Research Synthesis Methods (2026), **17**: 389–402
doi:10.1017/rsm.2025.10062

Research Synthesis Methods

REVIEW

Reflections on the I-squared index for measuring inconsistency in meta-analysis

Julian P. T. Higgins [1,2] and José A. López-López [3,4]

[1]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK
[2]National Institute for Health and Care Research Applied Research Collaboration West at University Hospitals Bristol and Weston NHS Foundation Trust, Bristol, UK
[3]Department of Basic Psychology and Methodology, University of Murcia, Murcia, Spain
[4]Murcian Institute of Biomedical Research, IMIB-Arrixaca, Murcia, Spain

Corresponding author: Julian P. T. Higgins; Email: julian.higgins@bristol.ac.uk

Received: 15 November 2024; **Revised:** 5 September 2025; **Accepted:** 13 October 2025; published online 29 December 2025

Keywords: combining information; evidence synthesis; heterogeneity; I-squared; inconsistency; meta-analysis

Abstract

The I-squared index was proposed in 2002 as a measure to help understand the consistency of study results in a meta-analysis. It was developed to overcome some of the limitations of existing measures, principally the chi-squared test for heterogeneity and the between-study variance as estimated in a random-effects meta-analysis. I-squared measures approximately the proportion of total variability in results that is due to true heterogeneity rather than random error; it is also conveniently interpreted as a measure of inconsistency in the results of the studies. The index has become extremely widely used, although it is often misinterpreted as an absolute measure of the amount of heterogeneity, which it is not. Here, we discuss the I-squared index and the different ways it can be defined, computed, and interpreted. We introduce a new interpretation of I-squared as a weighted sum of squares, which we propose may be helpful when setting up simulation studies. We discuss some of the extensions and repurposes that have been proposed for I-squared and offer some recommendations on the appropriate use of the index in practice.

Highlights

What is already known?

- The I-squared index is widely used in relation to the identification or description of heterogeneity in meta-analysis.
- The role of I-squared has been questioned, partly because it is often misinterpreted: It measures inconsistency in findings, which is not the same as variability in effect sizes.

What is new?

- We provide a technical overview and our reflections on the I-squared index, including its definition, computation, and interpretation, and we offer a new definition that might be particularly useful when designing simulation studies.

Potential impact for RSM readers

- Given the routine use of the I-squared index in diverse research fields, it is important that users have a proper understanding of what it can, and cannot, portray.

This article was awarded Open Materials badge for transparent practices. See the Data availability statement for details.

© The Author(s), 2025. Published by Cambridge University Press on behalf of The Society for Research Synthesis Methodology. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/rsm.2025.10062 Published online by Cambridge University Press

1. Introduction

Meta-analysis is the statistical integration of results from a set of studies considered to be sufficiently similar for the combined estimate to be meaningful. If the studies are very similar such that they can be assumed to have addressed the same research question, their results might be expected to differ only because of sampling error. If the studies are all very large, the variation among their results would then be small. However, in practice, studies tend to differ in aspects such as the sample characteristics, ways in which treatments or exposures are defined or measured, and ways in which outcomes are measured. As a result, some degree of additional variation among the results is to be expected.[1,2] Such variation beyond what can be attributed to chance alone is usually called *statistical heterogeneity* or (in statistical texts) simply *heterogeneity*. It may be measured using, for example, the standard deviation of true effects across studies on the scale in which the effect size is measured (e.g., as a mean, a mean difference, or (log) odds ratio).

The I^2 index is a descriptive statistic that is used in relation to heterogeneity among studies in a meta-analysis. It aims to quantify *inconsistency*, a term we use for a construct that is distinct from (statistical) heterogeneity. Inconsistency relates to the extent to which the statistical findings of the studies disagree with each other after taking into account sampling error. Interpretation of inconsistency measures such as the I^2 index does not depend on the scale on which the result is measured. As such, they cannot be used to infer the actual (or absolute) amount of heterogeneity in the effect sizes across studies. The I^2 index is typically expressed as a percentage, with a value of 0% indicating that the only source of variability among the results reported in the primary studies is sampling error (so there is no heterogeneity) and values approaching 100% indicating that nearly all variability is due to heterogeneity among the studies rather than sampling error.

The index was introduced under the nomenclature I^2 by Higgins and Thompson (HT) in 2002, motivated by a desire to find an alternative to either the standard χ^2 test for heterogeneity (whose interpretation depends importantly on the number of studies) or the between-study variance as estimated in a random-effects meta-analysis (whose interpretation depends importantly on the measurement scale). The I^2 index was promoted in a *BMJ* paper by the same authors, along with Altman and Deeks, the following year.[3,4] The notion of "proportion of total variance due to between-study variance" had previously been mentioned in passing by Takkouche and colleagues.[5] Reporting of the I^2 index in meta-analyses has increased year by year, extending far beyond the initial context for which it was initially proposed, and the measure has been widely implemented in different research fields.[6–10] The index is now extremely widely used; the 2002 and 2003 papers frequently appear in lists of very highly cited papers.[11,12] The journal *Nature* recently ranked the *BMJ* paper as the twentieth most cited of the twenty-first century.[13]

Although I^2 has been welcomed by some,[14,15] its popularity has been accompanied by some common misconceptions and misinterpretations that hamper its usefulness in practice.[16,17] When interpreted correctly as a measure of inconsistency rather than heterogeneity, the index has a useful place in the meta-analysis toolkit and its idea has been extended to other applications. In this paper, we provide a technical discussion of the I^2 index, providing extensive references to related papers and our reflections from several years experience of working in the research synthesis field. We present different definitions of the I^2 index, clarify its interpretation, address expressions of uncertainty, discuss extensions, and close with our suggestions on how the use and reporting of I^2 might be improved in future meta-analyses.

2. Definitions of I^2

2.1. Standard (Q-based) definition of I^2

Perhaps the most familiar definition of the I^2 index is based on the Q statistic, which is another commonly employed statistic used to examine the variability among the results included in the meta-analysis. In this approach, I^2 is defined as

https://doi.org/10.1017/rsm.2025.10062 Published online by Cambridge University Press

$$I^2 = \frac{Q - (k-1)}{Q} \times 100\%, \tag{1}$$

where k is the number of studies and Q is the χ^2 statistic proposed by Cochran and defined as follows[18]:

$$Q = \sum_i w_i \left(y_i - \overline{y}_w\right)^2. \tag{2}$$

Here, y_i is the result from the i^{th} study; $w_i = 1/\sigma_i^2$ is the weight assigned to the result from the i^{th} study, with σ_i^2 being the within-study variance reflecting the sampling error in y_i and typically assumed to be known in meta-analysis; and $\overline{y}_w = \sum w_i y_i / \sum w_i$ is an inverse-variance-weighted average of the study-specific estimates. Under the simplifying assumption that weights are known and identical for every study, the Q statistic follows a chi-squared distribution with $k-1$ degrees of freedom.[19] Although the Q statistic has been widely used as a hypothesis test to detect heterogeneity among the individual results, it suffers from low statistical power in most meta-analytic applications and therefore can easily lead to misleading conclusions.[2,20]

Because the value of Q can be smaller than its degrees of freedom, the I^2 index defined in (1) can be negative. Therefore, in practice, I^2 is customarily truncated at 0%, such that negative numbers (indicating less variation than would be expected by chance alone) are not allowed.

