Hedges' *d;* meta-analysis; sample size; weights.

Received 10 July 2018; accepted 16 July 2018. Corresponding Editor: Debra P. C. Peters.

[4] Present address: Department of Biology, East Carolina University, N108 Howell Science Complex, Greenville, North Carolina 27858, USA.
† **E-mail:** eahamman@gmail.com

Introduction

Meta-analysis allows researchers to make broad conclusions and evaluate factors that influence the outcomes of experiments by quantitatively synthesizing data from primary studies. Papers based on meta-analyses are becoming increasingly prevalent in ecology and often have a large influence in our discipline (Cadotte et al. 2012, Gurevitch and Koricheva 2013, Lortie 2014, Gurevitch et al. 2018). Therefore, it is essential to evaluate the conditions under which established methods are valid (Whittaker 2010, Lajeunesse 2015). If they are not valid under conditions commonly encountered in actual datasets, then it is important to seek alternative approaches.

In meta-analysis, a quantitative estimate of the phenomenon being studied (i.e., the effect size) is extracted from each study. The observed effect sizes will vary among studies due to systematic (i.e., fixed) effects, real but unexplained variation about what is expected given the fixed effects (called among-study variance, τ^2), and experimental error (called within-study variance, v_i). In general, the within-study variance depends on the variability among replicates and the sample size

(number of replicates) for each original study. The presence of both variances (τ^2 and v_i) results in a random-effects model. If heterogeneity is assumed absent ($\tau^2 = 0$), we have a fixed-effects model. The random-effects model is recommended in ecology because of the large amount of heterogeneity among the studies being synthesized (Gurevitch and Hedges 1993, 1999, Senior et al. 2016).

To obtain a pooled effect size averaged over all studies (potentially conditional on covariates), meta-analyses typically weight each observed effect size based on the precision of each estimate, with more precise estimates receiving greater weights. The most common and often most efficient (reducing the error in the estimation) weight is the reciprocal of the sum of the within-study and among-study variances (Hedges and Olkin 1985, Gurevitch et al. 2001, Lajeunesse 2010):

$$w_i = \frac{1}{v_i + \tau^2} \quad (1)$$

where w_i is the weight assigned to the effect size from study i, v_i is estimated from the data prior to analysis (and is unique to each study), and τ^2 is estimated during the meta-analysis (and is common to all studies).

There are many ways to define ecological effects (Osenberg et al. 1997, 1999, Koricheva et al. 2013). This paper focuses on one of the most common metrics used in ecology (used in approximately 25–30% of studies: Nakagawa and Santos 2012, Koricheva and Gurevitch 2014), the standardized mean difference between two treatments (i.e., Hedges' d: Hedges 1981, 1982, 1983):

$$d_i = \frac{\bar{X}_{i,T} - \bar{X}_{i,C}}{s_i} J_i \quad (2)$$

where $\bar{X}_{i,T}$ is the mean of the Treatment, $\bar{X}_{i,C}$ is the mean of the Control group, s_i is the pooled standard deviation, $J_i = \left(1 - \frac{3}{4(n_{i,C} + n_{i,T} - 2) - 1}\right)$ is a correction for small sample bias (Hedges 1981), and d_i is an estimate of δ_i, the true study-specific effect size (i.e., standardized difference). The within-study variance of d_i is usually estimated as follows:

$$v_i = \frac{n_{i,T} + n_{i,C}}{n_{i,T} n_{i,C}} + \frac{d_i^2}{2(n_{i,T} + n_{i,C})} \quad (3)$$

where $n_{i,T}$ and $n_{i,C}$ are the number of replicates for the Treatment and Control groups (Hedges 1981). Equation 3 is based on the sampling distribution of d from Eq. 2, which equals a constant time a non-central t-distribution (Hedges 1981, 1982). In this equation, d (as a replacement of δ) attempts to account for the increase in variance of d due to the increase in magnitude of δ (Hedges 1981, 1982).

In some cases, this standard weighting approach can introduce problems. Because the estimate of the effect size for study i, d_i (Eq. 2), is used to estimate its variance (Eq. 3), any error in the estimation of the true standardized difference for a study, δ_i, will be propagated into error in the estimated variance and thus into the weight given to the study (Hedges 1982). Due to sampling error, some studies will provide estimated effect sizes (d_i) that are smaller in magnitude (closer to zero) than δ_i and others will yield estimated effects that are larger in magnitude than δ_i. As a result, the weights will be too low for studies that overestimate the magnitude of the effect but too high for studies that underestimate the magnitude. When the true pooled effect is non-zero, these errors lead to bias in the estimated pooled effect size (Hedges 1982, 1983). Similar biases can also arise in random-effects meta-analysis using other metrics, such as the proportion of successful trials (e.g., survival; Böhning et al. 2002). Because of this possible bias, Hedges (1982) proposed an alternative estimator of the within-study variance (in place of Eq. 3) that avoided using the estimate d_i in its calculation (see also Appendix S1):

$$v_i = \left(1 - \frac{3}{4(n_{i,T} + n_{i,C} - 2) - 1))}\right)^2 \times \left(\frac{n_{i,T} + n_{i,C} - 2}{\frac{n_{i,T} \times n_{i,C}}{n_{i,T} + n_{i,C}}(n_{i,T} + n_{i,C} - 4)}\right) \quad (4)$$

Despite long-standing knowledge that Eq. 3 can lead to bias, and the potential for Eq. 4 to reduce this bias, Eq. 3 remains the standard approach for meta-analyses that use Hedges' d (e.g., Hillebrand and Gurevitch 2014). Fortunately, for fixed-effects meta-analysis the bias (using Eq. 3) is generally small (Hedges 1982, Sánchez-Meca and Marín-Martínez 1998). Based on this, Hedges (1983) argued (but without direct evidence) that the bias would be small in random-effects meta-analyses unless the among-study variance was large, and recent simulations found relatively small bias using Eq. 3 (Van Den Noortgate and Onghena

2003, Johnson et al. 2013). However, these simulations explored ranges of conditions typical in other fields (e.g., medical sciences and psychology) that may not reflect the conditions under which meta-analyses are conducted in ecology. For example, both Johnson et al. (2013) and Sánchez-Meca and Marín-Martínez (1998) used a minimum average replication of 30, and Van Den Noortgate and Onghena (2003) used 10, 25, and 50 replicates. Ecological studies often have smaller levels of replication (e.g., averaging 8–9, but often with as few as 2–4 replicates per treatment: Hillebrand and Gurevitch 2014).

Furthermore, heterogeneity in ecological studies (i.e., the magnitude of variation in the actual effect sizes among studies) may be much higher than in other disciplines, such as medicine (Senior et al. 2016). Importantly, the magnitude of bias in Hedges' *d* increases as among-study heterogeneity increases (Böhning et al. 2002). Unfortunately, the ability to generalize from Böhning et al. (2002) to ecological contexts is limited because they considered effects of heterogeneity under just one, relatively high level of replication ($\bar{n}$ = 18), and used the now outdated Dersimonian-Laird estimator of among-study variance. Therefore, there is a need to evaluate the performance of random-effects meta-analyses using different weighting schemes under a range of ecologically relevant conditions.

In this paper, we simulated data and conducted meta-analyses using Hedges' *d* with three different methods: (1) analyses using inverse-variance weights (Eq. 3); (2) analyses using the sample size-based approximation (Eq. 4; Hedges 1982); and (3) unweighted analyses. Our primary objectives were to (a) determine whether the inverse-variance method (Eq. 3) leads to substantial bias in situations that are characteristic of ecological meta-analyses (low replication and high heterogeneity) and (b) evaluate whether alternative weighting approaches eliminate or reduce such bias without a substantial loss of efficiency. If effective alternatives exist, this would suggest a way forward for future ecological meta-analyses.

Methods

We compared the performance of the three weighting schemes when applied to data that were simulated over an ecologically plausible range of conditions. In the simulations, we simultaneously varied four elements of an ecological meta-analytic dataset: replicates per study (n), number of studies in the meta-analysis (k), true pooled effect size (δ), and true among-study variance (τ^2). We generated the data for each of the k studies by drawing n_i replicates for each of two treatments. For each of our simulated experiments, we drew the sample size for each study from a negative binomial distribution with a specified mean, $\bar{n}$, and with a fixed level of over-dispersion (θ = 1.55: Appendix S2), based upon data in Levin (1992). We set all generated samples sizes of 1 or 2 to 3. We assumed that the Control (C) and Treatment (T) groups each had identical among-replicate variance, σ^2, equal to 1.0 in all simulations. We fixed σ^2 at a single value and varied the among-study variation, τ^2, as we were interested in the effect of the relative magnitude of these two variances. The true means for the Treatment and Control groups in the ith simulated study (i.e., what the observed mean converges to as $n_i \to$ infinite) differed by $\delta\sigma + \eta_i$, where η_i is the random-effect associated with study i. Thus, $C \sim N(\mu, \sigma^2)$ and $T \sim N(\mu + \delta\sigma + \eta, \sigma^2)$, where μ is the control mean (=1 in our simulations), and $\eta \sim N(0, \tau^2)$. We repeated this process for k studies to generate each meta-analytic dataset.

We simulated data over a range of ecologically relevant conditions defined by a fully crossed design that varied sample size (assumed equal for the two treatments within a study: $\bar{n}$ = {4, 6, 8, 10, 12, 14, 16, 20, 25}), magnitude of the true difference between treatments ($\delta\sigma$ = {0, 0.1, 0.15, 0.25, 0.35, 0.5, 0.6, 0.75, 1, 1.25, 1.5, 2.5}), the number of studies in the meta-analysis (k = {5, 10, 15, 25, 35, 45, 55, 75, 100, 125}), and the among-study variance (τ^2 = {0, 0.1, 0.5, 1, 2.5, 5, 10}). These levels were chosen to represent ecologically relevant conditions for a meta-analysis (See Appendix S2 for rationale). For each combination of factors, we generated 10,000 replicate meta-analytic datasets.

Each meta-analytic dataset was analyzed with one of three approaches: (1) weighting by the traditional inverse-variance estimator (using Eqs. 1 and 3); (2) weighting by the sample size-based approximation of Hedges (1982; Eqs. 1 and 4); and (3) unweighted. We completed all analyses (code

available in Data S1) using R Statistical Software (R Core Team 2018, version 3.2.2). To perform the random-effect meta-analyses using the inverse-variance or sample size-based weighting, we used the function "rma" in R's metafor package (Viechtbauer 2010) and estimated the among-study variance with the restricted maximum likelihood method. We calculated confidence intervals using the Knapp-Hartung correction by setting option knha = TRUE (Viechtbauer 2010); this is particularly important for small k and is analogous to using a t- vs. Z-distribution to construct confidence intervals. For each unweighted analysis, we calculated the unweighted mean effect across the k studies and calculated a confidence interval using a t-distribution with df = $k-1$. We used our own R code for the unweighted analysis rather than using metafor because, although metafor will calculate unweighted means, it still uses the within-study variances to estimate the confidence interval for the unweighted means.