2.2. Conceptual definition of I^2

A broader definition of I^2 relates to the underlying parameters rather than to the test statistic Q. Although the following definition appears to be parametric on the surface, it is not strictly parametric, as we shall discuss. We define a quantity ι^2 as

$$\iota^2 = \frac{\tau^2}{\tau^2 + \sigma^2} \times 100\%, \tag{3}$$

where τ^2 is the between-study variance parameter and σ^2 is a measure of within-study variance (i.e., the sampling error in estimation of y_i). We use a Greek letter here to distinguish the parametric form, ι^2, from a statistic computed from data, for which we use I^2.

Numerous estimators of between-study variance, τ^2, have been described in the meta-analytic literature.[21] A moment-based estimate, $\widehat{\tau}^2_{DL}$, introduced by DerSimonian and Laird, was popular for many years, although others are now considered preferable.[22,23] The moment-based estimate is given by[24]

$$\widehat{\tau}^2_{DL} = \max\left(\frac{Q - (k-1)}{\sum w_i - \frac{\sum w_i^2}{\sum w_i}}, 0\right).$$

The within-study variance quantity σ^2 is a clearly defined parameter only if all studies have the same within-study variance. Otherwise, σ^2 needs to be interpreted as a "typical" within-study variance across the studies included in the meta-analysis. Different choices are available for estimating such a typical variance, including the arithmetic mean, the median, and the geometric mean of the σ_i^2. HT proposed the quantity:

$$\widehat{\sigma}^2_{HT} = \frac{(k-1)\sum w_i}{(\sum w_i)^2 - \sum w_i^2} \tag{4}$$

with $w_i = 1/\sigma_i^2$.[3] Takkouche, Cadarso-Suárez, and Spiegelman (TCS) used[5]

$$\widehat{\sigma}^2_{TCS} = k \times Var\left(\overline{y}_w\right). \tag{5}$$

https://doi.org/10.1017/rsm.2025.10062 Published online by Cambridge University Press

Since $Var\left(\overline{y}_w\right) = 1/\sum w_i$ and $w_i = 1/\sigma_i^2$, it can be seen that $\widehat{\sigma}^2_{TCS}$ is the harmonic mean of the σ_i^2 values.

Given that there are several options to estimate the between-study variance τ^2 and several options to estimate a typical within-study variance σ^2, numerous values of I^2 can be derived using Equation (3). The Q-based definition of I^2 in Equation (1) corresponds to the choice of the DerSimonian and Laird estimator, $\widehat{\tau}^2_{DL}$, and the use of $\widehat{\sigma}^2_{HT}$ from Equation (4). We refer to this as the "method-of-moments" (MM) definition:

$$I^2_{MM} = \frac{\widehat{\tau}^2_{DL}}{\widehat{\tau}^2_{DL} + \widehat{\sigma}^2_{HT}} = \frac{\frac{Q-(k-1)}{\sum w_i - \frac{\sum w_i^2}{\sum w_i}}}{\left(\frac{Q-(k-1)}{\sum w_i - \frac{\sum w_i^2}{\sum w_i}}\right) + \left(\frac{(k-1)\sum w_i}{(\sum w_i)^2 - \sum w_i^2}\right)} = \frac{Q-(k-1)}{Q}. \tag{6}$$

2.3. Heuristic definition of I^2

Noting the broad interpretation of I^2 as the proportion of total variation in observed results that arises from between-study heterogeneity,[3,25] a further option to define I^2 is as follows:

$$I^2 = \frac{\widehat{\tau}^2}{Var\left(y_i\right)} \times 100\%. \tag{7}$$

The denominator in (7) is just the sample variance of the study estimates. Under the usual additive heterogeneity model, rather than a multiplicative heterogeneity model,[26] this can be regarded as conceptually equivalent to the sum of the between-study (τ^2) and within-study (σ^2) variance components in the denominator of Equation (3).[25]

2.4. General definition of I^2

The conceptual definition presented in Equation (3) is analogous to using weighted sums of squares as follows:

$$\iota^2 = \frac{WSS_b}{WSS_b + WSS_w} \times 100\%. \tag{8}$$

In this expression, $WSS_b = \sum w_i(\theta_i - \mu)^2$ is the between-study weighted sum of squares and quantifies the dispersion of the effect parameters from each study (θ_i) from the overall mean μ. The term $WSS_w = \sum w_i(y_i - \theta_i)^2$ is the within-study weighted sum of squares, which accounts for the dispersion of the estimates (y_i) from the effect parameters θ_i for each study.

We can verify the conceptual equivalence of (8) with (3) and (7) by noting that

$$\begin{aligned} WSS_{\text{total}} &= \sum w_i(y_i - \mu)^2 \\ &= \sum w_i(y_i - \theta_i + \theta_i - \mu)^2 \\ &= \sum w_i(y_i - \theta_i)^2 + \sum w_i(\theta_i - \mu)^2 + 2\sum w_i\left(y_i - \theta_i\right)\left(\theta_i - \mu\right), \end{aligned} \tag{9}$$

where $\sum w_i(\theta_i - \mu)^2$ corresponds to WSS_b and $\sum w_i(y_i - \theta_i)^2$ corresponds to WSS_w. The final term, $2\sum w_i\left(y_i - \theta_i\right)\left(\theta_i - \mu\right)$, represents residual covariation between the within- and between-study variance components. When within-study variances are uncorrelated with effect sizes, this covariation will be zero and the WSS_{total} is equivalent to the sum of WSS_b and WSS_w.

The general definition of I^2 presented in Equation (8) provides a framework to specify the "true" value of the index (ι^2) in simulation studies. We used this definition for the study presented in the next section.

https://doi.org/10.1017/rsm.2025.10062 Published online by Cambridge University Press

3. Illustration of the properties of different I^2 estimators

3.1. Application to two real data sets

Simulation studies indicate that alternatives to $\widehat{\tau}^2_{DL}$ have better properties,[22,23] so the question arises as to whether using an alternative estimator for τ^2 in (3) would be preferable for computing I^2. In addition, since different options are available for estimating σ^2, it is not immediately clear which should be used. In a later section, we examine some of these issues in a simulation study. Here, we illustrate the impact that different choices can have through application to two examples, using the *metafor* R package for the calculations.[27]

The first example is a set of 13 clinical trials of Bacillus Calmette–Guerin (BCG) vaccine to prevent tuberculosis,[28] using data presented by Berkey and colleagues.[29] The effect measure for this example is the log risk ratio comparing the risk of tuberculosis in the vaccinated versus unvaccinated groups. In Table 1, we list different values of I^2 for different estimators of τ^2 and σ^2 as well as using the simple sample variance of the y_i as the denominator (Equation (7)). A key point we seek to illustrate here is that I^2 values range markedly, from a minimum of 47.4% to a maximum of 94.2%. The usual value, based on the DerSimonian–Laird estimator of τ^2 ($\widehat{\tau}^2_{DL}$) and the HT "typical variance" formula ($\widehat{\sigma}^2_{HT}$), is 92.1%.