For each meta-analysis, we recorded the estimated pooled effect size ($\bar{d}$) and the 95% confidence interval, and determined if the confidence interval contained the true pooled effect size (δ). After running all the meta-analyses for each parameter combination, we estimated the bias (as the average difference between the observed and the true pooled effect size: $\bar{d} - \delta$), coverage (the proportion of 10,000 meta-analyses in which the confidence interval contained the true value: e.g., a 95% confidence interval should contain the true value 95% of the time), and the efficiency (as the root mean square error). A meta-analysis procedure is more efficient when the error (including bias) in the estimates it produces is low. For each set of simulation conditions, we calculated 95% confidence intervals for the estimated bias using a Student's t-distribution and obtained the 95% confidence intervals for the estimated coverage using Wilson binomial confidence intervals implemented in the R package binom (Dorai-Raj 2014).

Results

We summarize results by reporting effects of varying $\bar{n}$, k, δ, and τ^2, one at a time, while holding all other conditions to a plausible value for ecological analysis (i.e., $\delta = 1$, $\bar{n} = 8$, $k = 55$, $\tau^2 = 1$; Appendix S2). Results using different parameter combinations are presented in Appendix S3.

Bias, coverage, and efficiency depended on the type of weighting scheme used. For example, analyses that were unweighted or based on sample size weighting never exhibited bias distinguishable from zero and yielded coverages that were close to the nominal level of 95% at all but the lowest values of k (Fig. 1). In contrast, analyses based on the inverse-variance weights were often biased and often yielded coverage that was far below the nominal level. Bias and coverage associated with the inverse-variance method depended greatly on the characteristics of the data upon which the meta-analyses were based.

Effect size

When there was no difference between treatment means ($\delta = 0$), the inverse-variance method exhibited no bias and coverage was close to the nominal level of 95%. However, as the true effect size, δ, increased, bias increased in magnitude (Fig. 1A), and coverage decreased (Fig. 1B) far below the nominal level (0.95). At small effect sizes ($\delta < 0.8$), the inverse-variance weight yielded the most efficient estimator (Fig. 1C); however, for larger effect sizes ($\delta > 1.0$), bias was sufficiently large that the inverse-variance method was no longer most efficient. The sample size-based estimator was always more efficient than the unweighted estimator, although these differences were small (Fig. 1C).

The number of studies

The number of studies in a meta-analysis (k) had little effect on bias but did affect coverage and efficiency (Fig. 1D–F). Increasing the number of studies reduced the size of the confidence intervals. Thus, in the presence of bias (i.e., reduced accuracy), the increased precision (due to increased k) of the estimated effect led to reduced coverage (Fig. 1D, E).

Sample size

Performance of the inverse-variance method also was sensitive to sample size ($\bar{n}$). Bias was greater, and coverage was lower, when sample sizes were small because the effect size was less accurately estimated (Fig. 1G–I), and thus a larger error in the estimate of the effect was propagated into error in the weights. For example, when $k = 55$ and $\delta = 1$, decreasing the average sample size from 25 to 4 increased the magnitude

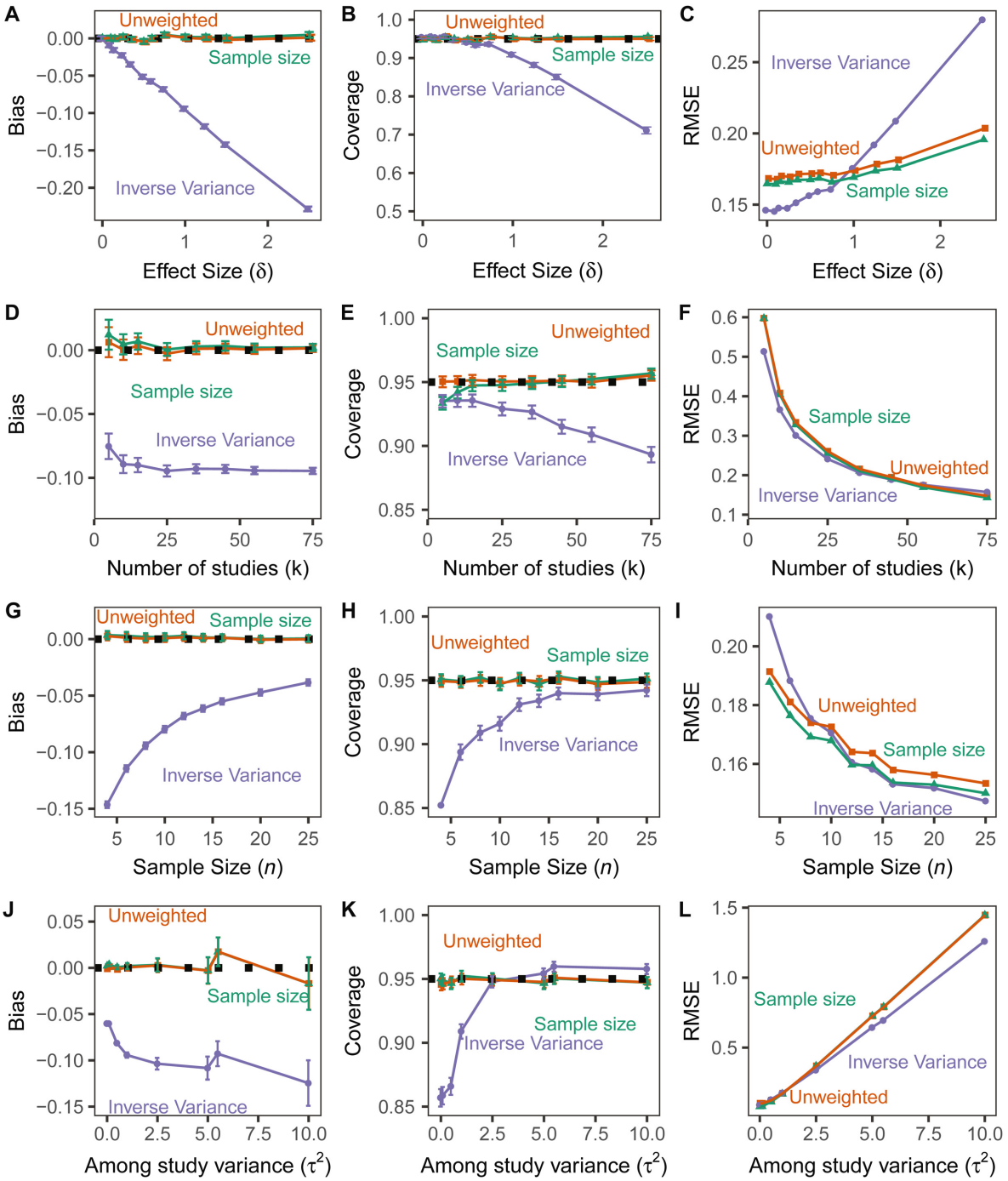

Fig. 1. Bias (A, D, G, and J), coverage (B, E, H, and K), and efficiency (C, F, I, and L) for three weighting schemes (unweighted, sample size, and inverse-variance). Unless varied on the *x*-axis, 55 studies were included in the meta-analysis ($k = 55$), $\delta = 1$, $\sigma = 1$, and $\bar{n} = 8$. Results are given for meta-analyses conducted with inverse-variance weighting (purple circles), sample size weighting (orange squares), and unweighted meta-analyses (green triangles). The dashed line in Panels A, D, G, and J indicates a bias of 0; the dashed line in Panel B indicates 95% coverage. Uncertainty in estimated average bias and coverage is represented with 95% confidence intervals based on 10,000 simulations. Efficiency is represented as the root mean squared error (RMSE).

of the bias by 3.8-fold (Fig. 1G). This change in bias reduced coverage from 94% to 85%. (Fig. 1H). When average sample size was large ($\bar{n} > 20$), the inverse-variance weighted estimate was most efficient, but for smaller sample sizes ($\bar{n} < 8$), the sample size weighted estimate was more efficient (Fig. 1I).

Among-study variance

The magnitude of the among-study variance also affected bias, coverage, and efficiency of the inverse-variance weighted estimator (Fig. 1J–L): Bias increased as among-study variance increased, but coverage also increased (eventually exceeding 95% at large values of among-study variance, Fig. 1K). At high levels of among-study variance, inverse-variance weighting was the most efficient (Fig. 1L), despite being the most biased (Fig. 1J).

Interactions

The above summaries ignore interactions among the parameters (but see Appendix S3). For example, the effect of the number of studies (k) depended upon sample size ($\bar{n}$). When $\bar{n}$ was small, bias was large (Fig. 1A) and thus increasing k led to reduced coverage (because greater precision of a biased estimate led to lower coverage). However, when $\bar{n}$ was large, bias was negligible and thus increasing k did not have as large an effect on coverage (e.g. compare Panel I in Appendix S3: Figs. S1, S3 at low and high $\bar{n}$). In the worst-case scenarios for bias and coverage, bias can be as large as −0.4 (16%; Panel G in Appendix S3: Fig. S9) and coverage can be as low as 3% (Panel H in Appendix S3: Fig. S3).

Discussion

Meta-analyses based on Hedges' d are known to be biased for two reasons: (1) The metric itself is biased, which is why Hedges (1981) introduced the small sample correction, J, into Eq. 1; and (2) the choice of a weighting scheme can introduce additional bias (e.g., Hedges 1982). In our simulations, the first form of bias was minor as the unweighted approach yielded approximately unbiased estimates (Fig. 1). This is expected because Hedges' (1981) correction adequately reduces this source of bias. In contrast, our study demonstrated that, for scenarios common to ecological meta-analyses, the use of Hedges' d in combination with the inverse-variance weighting scheme (Eq. 1 in combination with Eq. 3) leads to biased estimates of pooled effect sizes and low coverage of the associated confidence intervals. The low coverage could have resulted from the bias alone, or in combination with inappropriately narrow confidence intervals. Our analyses (Appendix S3: Fig. S1) suggest that the effect is largely driven by the bias. The bias results from the correlation between the estimated effect sizes and their estimated variances, which stems, in part, from error in the estimate of the effect that is propagated into error in the within-study variance and thus the weight. This problem is worse when sample sizes of the original studies are low (thus increasing the magnitude of the shared error) and when the true effect sizes are large (Fig. 1A, G).

Understanding why there is an absence of bias when $\delta = 0$ helps elucidate why the inverse-variance method leads to bias under other conditions. Due to sampling error, the estimated effect size from a study will deviate from its true effect size, either because of variation in the sampled means (the numerator in Eq. 1) or error in the estimate of the sample variance (i.e., the denominator in Eq. 1). That error gets propagated into the weights based on the use of the estimate of the within-study variance, v_i, in Eq. 2. Thus, when δ is large, studies that underestimated the true effect size would be weighted more heavily than studies that overestimated the true effect size, leading to bias (Fig. 1A). However, when $\delta = 0$, the errors about δ would be symmetrical, eliminating the bias. Importantly, we use meta-analysis in large part to estimate δ under the hypothesis that it is non-zero; thus, it seems counterproductive to apply a method that is only unbiased in the absence of an effect (i.e., when $\delta = 0$).

We found substantial bias (e.g., on the order of 10% of the true mean effect) and substantial under-coverage (often < 90% instead of 95%) using a random-effects model and data with characteristics often seen in ecological meta-analyses. The bias of Hedges' d has been previously recognized (Hedges 1982, 1983, Sánchez-Meca and Marín-Martínez 1998, Van Den Noortgate and Onghena 2003, Sánchez-Meca and Marín-Martínez 2008), yet this bias remains largely unknown in the ecological literature. Hedges

(1982) first highlighted the problem but concluded that estimation of the overall effect was quite accurate for a fixed-effects model with a range of true effects from 0.25 to 1.25 and for sample sizes ≥10, and Hedges (1983) inferred indirectly that bias would be low in a random-effect model, unless among-study variance was large. Of course, ecologists typically apply random-effects models, synthesize studies with fewer than 10 replicates per treatment (Hillebrand and Gurevitch 2014), and conduct meta-analyses characterized by very large among-study variance (Senior et al. 2016).

Importantly, although bias arises from effects of the within-study variance term on the weights, this effect is produced by sampling error as well as true study-specific variation in effects (i.e., by the among-study variance). This arises from the use of the observed effect (d_i) in the estimation of the within-study variance, an issue that was emphasized by Böhning et al. (2002) who argued for using a single (i.e., common) estimate of the expected effect size in Eq. 3. Similar logic led Hedges (1982) to suggest the sample size weighting expression in Eq. 4, which is based on Eq. 3 but with δ fixed at zero. Importantly, neither Hedges' nor Bohning et al.'s recommendations have been embraced in the ecological literature. A downside to using weights based on Eq. 4 is that they will understate the within-study variance when average δ is substantially different from zero.

Recognizing that d_i deviates from the overall effect (δ) due to sampling error (reflected in v_i) and because δ_i deviates from δ, due to among-study variation, helps explain why increasing τ^2 had little effect on bias (Fig. 1J). On the one hand, it might seem that high τ^2 should swamp v_i, essentially leading to an unweighted analysis. This does not occur because as τ^2 increases, the extent to which the v_i vary among the studies also increases. This occurs because d_i influences v_i through Eq. 3, and increased variation in δ_i leads to increased variation in d_i. Only when sample size is large, will the within-study variance become uniformly small across all studies, causing the among-study variance to dominate in the calculation of weights and eliminate the bias (increasing n always reduced bias using the inverse-variance method: Fig. 1G; Appendix S3). However, weighting by Eq. 3 only avoids bias when sample sizes are large enough that the resulting approach is essentially an unweighted analysis.

Hedges' d is not the only effect size with a correlation between the estimates of the effect size and variance. This correlation also exists for the correlation coefficient, r, another common effect size metric used in ecological meta-analyses, in which the variance of r includes an estimate of r. Many researchers transform r to a Z-statistic for meta-analysis, often citing methods papers that recommend this approach (e.g., Rosenberg et al. 2000, Borenstein et al. 2009, Koricheva et al. 2013). However, these papers seldom refer to the original rationale for this transformation: that is, to remove the bias caused by the correlation between r and its variance (Fisher 1921). The estimated variance of Z is a function only of the sample size. In addition to the possible solutions proposed by Hedges (1983) and Böhning et al. (2002), Van Den Noortgate and Onghena (2003) suggested using an empirical Bayes estimator of within-study variance, which reduces, but does not completely eliminate, the bias. Doncaster and Spake (2017) recommend an adjusted precision weight, although in their study case the associated bias is due to error in estimating v_i with small sample size independent of the correlation between effect size and variance estimate.

We suspect that other metrics and similar approaches may suffer from related problems. For example, Fletcher and Dixon (2011) showed that weighted regression (and by extension, probably meta-regression) has lower coverage than unweighted methods, partly due to problems with the estimation of weights. We suggest that future meta-analyses carefully consider whether weights are correlated with effect size estimates and, if so, consider using alternative approaches that avoid this, such as sample size-based approaches (e.g., Hungate et al. 2009) or unweighted analyses, although unweighted analyses pose other issues in some cases (e.g., meta-analyses of absolute values [or magnitudes] will be biased if unweighted: Morrissey 2016). Further statistical research is needed to explore weighting alternatives for ecological meta-analyses.

Despite prior concerns about the suitability of Hedges' d based on conceptual grounds (Osenberg et al. 1997, 1999), it remains one of the most commonly used measures of effect in ecological meta-analyses (e.g., Hillebrand and Gurevitch

2014) and is typically paired with the inverse-variance weighting scheme (Eq. 1 in combination with Eq. 3). Given prevalent bias under ecologically relevant conditions, we recommend that, if Hedges' d is used, weights should not be based on within-study variance estimates from Eq. 3. Our results further suggest relatively little difference between unweighted analyses and those based on sample sizes (Eq. 4) under the conditions of our simulations, both of which performed relatively well. This similarity likely arose due to small variation in weights (i.e., sample sizes), leading to an approximately unweighted analysis. However, there may be situations in which sample sizes vary more and thus weighting based upon sample sizes (Eq. 4) would produce greater benefits with respect to efficiency. Alternatively, an average standardized mean difference could be obtained initially with a sample size or unweighted approach, and this average could then be applied to all studies (in place of d_i in Eq. 3) which reduces bias but produces a less efficient estimator (Sánchez-Meca and Marín-Martínez 1998). An additional benefit of sample size weights (or an unweighted analysis) over the inverse-variance approach to weighting is that it does not require estimates of among-replicate variation, which are often not reported in the ecological literature (Gurevitch and Hedges 1999, Lajeunesse and Forbes 2003).

Acknowledgments

We thank the Osenberg laboratory for helpful discussion. This research was supported, in part, by the U.S. Department of Energy, Office of Science, Biological and Environmental Research Program, under Award Number DE-SC-0010632, the National Science Foundation (DEB-1655426 and DEB-1655394), the Georgia Advanced Computing Resource Center, the Partnership for Ecosystem Research and Managemental Michigan State University (MSU) (with support from the Michigan DNR), the Quantitative Fisheries Center at MSU, and AgBioResearch of Michigan State University (SDP). This is QFC publication 2018-16.

Literature Cited

Böhning, D., U. Malzahn, E. Dietz, P. Schlattmann, C. Viwatwongkasem, and A. Biggeri. 2002. Some general points in estimating heterogeneity variance with the Dersimonian-Laird estimator. Biostatistics 3:445–457.

Borenstein, M., L. V. Hedges, J. P. T. Higgins, and H. R. Rothstein. 2009. Introduction to Meta-Analysis. John Wiley & Sons Ltd, Chichester, UK.

Cadotte, M. W., L. R. Mehrkens, and D. N. L. Menge. 2012. Gauging the impact of meta-analysis in ecology. Evolutionary Ecology 26:1153–1167.

Doncaster, C. P., and R. Spake. 2017. Correction for bias in meta-analysis of little replicated studies. Methods in Ecology and Evolution 9:634–644.

Dorai-Raj, S. 2014. Binomial confidence intervals for several parameterizations. R package. https://cran.r-project.org/web/packages/binom/index.html

Fisher, R. A. 1921. On the 'probable error' of a coefficient of correlation deduced from a small sample. Metron 1:1–32.

Fletcher, D., and P. M. Dixon. 2011. Modelling data from different sites, times or studies: weighted vs. unweighted regression. Methods in Ecology and Evolution 3:168–176.

Gurevitch, J., P. S. Curtis, and M. H. Jones. 2001. Meta-analysis in ecology. Advances in Ecological Research 32:199–247.

Gurevitch, J., and L. V. Hedges. 1993. Meta-analysis: combining the results of independent studies in experimental ecology. Pages 378–398 *in* S. Scheiner and J. Gurevitch, editors. The design and analysis of ecological experiments. Chapman & Hall, New York, New York, USA.

Gurevitch, J., and L. V. Hedges. 1999. Statistical issues in ecological meta-analyses. Ecology 80:142–1149.

Gurevitch, J., and J. Koricheva. 2013. Conclusions: past, present, and future of meta-analysis in ecology and evolution. Pages 426–432 *in* J. Koricheva, J. Gurevitch, and K. Mengersen, editors. Handbook of meta-analysis in ecology and evolution. Princeton University Press, Princeton, New Jersey, USA.

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

Hedges, L. V. 1981. Distribution theory for Glass's estimator of effect size and related estimators. Journal of Educational Statistics 6:107–128.

Hedges, L. V. 1982. Estimation of effect size from a series of independent experiments. Psychological Bulletin 92:490–499.

Hedges, L. V. 1983. A random effects model for effect sizes. Psychological Bulletin 93:388–395.

Hedges, L. V., and I. Olkin. 1985. Statistical methods for meta-analysis. Academic Press, Orlando, Florida, USA.