The second example is a set of results from 19 studies examining how teachers' expectations influence their pupils' intelligence quotient (IQ) scores, a phenomenon sometimes labeled as the "Pygmalion effect" using data provided by Raudenbush (Table 2).[30] The effect measure for this example is the standardized mean difference. Data were obtained from the metadata R package.[31] Again, I^2 values range notably, from a minimum of 9.8% to a maximum of 76.1%, with the usual value using $\widehat{\tau}^2_{DL}$ and $\widehat{\sigma}^2_{HT}$ being I^2_{MM} = 49.8%. The question then arises as to whether any of these variants is "more correct" than others, and we turn to this question in the following section by conducting a small simulation study.

3.2. A simulation study

3.2.1. Simulation design

We conducted a small simulation study to examine the different approximations to the "typical" within-study variance in Equations (3) and (8) in the context of a very large meta-analysis (1,000 studies). For each meta-analysis, we drew k = 1,000 effect parameters θ_i from a normal distribution with μ = 0 and

Table 1. *Different values of* $\boldsymbol{I^2}$ *obtained using the different definitions, with different choices of estimators for* $\boldsymbol{\sigma^2}$ *and* $\boldsymbol{\tau^2}$*, applied to studies of BCG vaccine.*

		Choice for $\widehat{\sigma}^2$					
		Arithmetic mean	Median	Geometric mean	HT	TCS	Proportion of total sample variance
Choice for $\widehat{\tau}^2$	DL	66.9%	81.2%	82.8%	**92.1%**	93.5%	64.1%
	Hedges	68.3%	82.1%	83.6%	92.6%	93.9%	68.3%
	HS	59.9%	76.2%	78.0%	89.6%	91.5%	47.4%
	SJ	69.3%	82.9%	84.3%	92.9%	94.2%	71.8%
	ML	64.7%	79.7%	81.3%	91.4%	92.9%	58.2%
	REML	67.2%	81.4%	83.0%	92.2%	93.6%	65.1%
	EB/PM	67.5%	81.7%	83.2%	92.3%	93.7%	66.1%

Note: The bold entry is the "standard" choice ($\boldsymbol{I^2_{MM}}$) as introduced by Higgins and Thompson.
HT, Higgins and Thompson; TCS, Takkouche, Cadarso-Suárez, and Spiegelman; DL, DerSimonian–Laird; HS, Hunter–Schmidt; SJ, Sidik–Jonkman; MLE, maximum likelihood; REML, restricted maximum likelihood; EB, empirical Bayes; PM, Paule–Mandel.

https://doi.org/10.1017/rsm.2025.10062 Published online by Cambridge University Press

Table 2. *Different values of* $\boldsymbol{I^2}$ *obtained using different definitions, with different choices of estimators for* $\boldsymbol{\sigma^2}$ *and* $\boldsymbol{\tau^2}$*, applied to studies on the "Pygmalion effect".*

		Choice for $\widehat{\sigma}^2$					
		Arithmetic mean	Median	Geometric mean	HT	TCS	Proportion of total sample variance
Choice for $\widehat{\tau}^2$	DL	34.9%	48.1%	42.7%	**49.8%**	50.6%	20.1%
	Hedges	62.4%	74.2%	69.8%	75.4%	76.1%	62.4%
	HS	31.6%	44.5%	39.2%	46.1%	47.0%	17.4%
	SJ	61.4%	73.4%	68.9%	74.7%	75.3%	59.9%
	ML	20.6%	31.0%	26.6%	32.4%	33.2%	9.8%
	REML	28.0%	40.3%	35.1%	41.9%	42.7%	14.6%
	EB/PM	51.7%	65.0%	59.8%	66.4%	67.2%	40.2%

Note: The bold entry is the "standard" choice ($\boldsymbol{I^2_{MM}}$) as introduced by Higgins and Thompson.
HT, Higgins and Thompson; TCS, Takkouche, Cadarso-Suárez, and Spiegelman; DL, DerSimonian–Laird; HS, Hunter–Schmidt; SJ, Sidik–Jonkman; MLE, maximum likelihood; REML, restricted maximum likelihood; EB, empirical Bayes; PM, Paule–Mandel.

$\tau^2 = 0.001$ and then generated individual scores on a continuous outcome for two-arm studies from normal distributions with means θ_i and 0 for the experimental and control arms, respectively, and with a between-individual variance of 1 for both arms. The effect measure for each study was Hedges' g, that is, the standardized mean difference as presented by Hedges and Olkin.[32] We assigned arm sizes of 300, 350, 400, 450, and 1,000 individuals, with each value being assigned to 200 studies in the meta-analysis, with equal arm sizes for the two arms within each study. Total study sizes were therefore 600, 700, 800, 900, or 2,000 individuals, and the corresponding within-study variances (σ_i^2) were 0.0067, 0.0057, 0.0050, 0.0044, or 0.0020.

We simulated 10,000 meta-analyses using R version 4.3.3.[33] For each simulated meta-analysis, we computed the within-study effect estimate (y_i) and resulting "true" I^2 using a general weighted sums of squares approach (Equation (8)) based on known values of μ and σ_i^2 and simulated values of θ_i. We compared these true values with the values yielded by replacing τ^2 in Equation (3) with the variance of the effect parameters θ_i (to avoid selecting a specific estimator for τ^2) and using different estimators of σ^2 in Equation (3): arithmetic mean, median, HT (Equation (5)), and Takkouche and colleagues (Equation (6)). We also included the method using the total sampling variance of the observed effect estimates (Equation (7)).