Hillebrand, H., and J. Gurevitch. 2014. Meta-analysis results are unlikely to be biased by differences in

variance and replication between ecological lab and field studies. Oikos 123:794–799.
Hungate, B. A., K. J. Van Groenigen, J. Six, J. D. Jastrow, Y. Luo, M. A. De Graaff, C. van Kessel, and C. W. Osenberg. 2009. Assessing the effect of elevated carbon dioxide on soil carbon: a comparison of four meta-analyses. Global Change Biology 15:2020–2034.
Johnson, B. T., and T. B. Huedo-Medina. 2013. Meta-Analytic Statistical Inferences for Continuous Measure Outcomes as a Function of Effect Size Metric and Other Assumptions. AHRQ Publication No. 13-EHC075-EF.
Koricheva, J., and J. Gurevitch. 2014. Uses and misuses of meta-analysis in plant ecology. Journal of Ecology 102:828–844.
Koricheva, J., J. Gurevitch, and K. Mengersen. 2013. Handbook of Meta-analysis in Ecology and Evolution. Princeton University Press, Princeton, New Jersey, USA.
Lajeunesse, M. J. 2010. Achieving synthesis with meta-analysis by combining and comparing all available studies. Ecology 91:2561–2564.
Lajeunesse, M. J. 2015. Bias and correction for the log response ratio in ecological meta-analysis. Ecology 96:2056–2063.
Lajeunesse, M. J., and M. R. Forbes. 2003. Variable reporting and quantitative reviews: comparison of three meta-analytical techniques. Ecology Letters 6:448–454.
Levin, S. A. 1992. The problem of pattern and scale in ecology. Ecology 73:1943–1967.
Lortie, C. J. 2014. Formalized synthesis opportunities for ecology: systematic reviews and meta-analyses. Oikos 123:897–902.
Morrissey, M. B. 2016. Meta-analysis of magnitudes, differences and variation in evolutionary parameters. Journal of Evolutionary Biology 29:1882–1904.
Nakagawa, S., and E. S. A. Santos. 2012. Methodological issues and advances in biological meta-analysis. Evolutionary Ecology 26:1253–1274.
Osenberg, C. W., O. Sarnelle, and S. D. Cooper. 1997. Effect size in ecological experiments: the application of biological models in meta-analysis. American Naturalist 150:798–812.
Osenberg, C. W., O. Sarnelle, S. D. Cooper, and R. D. Holt. 1999. Resolving ecological questions through meta-analysis: goals, metrics, and models. Ecology 80:1105–1117.
R Core Team. 2018. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
Rosenberg, M. S., D. C. Adams, and J. Gurevitch. 2000. MetaWin: statistical software for meta-analysis. Version 2.0. Sinauer Associates, Sunderland, Massachusetts, USA.
Sánchez-Meca, J., and F. Marín-Martínez. 1998. Weighting by inverse variance or by sample size in meta-analysis: a simulation study. Educational and Psychological Measurement 58:211–220.
Sánchez-Meca, J., and F. Marín-Martínez. 2008. Confidence intervals for the overall effect size in random-effects meta-analysis. Psychological Methods 13: 31–48.
Senior, A. M., C. E. Grueber, T. Kamiya, M. Lagisz, K. O'Dwyer, E. S. A. Santos, and S. Nakagawa. 2016. Heterogeneity in ecological and evolutionary meta-analyses: its magnitude and implications. Ecology 97:3293–3299.
Van Den Noortgate, W., and P. Onghena. 2003. Estimating the mean effect size in meta-analysis: bias, precision, and mean squared error of different weighting methods. Behavior Research Methods, Instruments, & Computers 35:504–511.
Viechtbauer, W. 2010. Conducting meta-analyses in R with the metafor package. Journal of Statistical Software 36:1–48.
Whittaker, R. J. 2010. Meta-analyses and mega-mistakes: calling time on meta-analysis of the species richness–productivity relationship. Ecology 91:2522–2533.

Supporting Information

Additional Supporting Information may be found online at: http://onlinelibrary.wiley.com/doi/10.1002/ecs2.2419/full

Appendix S1: Sample-size based weights

Bias in meta-analyses using Hedges's *d*
Hamman, EA, P Pappalardo, JR Bence, S Peacor, and CW Osenberg.
Ecosphere

In our analyses we included an approximation for the variance of Hedges's *d* (Eq. 4 in main text) based on the sample-size estimator proposed by Hedges (1982):

$$v_i = \left(1 - \frac{3}{4(n_{i,T}+n_{i,C}-2)-1)}\right)^2 \left(\frac{n_{i,T}+n_{i,C}-2}{\frac{n_{i,T}*n_{i,C}}{n_{i,T}+n_{i,C}}\left(n_{i,T}+n_{i,C}-4\right)}\right) \qquad \text{(Eq. S1-1)}$$

Another equation, proposed by Hedges and Olkin (1985), also uses experimental sample size to approximate the variance of *d*.

$$v_i = \frac{n_{i,T}+n_{i,C}}{n_{i,T}*n_{i,C}} \qquad \text{(Eq. S1-2)}$$

At large sample sizes, weights based on these two approaches are very similar. However, for small sample sizes Eq. S1-1 (Eq. 4 in main text), better approximates the variance of *d*. In addition, use of Eq. S1-2 yields weights that are less variable among studies with different sample sizes, so the effect of using Eq. S1-2 should yield results that are intermediate to those obtained using Eq. S1-1 or using an unweighted approach. Because our simulations demonstrated very little difference between unweighted analyses and those based on Eq. S1-1, we conclude that results based on Eq. S1-2 would have deviated little from the results we report in the main text based on Eq. S1-1.

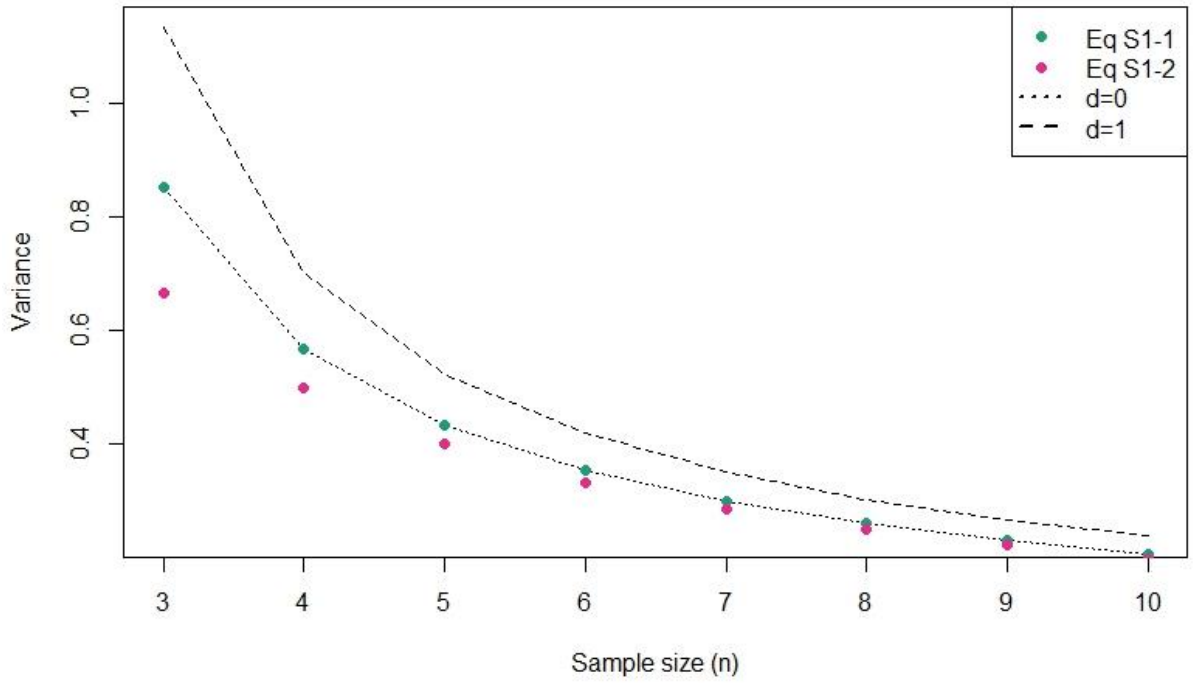

Figure S1-1: Variance estimates of Hedges's *d* using Eq. A1 (green dots, and estimate used in sample-size weight in simulations), Eq. S1-2 (pink dots). When the true standardized difference is 0 and it is estimated accurately (small dashes), there is no difference between the estimated variance using Eq. S1-1 and Eq. 3 (in the main text). However, when the true *d* is non-zero and is estimated accurately (long dashes), both sample-size based formulae for the variance result in underestimates of the variance of d, although Eq. S1-1 is more accurate.

References:

Hedges, L. V. 1982. Estimation of effect size from a series of independent experiments.

Psychological Bulletin **92**:490-499.

Hedges, L. V., and I. Olkin. 1985. Statistical methods for meta-analysis. Academic Press,

Orlando, Florida, USA

Appendix S2: Justification for Parameter Estimates

Bias in meta-analyses using Hedges's *d*
Hamman, EA, P Pappalardo, JR Bence, S Peacor, and CW Osenberg.
Ecosphere

For the results in Figure 1, we aimed to represent characteristics of ecological meta-analyses that would allow us to explore the bias present in Hedges' *d* under conditions typical of ecological meta-analyses. Below we detail for each variable how we chose the parameter range likely to represent an ecological meta-analysis, and justify the value for each parameter that we used to explore effects of variation in the other parameters.

Mean experimental sample size ($\bar{n}$)

In our simulations, we used $\bar{n}$ = {4, 6, 8, 10, 12, 14, 16, 20, 25}, and focused on the results based on $\bar{n}$ = 8. In an extensive survey of ecological meta-analyses, Hillebrand and Gurevitch (2014: in their Figure 1C) report that both field and lab studies had an approximate median of 8 replicates, with a range from 1-40 (although most studies fell in the range of 5-10). To characterize the distribution of n_is (around a given $\bar{n}$) we fit a negative binomial distribution to Figure 14 of Levin 1992, which yielded a dispersion parameter of θ=1.55; we used this estimate in our simulations to obtain the variation in sample sizes among studies in a meta-analysis.

Number of studies included in a meta-analysis (k)

In our simulations, we used *k* = {5, 10, 15, 25, 35, 45, 55, 75, 100, 125}, and focused on the results based on k = 55, which is similar to the range used in Sanchez-Meca and Marin-Martinez's (2008) simulations (i.e., they used 5-100). We used k=55 studies as our reference for the main figure because it was closest to the midpoint of these ranges. Unpublished data by P. Pappalardo found this range was common for ecological meta-analyses, although the distribution of ecological studies was skewed towards smaller numbers of studies.

Effect size (δ)

In our simulations, we used $\delta\sigma$ = {0, 0.1, 0.15, 0.25, 0.35, 0.5, 0.6, 0.75, 1, 1.25, 1.5, 2.5}, and because we assumed σ^2=1.0, this range of absolute differences is the same as the range of standardized differences. We focused on the results based on $\delta\sigma = \delta = 1$. Moller and Jennions (2002) surveyed ecological, evolutionary, and physiological meta-analyses and found that d ranged from 0.22 to 1.70, with disciplinary averages ranging from 0.5 to 0.84. Hillebrand and Gurevitch (2014) surveyed 865 experiments on grazers and found |d| ranged from 0 to 12, with most studies having |d|<3. Thus, our simulations occurred over an ecologically relevant range of effect sizes and used a representative value of 1.