3.2.2. Simulation results

Table 3 shows the results of the simulation. Weighted sums of squares WSS_b and WSS_w (see Equation (8)) yielded values of 0.001 and 0.004, respectively, the former corresponding to the preset known value of the between-study variance component and the latter defining the true value of the within-study variance component (this coincides with the harmonic mean of the vector of five preset σ_i^2 values), leading to a true parameter value of $\iota^2 = 20.0\%$. Among the methods we compared, the formulae proposed by HT and by Takkouche and colleagues both yielded the correct average values of 0.004 for within-study variation. Conversely, using the median (0.005) or the arithmetic mean (0.005) led to the overestimation of within-study variation, resulting in an underestimation of ι^2 (by 3.3% and 2.7%, respectively). Lastly, the method using the total sampling variance of the observed effect estimates also underestimated the true value of I^2 with an average value of 17.4%. Therefore, the different methods proposed to quantify a representative value of the within-study variation across studies, σ^2, showed some discrepancies that affected the estimation of I^2 using Equation (3), with only the Takkouche and the HT methods performing close to the general formula presented in Equation (8). The lesson from our simulations is that only these two approaches should be considered (from those

https://doi.org/10.1017/rsm.2025.10062 Published online by Cambridge University Press

Table 3. *Results from the simulation study.*

Approach	Between studies	Within studies	I^2
WSS	0.001	0.004	20.0%
Variances: HT	0.001	0.004	20.0%
Variances: TCS		0.004	20.0%
Variances: Med		0.005	16.7%
Variances: AM		0.005	17.3%
Total variance	0.006		17.4%

WSS, weighted sums of squares; HT, Higgins and Thompson; TCS, Takkouche, Cadarso-Suárez, and Spiegelman (harmonic mean); Med, median; AM, arithmetic mean.

compared) for summarizing within-study variances across studies. Comparisons of these two methods in less favorable simulated scenarios can be found elsewhere.[34,35] Our simulations involved large numbers of studies so as to ensure that the between-study variance was well estimated. Further simulation studies might elucidate preferred estimators of between-study variance, although we would expect findings to align with existing recommendations for the choice of between-study variance.[22,23]

4. Expressing uncertainty in I^2

Values of the I^2 index can be highly volatile when based on a small number of studies.[36] In many fields, very few studies tend to be included in meta-analyses (with a median of 3 having been reported in the health area[37]). HT originally derived formulae to calculate a confidence interval around I^2 based on H^2, another metric for quantifying inconsistency in meta-analysis sometimes known as Birge's ratio.[3] An alternative, preferable, confidence interval for I^2 is based on a noncentral chi-squared distribution,[38] used by Orsini.[39] Hedges and Pigott note that this is the distribution of Q under a fixed-effects model, while under a random-effects model it has a gamma distribution.[38] These results are used as the default in the *metan* macro for Stata.[40]

Confidence intervals for I^2 may also be derived from confidence intervals for τ^2. Methods for these include the Q-profile method,[41,42] a generalized Q-statistic method,[43,44] and a profile-likelihood method,[45,46] all of which are implemented in the metafor package for R.[27,33] Additional approaches have been developed that do not require a known sampling distribution of the Q-statistic[47,48] an assumption which has been deemed unrealistic in most applied situations.[19] To the best of our knowledge, no simulation studies comparing all existing methods have been conducted to date, although an empirical comparison of the performance of some of them was recently conducted by Wang and colleagues.[49]

5. Interpretation of I^2

The I^2 index is often interpreted as the amount of heterogeneity among studies, that is, as a measure of the variability across studies in true effect sizes underlying the effect estimates. Despite being widespread in the literature, this interpretation is incorrect. Numerous papers have explained this misinterpretation, and we refer the reader to several clear expositions.[16,17,50–52] The I^2 index measures the amount of heterogeneity *in relation to the sampling error associated with the effect estimates feeding into the meta-analysis*. As such, it is similar in interpretation to the intraclass correlation coefficient in clustered data. In fact, the letter I was originally chosen to represent "intraclass," although we regard it as more constructive to consider the I to stand for "inconsistency."

I^2 may be interpreted as the proportion of variability in point estimates that is due to heterogeneity rather than sampling error. Informally, it can be considered to reflect the extent to which confidence intervals displayed in a forest plot do not overlap with each other. This is the rationale for the

https://doi.org/10.1017/rsm.2025.10062 Published online by Cambridge University Press

characterization of I^2 as a measure of *inconsistency*. If the results of the different studies are consistent with each other, their confidence intervals will have a high degree of overlap. If the results are inconsistent, there will be poor overlap in the confidence intervals. Note that the degree (or not) of overlap in confidence intervals is unrelated to the numbers on the effect measure axis. I^2 is invariant to a linear transformation of point estimates and confidence interval limits (in the metric used for the analysis, which is typically the logarithmic scale if the effect is measured on a ratio scale). Thus, I^2 cannot be a measure of absolute heterogeneity, which is a quantity that is specific to the actual numbers on the axis.

A fundamental property of I^2 is that it will increase as the study effect estimates get more precise. Specifically, as studies get larger, confidence intervals around their effect estimates get narrower, and the overlap in confidence intervals across studies reduces, leading to larger values of I^2. This is apparent from Equation (3): For a given set of true effect sizes θ_i (and hence fixed Var $(\theta_i) = \tau^2$), the value of ι^2 increases as the typical within-study variance, σ^2, decreases.[17]

A much lesser problem in the interpretation of the I^2 index is that it has a minor dependence on the number of studies (k) in the meta-analysis, as simulation studies have shown. Specifically, I^2 has been reported to yield biased estimates when the number of studies, k, is very small,[25] similar to the pseudo-R^2 statistic, another ratio measure in meta-analysis that requires estimation of τ^2.[53] Furthermore, the I^2 index is closely related to the Q statistic, and hence, analogous to the low statistical power of the test based on the Q statistic, it has substantial uncertainty unless a large number of results is combined.[35,54] Performance of both the Q-statistic and the I^2 index has also been found to be affected by publication bias simulated by Augusteijn and colleagues, who additionally developed a web-based application to explore the potential impact of publication bias on the assessment of heterogeneity for a given meta-analytic data set.[55]

An important additional point worth noting here is that there are technical problems with I^2 arising from the fact that the within-study variances are not known but estimated with error. These problems, which will be more important when study estimates have low precision, are well articulated by Hoaglin.[19,56]

5.1. What is a large value of I^2?

It is tempting to seek ranges or thresholds for the interpretation of I^2: What is a large amount of inconsistency? The original paper proposing the index proposed that "mild heterogeneity might account for less than 30 per cent of the variability in point estimates, and notable heterogeneity substantially more than 50 per cent." The use of the word "heterogeneity" in these proposals, while correct, may have led readers to believe the descriptors "mild" and "notable" related to the amount of heterogeneity rather than the amount of inconsistency. This is unfortunate. The arguments were based informally on the connection between heterogeneity and inconsistency for situations in which the chi-squared test for heterogeneity achieves statistical significance (at a 5% level) for around 10–30 studies, with meta-analyses of randomized trials very much in mind.[3] The benchmarks did not directly consider the absolute magnitude of heterogeneity, as measured by τ^2.

The most widely cited paper about I^2 proposed "naive categorisation of values for I^2 would not be appropriate for all circumstances, although we would tentatively assign adjectives of low, moderate, and high to I^2 values of 25%, 50%, and 75%."[4] Again, these proposals were intended to apply only to randomized trials. Despite the caveats, these benchmarks have become widely (mis)used and applied to situations far beyond meta-analyses of trials.

In an attempt to curb the overzealous use of the thresholds, the *Cochrane Handbook for Systematic Reviews of Interventions* in 2008 proposed overlapping ranges, again intended for use only with randomized trials: 0% to 40% might not be important; 30% to 60% may represent moderate heterogeneity; 50% to 90% may represent substantial heterogeneity; 75% to 100% may represent considerable heterogeneity.[57] These suggestions were accompanied by clear guidance that "the importance of the observed value of I^2 depends on (i) magnitude and direction of effects and (ii) strength of evidence for

https://doi.org/10.1017/rsm.2025.10062 Published online by Cambridge University Press

heterogeneity (e.g., *P*-value from the chi-squared test, or a confidence interval for I^2)," advice which has again largely been ignored in practice.

Although these benchmarks can be helpful as a very rough guide, they should not be used to categorize meta-analysis data sets or to decide on whether, or how, to combine the results. We particularly discourage the use of thresholds for I^2 to decide between adopting fixed-effect(s) or random-effects meta-analysis models, because such decisions should be made on the basis of the question being asked rather than characteristics of the data.[58] Empirical investigations of typical values of I^2 can provide some useful pointers as to how a particular meta-analysis data set compares with others in related areas. Rhodes and colleagues reanalyzed 9,895 meta-analysis data sets from Cochrane reviews and fitted distributions to them.[59] For meta-analyses of trials with binary outcomes, the median value of I^2 was 22% for log odds ratios, with interquartile range (IQR) 12% to 39%. For continuous outcomes meta-analysis, the median was 40% for standardized mean differences, with IQR 15% to 73%. Such a difference between binary and continuous outcomes has been noted by others.[60] Differences have also been reported depending on the choice of effect measure within categorical[59,61] and continuous[54,59] outcomes, with, for example, inconsistency being higher for risk differences than for odds ratios or risk ratios.

Application of I^2 to results other than treatment effects from randomized trials can yield extreme results. An overview of 134 meta-analyses of prevalence studies found a median I^2 value of 96.9% (IQR: 90.5% to 98.7%).[62] An overview of 138 reliability generalization meta-analyses found a median I^2 value of 93.2% (IQR: 88.8% to 96.4%) when integrating untransformed alpha coefficients, with similar results for the transformations typically used in this field.[63] These very high values reflect the higher precision with which these quantities are estimated in relation to the underlying heterogeneity. For example, studies of prevalence can be very large, being based on large surveys or databases. In this situation, even trivially small differences in prevalence can lead to very large values of I^2. The utility of the I^2 index is therefore questionable in such situations, although as a statistic it cannot be said to be misleading or wrong.

6. Extensions and repurposing of I^2

The popularity of the I^2 index has led to several extensions for application to specific types of meta-analysis. One of these is to meta-analyze models with covariates (known as moderator analysis or meta-regression). Here, part of the between-study variation is hoped to be explained by covariates, with some residual variation that cannot be accounted for by the model covariates. In this context, I^2 is interpreted as the proportion of the residual between-study variation (rather than the total variation) due to heterogeneity as opposed to random sampling error.[64]

Other extensions for more complex models include variants for multilevel meta-analytic models[65] and multivariate meta-analysis,[66,67] with the latter being applicable also in the context of network meta-analysis.[68,69] Multivariate versions were initially based on normal–normal random-effects models and have been extended to binomial–normal random-effects models for use, for example, in meta-analysis of diagnostic test accuracy studies.[70] Analogs to I^2 have additionally been proposed for individual participant data meta-analysis and cluster-randomized trials.[71]

Vo and colleagues extend the use of the I^2 index to examine the impact of different sources of heterogeneity.[72] They partition the heterogeneity variance, τ^2, into heterogeneity due to differences in participant characteristics ("case mix") and heterogeneity due to other sources (largely methodological differences across the studies). They then propose versions of the I^2 index for each of these components separately, expressing each as a proportion of the total variation in study estimates. Note that, in practice, the separation of these two sources of heterogeneity can only be achieved with individual participant data.[72]

The I^2 index has been repurposed for application to other problems. Montori and colleagues apply the I^2 index to multiple waves of a survey, where at each wave an attempt is made to reach nonresponders to earlier waves.[8] The I^2 index calculated from answers to a survey question across

https://doi.org/10.1017/rsm.2025.10062 Published online by Cambridge University Press

multiple waves provides a measure of the inconsistency across waves due to response bias. Specifically, it describes the proportion of total variation in responses to the question, in successive waves, that is due to differences in the responders across waves rather than random variation within each wave. Finally, Bowden and colleagues adopted the I^2 index for use in the field of Mendelian randomization to quantify the strength of violation of an assumption made in two-sample Mendelian randomization.[73]

7. Alternative indices of heterogeneity

Alongside the I^2 index, popular statistics for assessing heterogeneity in a meta-analysis are the Q-statistic and the between-study variance parameter. The latter is easier to interpret when reported as the standard deviation, τ. This standard deviation may also be used to calculate prediction intervals around the overall effect estimate.[74] Prediction intervals provide a range of plausible values for the true effect in a further primary study similar to those already in the meta-analysis. They directly portray the amount of heterogeneity observed in the meta-analysis on the metric used to present the meta-analytic results. The use and reporting of prediction intervals is now widely advocated in the meta-analytic literature.[74–80]

Assessment of heterogeneity is a key aspect of any meta-analysis, so it is not surprising that several alternatives to these indices have been proposed.[81] Lin and colleagues[82] developed two indices intended for meta-analyses including outlying studies and reported a robust performance under such scenarios.[83] Crippa and colleagues proposed another scale-invariant measure, which, unlike the I^2 index, requires neither quantification of a typical value of the within-study variance nor an assumption of homogeneity of these quantities.[84] The proposed index quantifies the between-study heterogeneity relative to the variance of the summary random-effects estimate. A disadvantage of this measure is the potential for bias, especially with a small number of studies.[81] The coefficient of variation, calculated as the ratio between τ and either the fixed-effects[5] or random-effects[85] overall estimate, has been proposed as a complementary measure of heterogeneity, although it can be expected to yield very large values when the effect estimate is small, which hampers interpretation.[86]

Last, we note that HT originally proposed the I^2 index along with two other quantities.[3] The first of these was the H index or Birge's ratio, closely related to the standard (Q-based) definition of I^2, and with a modified version subsequently proposed.[35] The second was the R^2 index, a ratio of the variances of the random-effects and the fixed-effects overall estimates in the meta-analysis, subsequently rebranded as the diamond ratio.[87]

8. Discussion

Papers proposing the I^2 index for meta-analysis are among the most highly cited in the medical research literature, with the original papers in *Statistics in Medicine* and *BMJ* having been cited in excess of 35,000 and 60,000 times, respectively. In this paper, we have reviewed the properties and interpretations of the I^2 index. We overviewed different ways of defining the statistic and illustrated how different definitions can lead to different values.