Among-study variation (τ^2)

In our simulations, we used τ^2 = {0, 0.1, 0.5, 1, 2.5, 5, 10}, and focused on the results based on τ^2=1. To choose among-study variation that was relevant to ecological studies, we relied on the review by Senior et al. (2010), who summarized estimates of I^2 in ecological and evolutionary

meta-analyses. I^2 represents the proportion of total variance due to variation among studies (i.e., the proportion of the variation in d's that cannot be attributed to sampling error). They found a wide range of I^2 (from 0 to 100%), with a mean of 85%. Our values for τ^2 (with σ^2=1), reflect this range.

References:

Hillebrand, H. and J. Gurevitch. 2014. Meta-analysis results are unlikely to be biased by differences in variance and replication between ecological lab and field studies. Oikos, **123**:794–799.

Levin, S.A. 1992. The problem of pattern and scale in ecology: The Robert H. MacArthur award lecture. Ecology **73**(6):943-1967.

Moller, A.P. and M.D. Jennions. 2002. How much variance can be explained by ecologists and evolutionary biologists. Oecologia **132**(4): 492-500.

Sanchez-Meca, J. and F. Marin-Martinez. 2008. Confidence Intervals for the Overall Effect Size in Random-Effects Meta-Analysis. Psychological Methods **13**(1): 31-48.

Senior, A.M., C.E. Grueber, T. Kamiya, M. Lagisz, K. O'Dwyer, E.S.A. Santos, and S. Nakagawa. 2016. Heterogeneity in ecological and evolutionary meta-analyses: its magnitude and implications. Ecology **97**(12): 3293-3299.

Appendix S3: Expanded Results

Bias in meta-analyses using Hedges's *d*
Hamman, EA, P Pappalardo, JR Bence, S Peacor, and CW Osenberg.
Ecosphere

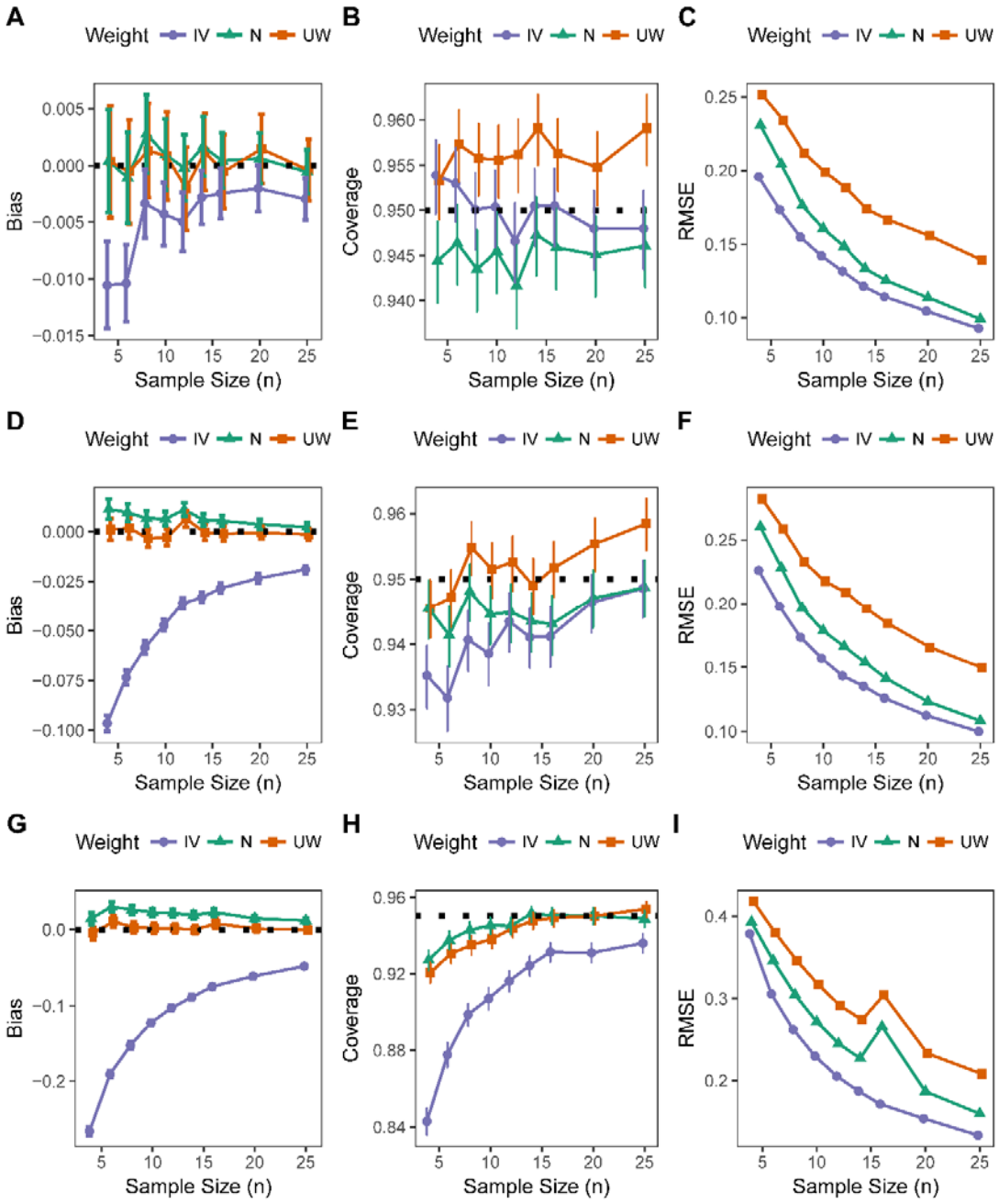

Figure S3-1: Bias, coverage and RMSE for simulations where among-study variance, τ^2=0, and the number of studies in the meta-analysis, k=10. Panels A-C represent simulations where δ=0.1, D-F simulations where δ=1, and G-I where δ=2.5. Purple circles represent meta-analyses conducted with inverse-variance weighting, orange squares sample size weighting, and green triangles unweighted meta-analyses. The dashed lines in Panels A, D, and indicate a bias of 0; the dashed lines in Panels B, E, and H indicate 95% coverage. Variability in bias and coverage is represented with 95% confidence intervals based on 10,000 simulations. Efficiency is represented as the root mean squared error (RMSE).

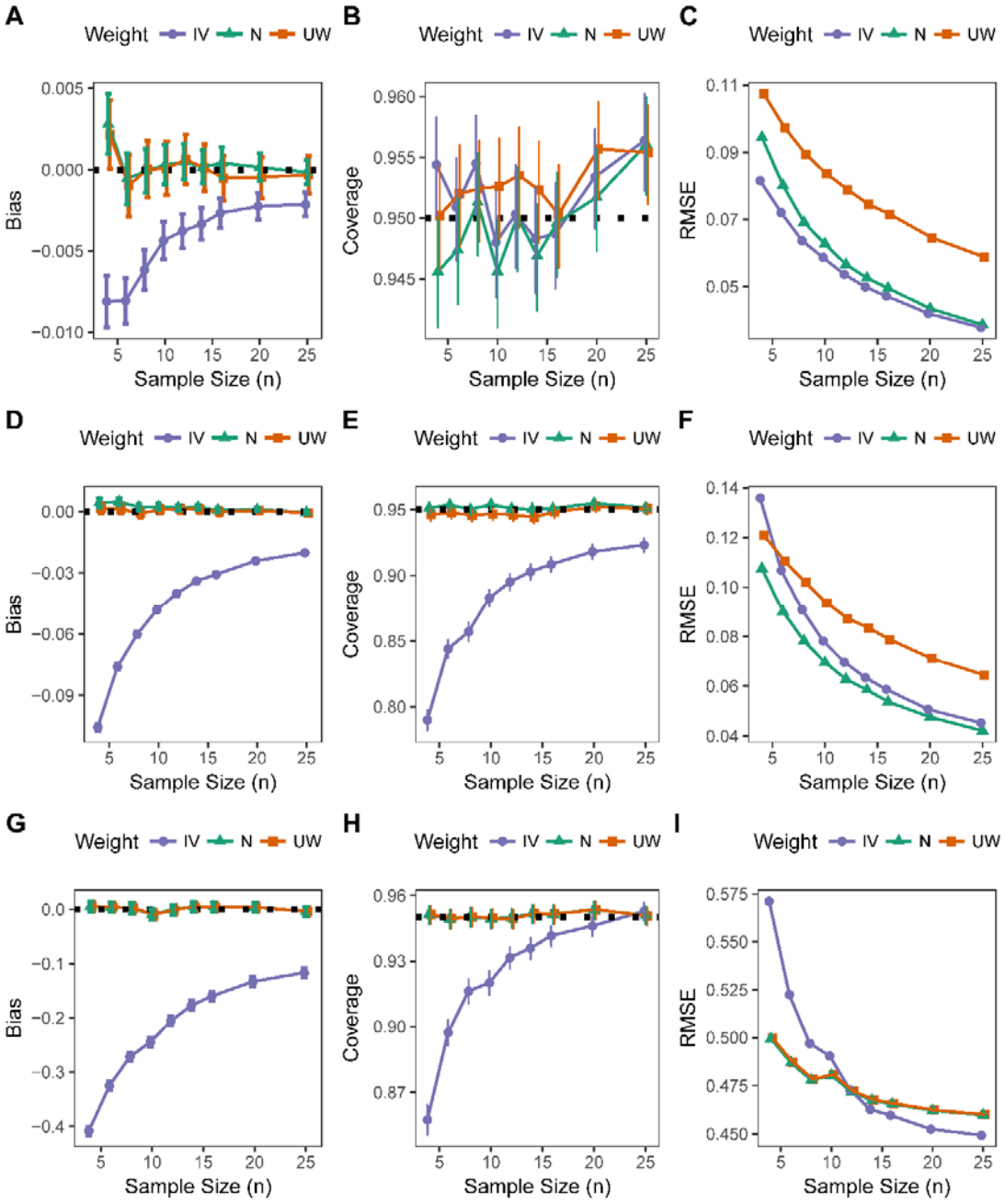

Figure S3-2: Bias, coverage and RMSE for simulations where among-study variance, τ^2=0, and the number of studies in the meta-analysis, k=55. Panels A-C represent simulations where δ=0.1, D-F simulations where δ=1, and where G-I δ=2.5. Purple circles represent meta-analyses conducted with inverse-variance weighting, orange squares sample size weighting, and green triangles unweighted meta-analyses. The dashed lines in Panels A, D, and indicate a bias of 0; the dashed lines in Panels B, E, and H indicate 95% coverage. Variability in bias and coverage is represented with 95% confidence intervals based on 10,000 simulations. Efficiency is represented as the root mean squared error (RMSE).

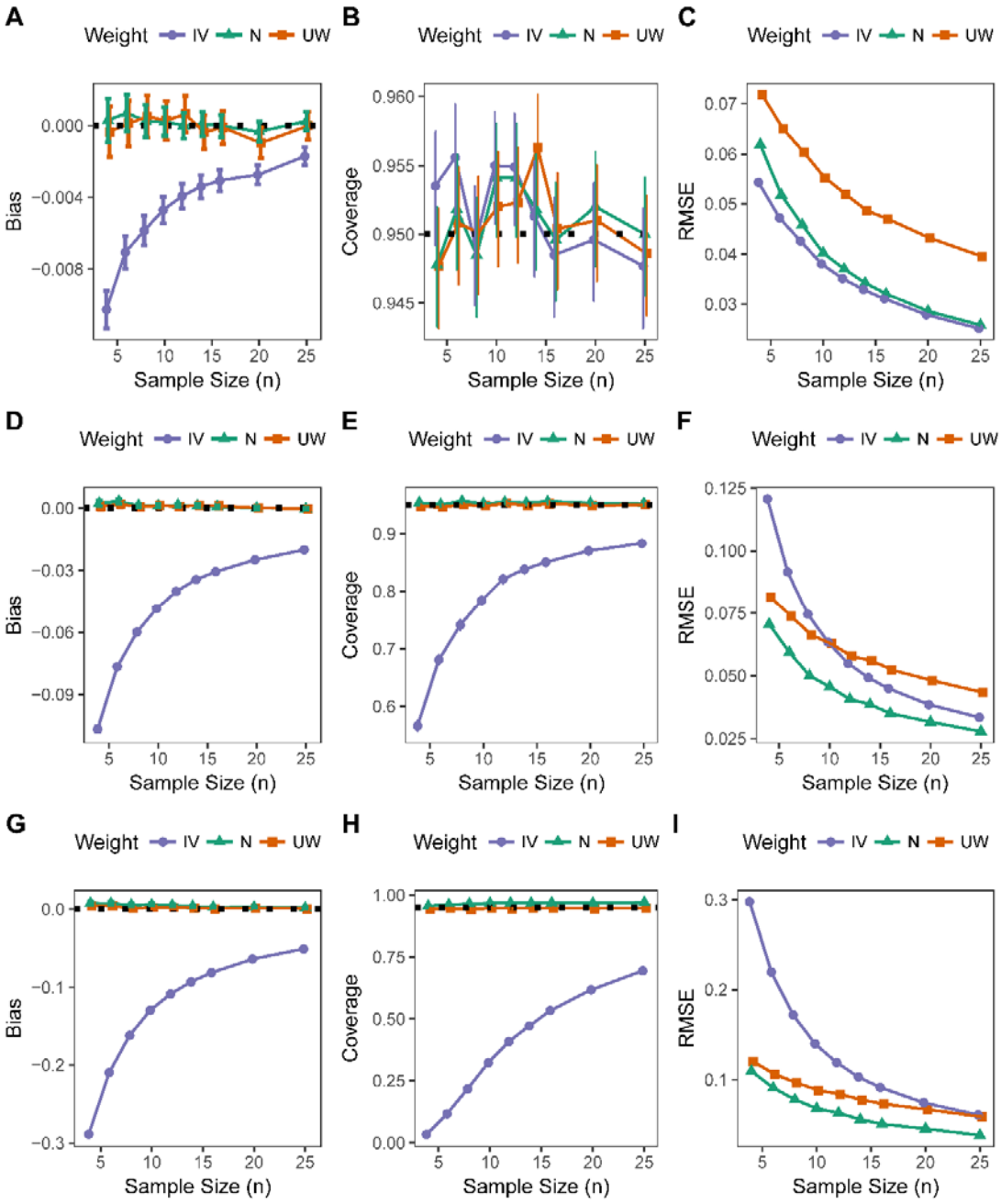

Figure S3-3: Bias, coverage and RMSE for simulations where among-study variance, τ^2=1, and the number of studies in the meta-analysis, k=125. Panels A-C represent simulations where δ=0.1, D-F simulations where δ=1, and G-I where δ=2.5. Purple circles represent meta-analyses conducted with inverse-variance weighting, orange squares sample size weighting, and green triangles unweighted meta-analyses. The dashed lines in Panels A, D, and indicate a bias of 0; the dashed lines in Panels B, E, and H indicate 95% coverage. Variability in bias and coverage is represented with 95% confidence intervals based on 10,000 simulations. Efficiency is represented as the root mean squared error (RMSE).

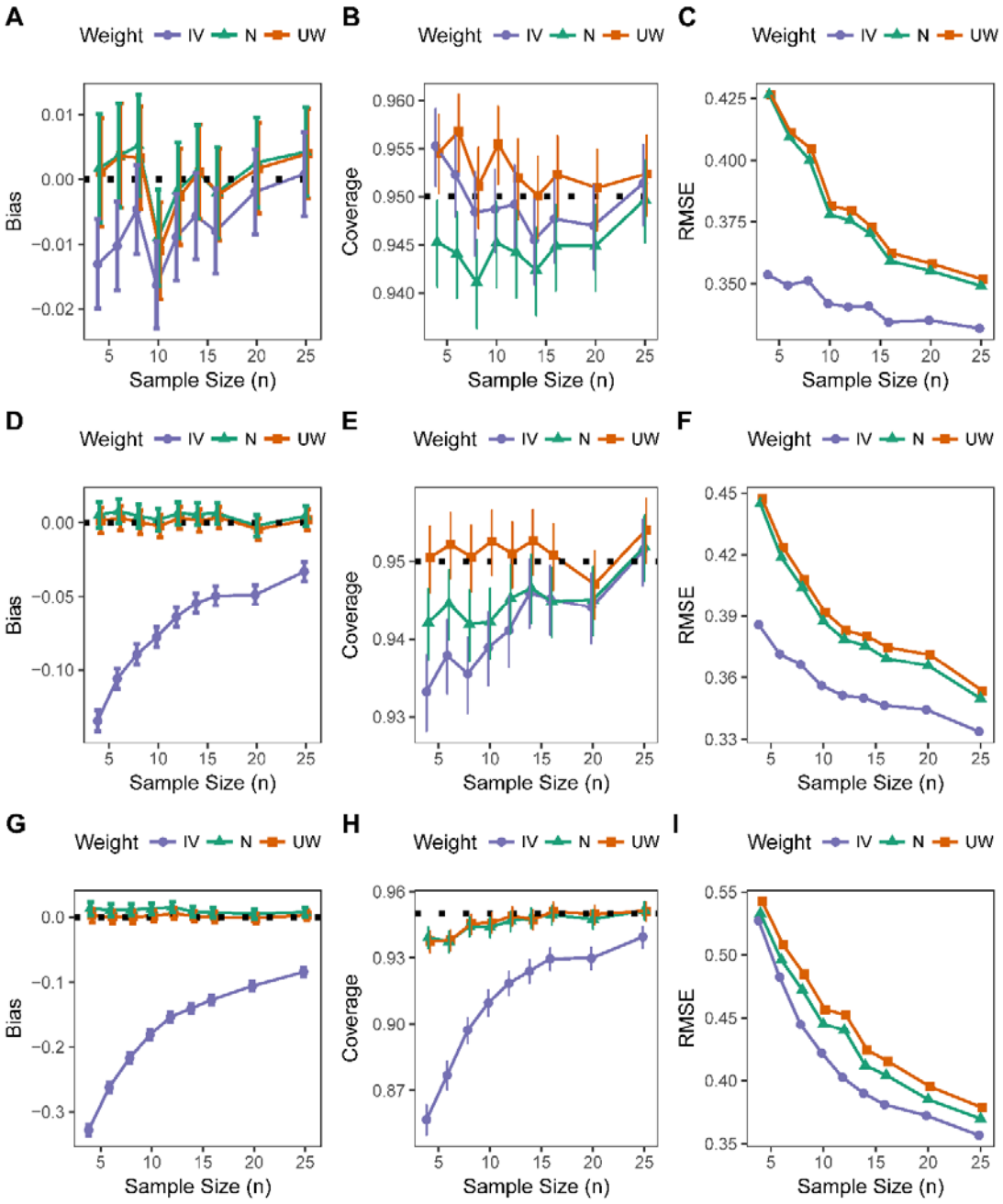

Figure S3-4: Bias, coverage and RMSE for simulations where among-study variance, τ^2=1, and the number of studies in the meta-analysis, k=10. Panels A-C represent simulations where δ=0.1, D-F simulations where δ=1, and G-I where δ=2.5. Purple circles represent meta-analyses conducted with inverse-variance weighting, orange squares sample size weighting, and green triangles unweighted meta-analyses. The dashed lines in Panels A, D, and indicate a bias of 0; the dashed lines in Panels B, E, and H indicate 95% coverage. Variability in bias and coverage is represented with 95% confidence intervals based on 10,000 simulations. Efficiency is represented as the root mean squared error (RMSE).