When a fixed-effects model is assumed, there is no estimation of the between-study variance, τ^2, and hence, we regard the standard (Q-based) definition to be a natural choice to compute I^2. Furthermore, if both fixed- and random-effects meta-analyses are planned, then the Q-based definition has the advantage that it will yield the same result for both statistical models. In the context of a random-effects meta-analysis, other definitions for I^2 are serious contenders. A particular problem is that the closest version to a parametric interpretation of the I^2 index (our conceptual definition) involves the notion of a "typical" within-study variance term, which is a poorly defined quantity. As a result, we propose a general definition of I^2 based on weighted sums of squares and use this as a basis for a small simulation study. We suggest that this new definition may have a useful role in a range of future simulation studies investigating methods for estimating heterogeneity parameters or mean meta-analytic effects.

https://doi.org/10.1017/rsm.2025.10062 Published online by Cambridge University Press

We believe the index plays a very valuable role in the meta-analysis toolkit, although we are concerned at the widespread misunderstanding and misinterpretation of it in practice. Common mistakes are to interpret the I^2 index as a measure of the absolute amount of heterogeneity across studies and to apply inappropriate thresholds for interpreting values of the index. It is important to remember that the I^2 index is a relative measure of inconsistency in results rather than an absolute measure of heterogeneity in effects. It intentionally depends on the precisions (or within-study variances) of the study estimates contributing to the meta-analysis. One consequence of this is that it cannot be used to measure the amount of heterogeneity; its value is wholly uninformative about variation in effect sizes on the scale used to measure the effect. We suspect that many users of the I^2 index have little appreciation of the actual degree to which the effects in their meta-analysis vary from study to study, which may lead them to inappropriate interpretations of their results. A second consequence is that, in some situations (mainly when studies are large), the value of I^2 can be expected to be large. This does not mean that use of the I^2 index is wrong, merely that it may not be the most helpful measure. To make the most of the imperfections in I^2, we offer a series of recommendations for its appropriate use (see Box).

Box. Recommendations for use of the I^2 index in meta-analyses.

1. Ensure I^2 is appropriately interpreted as a *relative* measure (of inconsistency) and not an *absolute* measure (of heterogeneity).
2. Recognize that, in many situations, large studies will naturally lead to large values of I^2. This is not a problem; it is inherent in the purpose of I^2 as a measure of inconsistency of results rather than heterogeneity of effects.
3. Accompany I^2 with some measure to convey the strength of evidence for whether heterogeneity is truly present. This could be, for example, a confidence interval for I^2 or an exact P value from the Q test for heterogeneity.
4. Avoid strong reliance on thresholds for interpreting I^2. Any thresholds that are used should be appropriate to the type of study, noting that commonly cited thresholds were derived specifically for clinical trials.
5. Be aware that I^2 may naturally be large for meta-analyses of single-group studies, particularly when true values can be variable in relation to the sampling errors involved in estimating them. This is particularly true, for example, in meta-analyses of prevalence estimates or sensitivity/specificity estimates in test accuracy studies.
6. To describe the actual amount of heterogeneity across studies, present the between-study standard deviation in effects (τ) rather than I^2. Note that this standard deviation is usefully portrayed using a prediction interval in the context of a random-effects analysis.

Author contributions. J.P.T.H. conceptualized the study; investigated the data; designed the methodology; administered the project; wrote the original draft; and wrote, reviewed, and edited the manuscript. J.A.L.-L. conceptualized the study; investigated the data; designed the methodology; wrote the original draft; and wrote, reviewed, and edited the manuscript.

Competing interest statement. J.P.T.H. codeveloped the I-squared index. The remaining author declares that no competing interests exists.

Data availability statement. The R code used for the examples and simulation can be found at https://osf.io/yp7u4/.

Funding statement. J.P.T.H. is a National Institute for Health and Care Research (NIHR) Senior Investigator (NIHR203807). He is supported by the NIHR Bristol Evidence Synthesis Group (NIHR153861) and the NIHR Applied Research Collaboration West (ARC West) at University Hospitals Bristol and Weston NHS Foundation Trust (NIHR200181). J.A.L.-L. was supported by the Regional Program for the Promotion of Scientific and Technical Research of Excellence (Action Plan 2022) of the Seneca Foundation—Science and Technology Agency of the Region of Murcia, Spain (Grant No. 22064/PI/22). The views expressed in this article are those of the authors and do not necessarily represent those of the funders.

https://doi.org/10.1017/rsm.2025.10062 Published online by Cambridge University Press