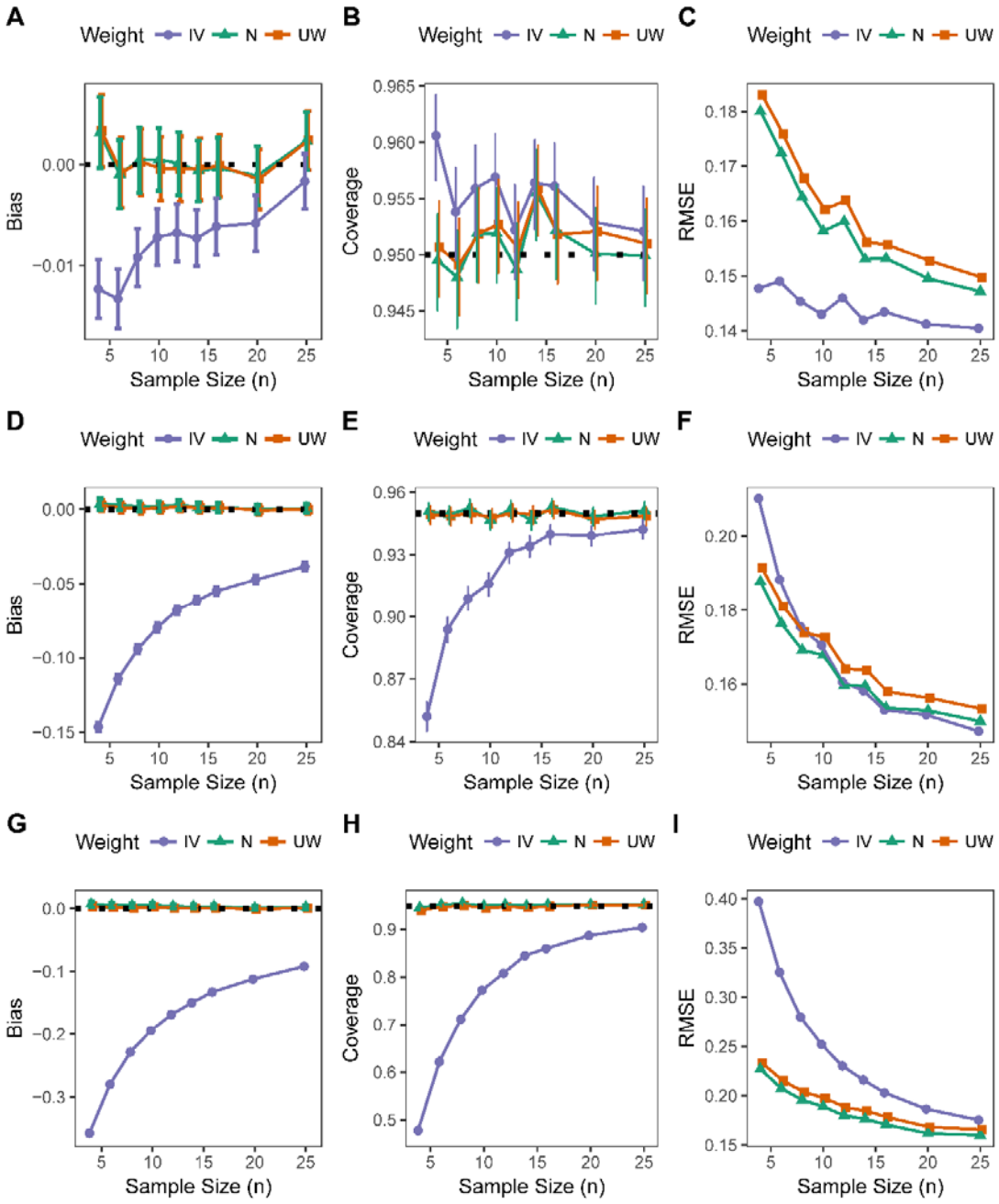

Figure S3-5: Bias, coverage and RMSE for simulations where among-study variance, τ^2=1, and the number of studies in the meta-analysis, k=55. Panels A-C represent simulations where δ=0.1, D-F simulations where δ=1, and G-I where δ=2.5. Purple circles represent meta-analyses conducted with inverse-variance weighting, orange squares sample size weighting, and green triangles unweighted meta-analyses. The dashed lines in Panels A, D, and indicate a bias of 0; the dashed lines in Panels B, E, and H indicate 95% coverage. Variability in bias and coverage is represented with 95% confidence intervals based on 10,000 simulations. Efficiency is represented as the root mean squared error (RMSE).

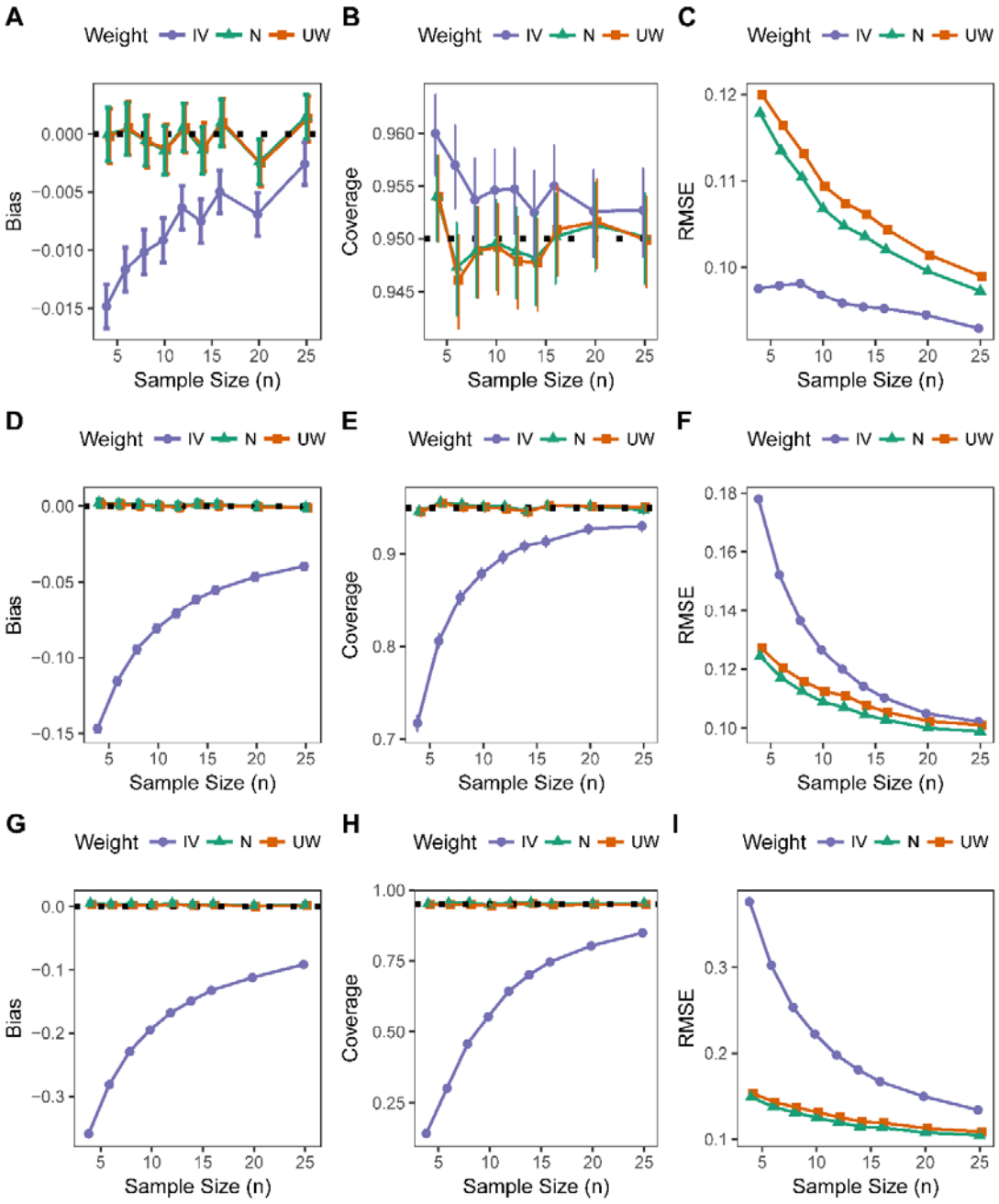

Figure S3-6: Bias, coverage and RMSE for simulations where among-study variance, τ^2=1, and the number of studies in the meta-analysis, k=125. Panels A-C represent simulations where δ=0.1, D-F simulations where δ=1, and G-I where δ=2.5. Purple circles represent meta-analyses conducted with inverse-variance weighting, orange squares sample size weighting, and green triangles unweighted meta-analyses. The dashed lines in Panels A, D, and indicate a bias of 0; the dashed lines in Panels B, E, and H indicate 95% coverage. Variability in bias and coverage is represented with 95% confidence intervals based on 10,000 simulations. Efficiency is represented as the root mean squared error (RMSE).