References

[1] Higgins JPT. Commentary: heterogeneity in meta-analysis should be expected and appropriately quantified. *Int J Epidemiol*. 2008;37(5): 1158–1160.
[2] Borenstein M, Hedges LV, Higgins JPT, Rothstein H. *Introduction to Meta-Analysis*. John Wiley & Sons; 2021.
[3] Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med*. 2002;21(11): 1539–1558.
[4] Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analysis. *BMJ*. 2003;327: 557–560.
[5] Takkouche B, CadarsoSurez C, Spiegelman D. Evaluation of old and new tests of heterogeneity in epidemiologic meta-analysis. *Am J Epidemiol*. 1999;150(2): 206–215.
[6] Hansen RA, Zeng P, Ryan P, et al. Exploration of heterogeneity in distributed research network drug safety analyses. *Res Synth Methods*. 2014;5(4): 352–370.
[7] Ioannidis JPA, Patsopoulos NA, Evangelou E. Heterogeneity in meta-analyses of genome-wide association investigations. *PLoS One*. 2007;2(9): e841.
[8] Montori VM, Leung TW, Walter SD, Guyatt GH. Procedures that assess inconsistency in meta-analyses can assess the likelihood of response bias in multiwave surveys. *J Clin Epidemiol*. 2005;58(8): 856–858.
[9] Senior AM, Grueber CE, Kamiya T, et al. Heterogeneity in ecological and evolutionary meta-analyses: its magnitude and implications. *Ecology*. 2016;97(12): 3293–3299.
[10] Villanueva EV, Zavarsek S. Evaluating heterogeneity in cumulative meta-analyses. *BMC Med Res Methodol*. 2004;4: 18.
[11] Uthman OA, Okwundu CI, Wiysonge CS, Young T, Clarke A. Citation classics in systematic reviews and meta-analyses: who wrote the top 100 most cited articles? *PLoS One*. 2013;8(10): e78517.
[12] Nietert PJ, Wahlquist AE, Herbert TL. Characteristics of recent biostatistical methods adopted by researchers publishing in general/internal medicine journals. *Stat Med*. 2013;32(1): 1–10.
[13] Pearson H, Ledford H, Hutson M, Van Noorden R. Exclusive: the most-cited papers of the twenty-first century. *Nature*. 2025;640(8059): 588–592.
[14] Ioannidis JP. Interpretation of tests of heterogeneity and bias in meta-analysis. *J Eval Clin Pract*. 2008;14(5): 951–957.
[15] Pathak M, Dwivedi SN, Deo SVS, Sreenivas V, Thakur B. Which is the preferred measure of heterogeneity in meta-analysis and why? A revisit. *Biostat Biomet Open Access J*. 2017;1(1): 14–20.
[16] Borenstein M, Higgins JP, T Hedges LV, Rothstein HR. Basics of meta-analysis: I^2 is not an absolute measure of heterogeneity. *Res Synth Methods*. 2017;8(1): 5–18.
[17] Rücker G, Schwarzer G, Carpenter JR, Schumacher M. Undue reliance on I^2 in assessing heterogeneity may mislead. *BMC Med Res Methodol*. 2008;8: 79.
[18] Cochran WG. The combination of estimates from different experiments. *Biometrics*. 1954;10: 101–129.
[19] Hoaglin DC. Misunderstandings about Q and 'Cochran's Q test' in meta-analysis. *Stat Med*. 2016;35(4): 485–495.
[20] Borenstein M. *Heterogeneity in MetaAanalysis*. In: Cooper H, Hedges LV, Valentine J, eds. *Handbook of Research Synthesis and Meta-Analysis*, 3rd Edition. Russell Sage Foundation; 2019: 453–468.
[21] Veroniki AA, Jackson D, Viechtbauer W, et al. Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Res Synth Meth*. 2016;7(1): 55–79.
[22] Langan D, Higgins JPT, Simmonds M. Comparative performance of heterogeneity variance estimators in meta-analysis: a review of simulation studies. *Res Synth Methods*. 2017;8(2): 181–198.
[23] Langan D, Higgins JPT, Jackson D, et al. A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. *Res Synth Methods*. 2019;10(1): 83–98.
[24] DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986;7: 177–188.
[25] von Hippel PT. The heterogeneity statistic I^2 can be biased in small meta-analyses. *BMC Med Res Methodol*. 2015;15: 35.
[26] Mawdsley D, Higgins JP, Sutton AJ, Abrams KR. Accounting for heterogeneity in meta-analysis using a multiplicative model-an empirical study. *Res Synth Methods*. 2017;8(1): 43–52.
[27] Viechtbauer W. Conducting meta-analyses in R with the metafor package. *J Stat Softw*. 2010;36(3): 1–48.
[28] Colditz GA, Brewer TF, Berkey CS, et al. Efficacy of BCG vaccine in the prevention of tuberculosis: meta- analysis of the published literature. *JAMA*. 1994;271(9): 698–702.
[29] Berkey CS, Hoaglin DC, Mosteller F, Colditz GA. A random-effects regression model for meta-analysis. *Stat Med*. 1995;14: 395–411.
[30] Raudenbush SW. Magnitude of teacher expectancy effects on pupil Iq as a function of the credibility of expectancy induction - a synthesis of findings from 18 experiments. *J Educ Psychol*. 1984;76(1): 85–97.
[31] White T, Nobles D, Senior A, Hamilton W, Viechtbauer W. metadat: meta-analysis datasets. R package version 1.2-0. 2022. Available at https://CRANR-projectorg/package=metadat.
[32] Hedges LV, Olkin I. *Statistical Methods for Meta-Analysis*. Academic Press; 1985.
[33] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; 2024. Available at https://www.R-project.org/.
[34] Böhning D, Lerdsuwansri R, Holling H. Some general points on the I^2-measure of heterogeneity in meta-analysis. *Metrika*. 2017;80: 685–695.
[35] Mittlbock M, Heinzl H. A simulation study comparing properties of heterogeneity measures in meta-analyses. *Stat Med*. 2006;25(24): 4321–4333.

https://doi.org/10.1017/rsm.2025.10062 Published online by Cambridge University Press

[36] Ioannidis JPA, Patsopoulos NA, Evangelou E. Uncertainty in heterogeneity estimates in meta-analyses. *BMJ*. 2007;335(7626): 914–916.
[37] Davey J, Turner RM, Clarke MJ, Higgins JPT. Characteristics of meta-analyses and their component studies in the Cochrane database of systematic reviews: a cross-sectional, descriptive analysis. *BMC Med Res Methodol*. 2011;11: 160.
[38] Hedges LV, Pigott TD. The power of statistical tests in meta-analysis. *PsycholMethods*. 2001;6(3): 203–217.
[39] Orsini N, Bottai M, Higgins J. HETEROGI: Stata module to quantify heterogeneity in meta-analysis. 2005. Available at http://econpapersrepecorg/software/bocbocode/s449201htm.
[40] Fisher D, Harris R, Bradburn M, et al. METAN: Stata module for fixed and random effects meta-analysis. 2022. Available at https://econpapersrepecorg/software/bocbocode/s456798htm.
[41] Knapp G, Biggerstaff BJ, Hartung J. Assessing the amount of heterogeneity in random-effects meta-analysis. *Biom J*. 2006;48(2): 271–285.
[42] Viechtbauer W. Confidence intervals for the amount of heterogeneity in meta-analysis. *Stat Med*. 2007;26(1): 37–52.
[43] Hartung J, Knapp G. On confidence intervals for the among-group variance in the one-way random effects model with unequal error variances. *J Stat Plan Infer*. 2005;127(1–2): 157–177.
[44] Jackson D. Confidence intervals for the between-study variance in random effects meta-analysis using generalised Cochran heterogeneity statistics. *Res Synth Methods*. 2013;4(3): 220–229.
[45] Hardy RJ, Thompson SG. A likelihood approach to meta-analysis with random effects. *Stat Med*. 1996;15: 619–629.
[46] Viechtbauer W, Lopez-Lopez JA. Location-scale models for meta-analysis. *Res Synth Methods*. 2022;13(6): 697–715.
[47] Kulinskaya E, Dollinger MB. An accurate test for homogeneity of odds ratios based on Cochran's Q-statistic. *BMC Med Res Methodol*. 2015;15: 49.
[48] Kulinskaya E, Dollinger MB, Bjorkestol K. Testing for homogeneity in meta-analysis I. The one-parameter case: standardized mean difference. *Biometrics*. 2011;67(1): 203–212.
[49] Wang Y, DelRocco N, Lin L. Comparisons of various estimates of the I^2 statistic for quantifying between-study heterogeneity in meta-analysis. *Stat Methods Med Res*. 2024;33(5): 745–764.
[50] Borenstein M. In a meta-analysis, the I-squared statistic does not tell us how much the effect size varies. *J Clin Epidemiol*. 2022;152: 281–284.
[51] Borenstein M. How to understand and report heterogeneity in a meta-analysis: the difference between I-squared and prediction intervals. *Integr Med Res*. 2023;12(4): 101014.
[52] Borenstein M. Avoiding common mistakes in meta-analysis: understanding the distinct roles of Q, I-squared, tau-squared, and the prediction interval in reporting heterogeneity. *Res Synth Methods*. 2024;15(2): 354–368.
[53] Lopez-Lopez JA, Marin-Martinez F, Sanchez-Meca J, Van den Noortgate W, Viechtbauer W. Estimation of the predictive power of the model in mixed-effects meta-regression: a simulation study. *Brit J Math Stat Psy*. 2014;67(1): 30–48.
[54] Huedo-Medina TB, Sánchez-Meca J, Marín-Martínez F, Botella J. Assessing heterogeneity in meta-analysis: Q statistic or I^2 index? *Psychol Methods*. 2006;11(2): 193–206.
[55] Augusteijn HEM, van Aert RCM, van Assen M. The effect of publication bias on the Q test and assessment of heterogeneity. *Psychol Methods*. 2019;24(1): 116–134.
[56] Hoaglin DC. Practical challenges of I(2) as a measure of heterogeneity. *Res Synth Methods*. 2017;8(3): 254.
[57] Higgins JPT, Thomas J, Chandler J, et al. *Cochrane Handbook for Systematic Reviews of Interventions*. John Wiley & Sons; 2019.
[58] Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Res Synth Methods*. 2010;1(2): 97–111.
[59] Rhodes KM, Turner RM, Higgins JPT. Empirical evidence about inconsistency among studies in a pair-wise meta-analysis. *Res Synth Methods*. 2016;7(6): 346–370.
[60] Alba AC, Alexander PE, Chang J, Maclsaac J, DeFry S, Guyatt GH. High statistical heterogeneity is more frequent in meta-analysis of continuous than binary outcomes. *J Clin Epidemiol*. 2016;70: 129–135.
[61] Papageorgiou SN, Tsiranidou E, Antonoglou GN, Deschner J, Jager A. Choice of effect measure for meta-analyses of dichotomous outcomes influenced the identified heterogeneity and direction of small-study effects. *J Clin Epidemiol*. 2015;68(5): 534–541.
[62] Migliavaca CB, Stein C, Colpani V, et al. Meta-analysis of prevalence: I^2 statistic and how to deal with heterogeneity. *Res Synth Methods*. 2022;13(3): 363–367.
[63] López-Ibáñez C, López-Nicolás R, Blázquez-Rincón DM, Sánchez-Meca J. Reliability generalization meta-analysis: comparing different statistical methods. *Curr Psychol*. 2024;43: 18275–18293.
[64] Harbord RM, Higgins JPT. Meta-regression in Stata. *Stata J*. 2008;8(4): 493–519.
[65] Nakagawa S, Santos ESA. Methodological issues and advances in biological meta-analysis. *Evol Biol*. 2012;26: 1253–1274.
[66] White IR. Multivariate random-effects meta-regression. *Stata J*. 2011;11: 255–270.
[67] Jackson D, White IR, Riley RD. Quantifying the impact of between-study heterogeneity in multivariate meta-analyses. *Stat Med*. 2012;31(29): 3805–3820.
[68] Balduzzi S, Rücker G, Nikolakopoulou A, Salanti G, Efthimiou O, Schwarzer G. netmeta: an R package for network meta-analysis using frequentist methods. *J Stat Softw*. 2023;106(2): 1–40.
[69] Jackson D, Barrett JK, Rice S, White IR, Higgins JPT. A design-by-treatment interaction model for network meta-analysis with random inconsistency effects. *Stat Med*. 2014;33(21): 3639–3654.