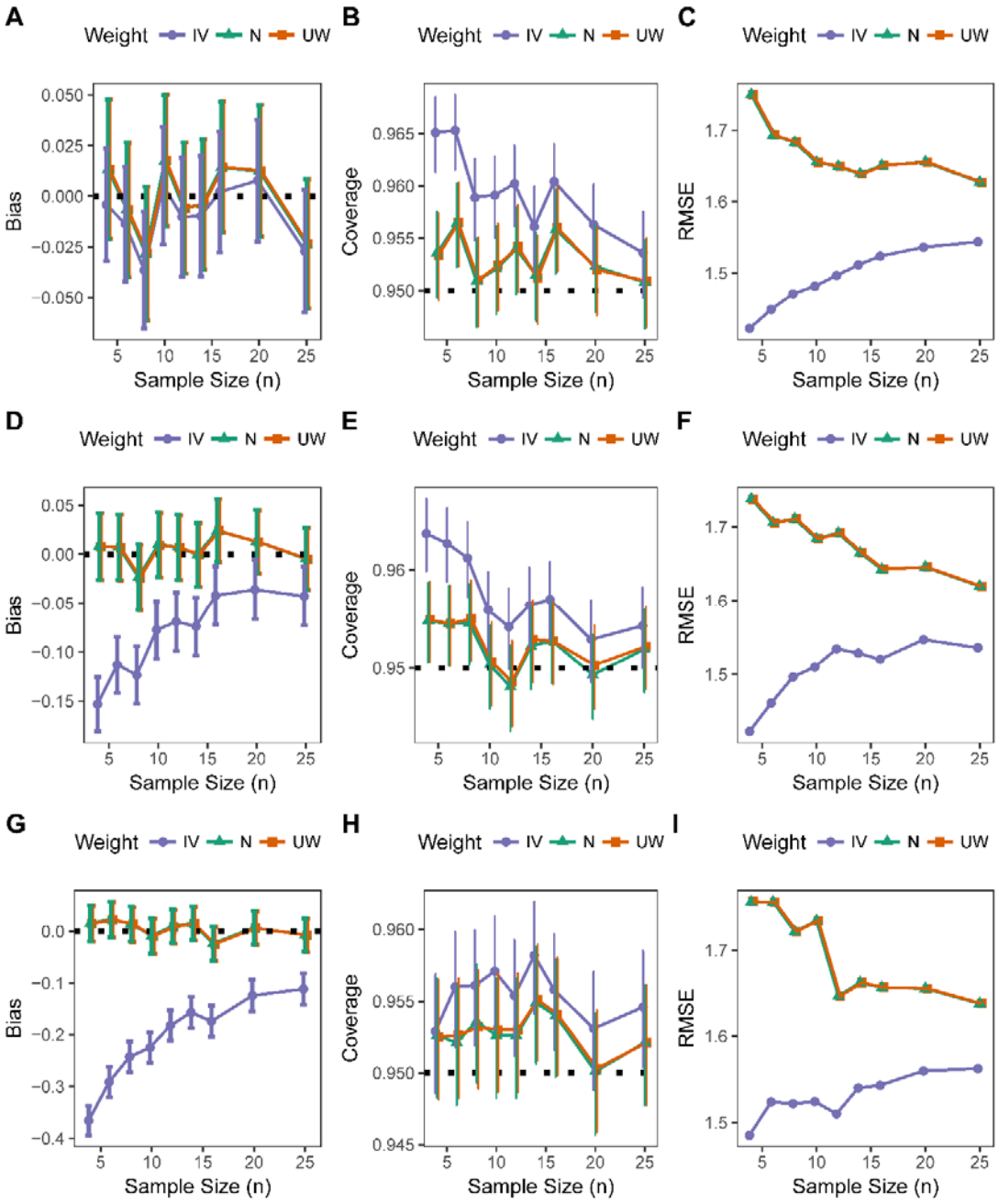

Figure S3-7: Bias, coverage and RMSE for simulations where among-study variance, τ^2=5, and the number of studies in the meta-analysis, k=10. Panels A-C represent simulations where δ=0.1, D-F simulations where δ=1, and G-I where δ=2.5. Purple circles represent meta-analyses conducted with inverse-variance weighting, orange squares sample size weighting, and green triangles unweighted meta-analyses. The dashed lines in Panels A, D, and indicate a bias of 0; the dashed lines in Panels B, E, and H indicate 95% coverage. Variability in bias and coverage is represented with 95% confidence intervals based on 10,000 simulations. Efficiency is represented as the root mean squared error (RMSE).

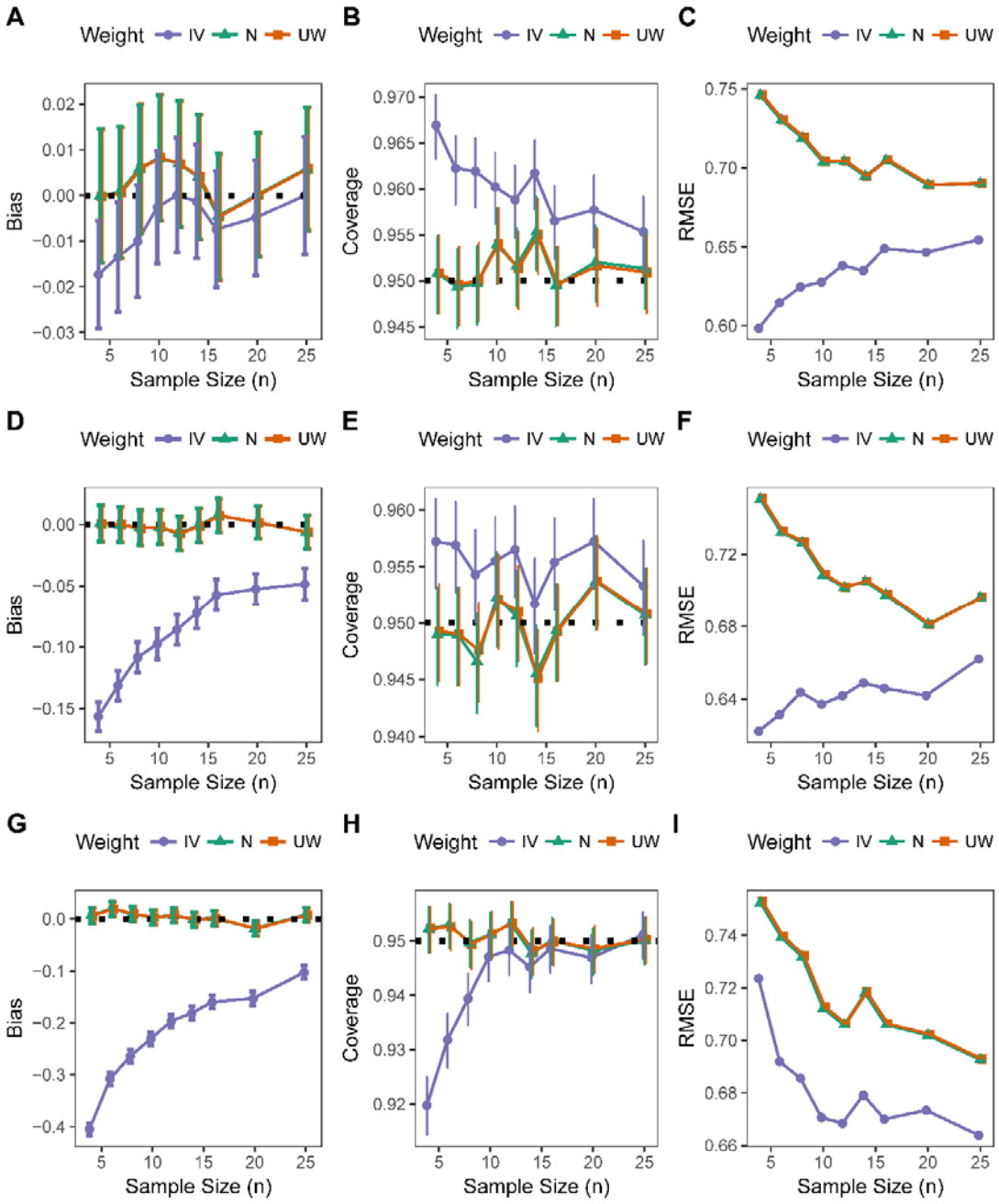

Figure S3-8: Bias, coverage and RMSE for simulations where among-study variance, τ^2=5, and the number of studies in the meta-analysis, k=55. Panels A-C represent simulations where δ=0.1, D-F simulations where δ=1, and G-I where δ=2.5. Purple circles represent meta-analyses conducted with inverse-variance weighting, orange squares sample size weighting, and green triangles unweighted meta-analyses. The dashed lines in Panels A, D, and indicate a bias of 0; the dashed lines in Panels B, E, and H indicate 95% coverage. Variability in bias and coverage is represented with 95% confidence intervals based on 10,000 simulations. Efficiency is represented as the root mean squared error (RMSE).

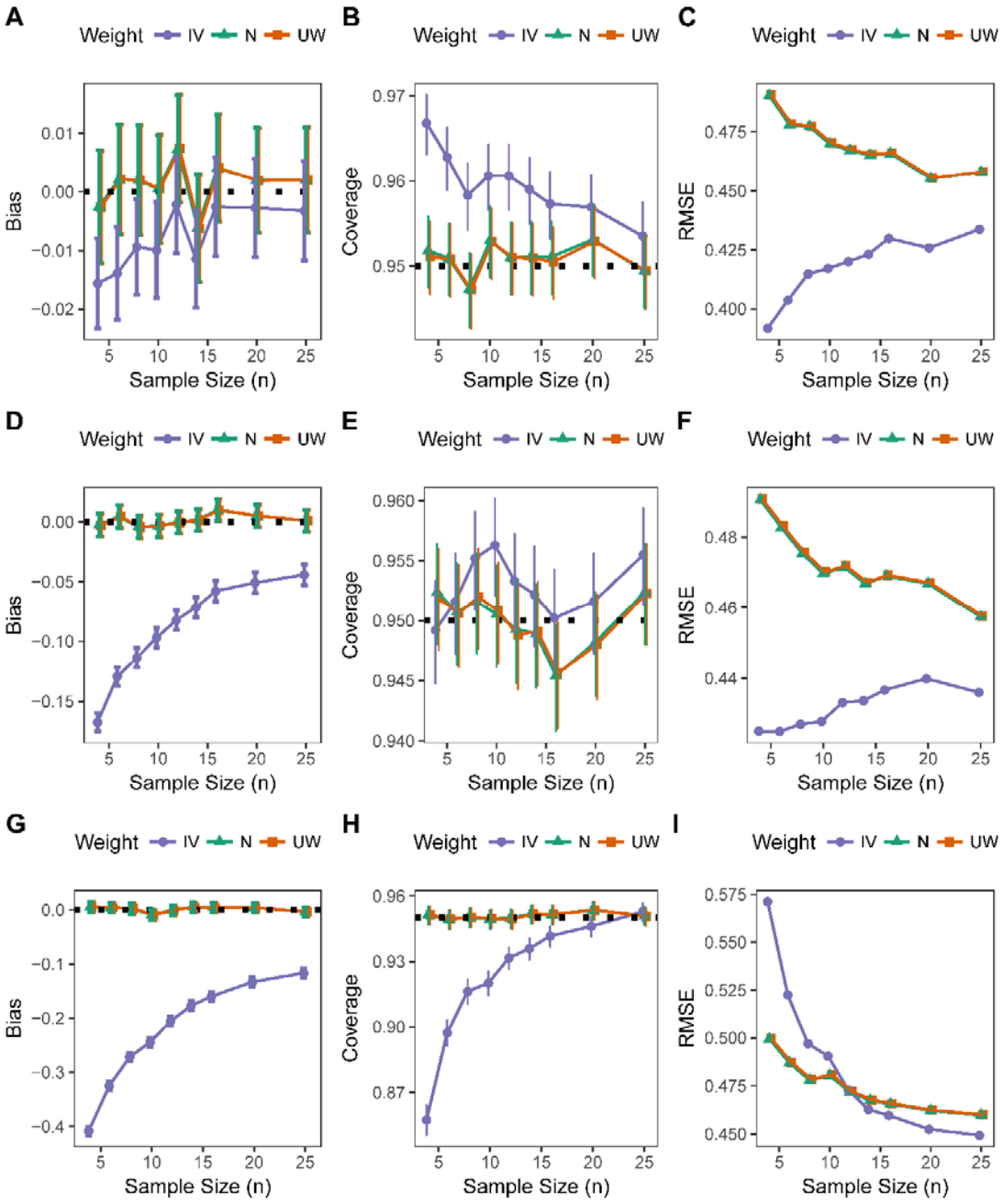

Figure S3-9: Bias, coverage and RMSE for simulations where among-study variance, τ^2=5, and the number of studies in the meta-analysis, k=125. Panels A-C represent simulations where δ=0.1, D-F simulations where δ=1, and G-I where δ=2.5. Purple circles represent meta-analyses conducted with inverse-variance weighting, orange squares sample size weighting, and green triangles unweighted meta-analyses. The dashed lines in Panels A, D, and indicate a bias of 0; the dashed lines in Panels B, E, and H indicate 95% coverage. Variability in bias and coverage is represented with 95% confidence intervals based on 10,000 simulations. Efficiency is represented as the root mean squared error (RMSE).