https://doi.org/10.1017/rsm.2025.10062 Published online by Cambridge University Press

[70] Zhou Y, Dendukuri N. Statistics for quantifying heterogeneity in univariate and bivariate meta-analyses of binary data: the case of meta-analyses of diagnostic accuracy. *Stat Med*. 2014;33(16): 2701–2717.
[71] Hemming K, Hughes JP, McKenzie JE, Forbes AB. Extending the I-squared statistic to describe treatment effect heterogeneity in cluster, multi-centre randomized trials and individual patient data meta-analysis. *Stat Methods Med Res*. 2021;30(2): 376–395.
[72] Vo T-T, Porcher R, Vansteelandt S. Assessing the impact of case-mix heterogeneity in individual participant data meta-analysis: novel use of I^2 statistic and prediction interval. *Res Methods Med Health Sci*. 2021;2(1): 12–30.
[73] Bowden J, Del Greco MF, Minelli C, Davey Smith G, Sheehan NA, Thompson JR . Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression: the role of the I^2 statistic. *Int J Epidemiol*. 2016;45(6): 1961–1974.
[74] Higgins JPT, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *J Roy Statist Soc Ser A*. 2009;172(1): 137–159.
[75] Chiolero A, Santschi V, Burnand B, Platt RW, Paradis G. Meta-analyses: with confidence or prediction intervals? *Eur J Epidemiol*. 2012;27(10): 823–825.
[76] Graham PL, Moran JL. Robust meta-analytic conclusions mandate the provision of prediction intervals in meta-analysis summaries. *J Clin Epidemiol*. 2012;65(5): 503–510.
[77] IntHout J, Ioannidis JP, Rovers MM, Goeman JJ. Plea for routinely presenting prediction intervals in meta-analysis. *BMJ Open*. 2016;6(7): e010247.
[78] Nagashima K, Noma H, Furukawa TA. Prediction intervals for random-effects meta-analysis: a confidence distribution approach. *Stat Methods Med Res*. 2019;28(6): 1689–1702.
[79] Partlett C, Riley RD. Random effects meta-analysis: coverage performance of 95% confidence and prediction intervals following REML estimation. *Stat Med*. 2017;36(2): 301–317.
[80] Stijnen T, White IR, Schmid CH. Analysis of univariate study-level summary data using normal models. In: Schmid CH, Stijnen T, White IR, eds. *Handbook of Meta-Analysis*. CRC Press; 2021: 41–64.
[81] Stogiannis D, Siannis F, Androulakis E. Heterogeneity in meta-analysis: a comprehensive overview. *Int J Biostat*. 2024;20(1): 169–199.
[82] Lin L, Chu H, Hodges JS. Alternative measures of between-study heterogeneity in meta-analysis: reducing the impact of outlying studies. *Biometrics*. 2017;73(1): 156–166.
[83] Ma X, Lin L, Qu Z, Zhu M, Chu H. Performance of between-study heterogeneity measures in the Cochrane Library. *Epidemiol*. 2018;29(6): 821–824.
[84] Crippa A, Khudyakov P, Wang M, Orsini N, Spiegelman D. A new measure of between-studies heterogeneity in meta-analysis. *Stat Med*. 2016;35(21): 3661–3675.
[85] Takkouche B, Khudyakov P, Costa-Bouzas J, Spiegelman D. Confidence intervals for heterogeneity measures in meta-analysis. *Am J Epidemiol*. 2013;178(6): 993–1004.
[86] Cairns M, Prendergast LA. On ratio measures of heterogeneity for meta-analyses. *Res Synth Methods*. 2022;13(1): 28–47.
[87] Cairns M, Cumming G, Calin-Jageman R, Prendergast LA. The diamond ratio: a visual indicator of the extent of heterogeneity in meta-analysis. *Br J Math Stat Psychol*. 2022;75(2): 201–219.

Cite this article: Higgins, JPT, López-López, JA. Reflections on the I-squared index for measuring inconsistency in meta-analysis. *Research Synthesis Methods*. 2026;17: 389–402. https://doi.org/10.1017/rsm.2025.10062

https://doi.org/10.1017/rsm.2025.10062 Published online by Cambridge University Press