CrossMark

Some general points on the I^2-measure of heterogeneity in meta-analysis

Dankmar Böhning[1] · **Rattana Lerdsuwansri**[2] ·
Heinz Holling[3]

Abstract Meta-analysis has developed to be a most important tool in evaluation research. Heterogeneity is an issue that is present in almost any meta-analysis. However, the magnitude of heterogeneity differs across meta-analyses. In this respect, Higgins' I^2 has emerged to be one of the most used and, potentially, one of the most useful measures as it provides quantification of the amount of heterogeneity involved in a given meta-analysis. Higgins' I^2 is conventionally interpreted, in the sense of a variance component analysis, as the proportion of total variance due to heterogeneity. However, this interpretation is not entirely justified as the second part involved in defining the total variation, usually denoted as s^2, is not an average of the study-specific variances, but in fact some other function of the study-specific variances. We show that s^2 is asymptotically identical to the harmonic mean of the study-specific variances and, for any number of studies, is at least as large as the harmonic mean with the inequality being sharp if all study-specific variances agree. This justifies, from our point of view, the interpretation of explained variance, at least for meta-analyses with larger number

✉ Dankmar Böhning
  d.a.bohning@soton.ac.uk

  Rattana Lerdsuwansri
  rattana.lerd@hotmail.com

  Heinz Holling
  holling@psy.uni-muenster.de

[1] Southampton Statistical Sciences Research Institute and Mathematical Sciences,
    University of Southampton, Southampton, UK

[2] Department of Mathematics and Statistics, Faculty of Science, Thammasat University,
    Rangsit Campus, Pathumthani, Thailand

[3] Statistics and Quantitative Methods, Faculty of Psychology and Sports Science,
    University of Münster, Münster, Germany

of component studies or small variation in study-specific variances. These points are illustrated by a number of empirical meta-analyses as well as simulation work.

Keywords Higgins I^2 · Explained variation due to heterogeneity · Meta-analysis · Harmonic mean

1 Introduction and background

Meta-analysis is a statistical methodology for the analysis and integration of results from individual, independent studies. In the last decades, meta-analysis developed a crucial role in many fields of science such as medicine and pharmacy, health science, psychology, and social science (Petitti 1994; Schulze et al. 2003; Böhning et al. 2008; Sutton et al. 2000; Egger et al. 2001). Consider the typical set-up in a meta-analysis: effect measure estimates $\hat{\theta}_1, \ldots, \hat{\theta}_k$ are available from k studies with associated variances $\sigma_1^2, \ldots, \sigma_k^2$, which are assumed known and equal to their sampling counterparts. Typically, the random effects model

$$\hat{\theta}_i = \theta + \delta_i + \epsilon_i$$

is employed where $\delta_i \sim N(0, \tau^2)$ is a normal random effect and $\epsilon_i \sim N(0, \sigma_i^2)$ is a normal random error, all random effects and errors being independent, and $\tau^2 \geq 0$. Furthermore, let $w_i = 1/\sigma_i^2$ and $W_i = 1/(\sigma_i^2 + \tau^2)$. The heterogeneity statistic Q is defined as

$$Q = \sum_{i=1}^{k} w_i (\hat{\theta}_i - \bar{\theta})^2,$$

where $\bar{\theta} = \sum_{i=1}^{k} w_i \hat{\theta}_i / \sum_{i=1}^{k} w_i$. The distribution of Q has been investigated including Biggerstaff and Jackson (2008) and a critical appraisal of Q is given by Hoaglin (2016). More importantly for this work, Q is also the basis of the DerSimonian–Laird estimator for the heterogeneity variance τ^2, which is given, in its untruncated form, by

$$\hat{\tau}^2 = \frac{Q - (k-1)}{\sum_{i=1}^{k} w_i - \sum_{i=1}^{k} w_i^2 / [\sum_{i=1}^{k} w_i]}.$$

It is also the foundation of Higgins' I^2 defined as

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

designed to provide a measure of quantifying the magnitude of heterogeneity involved in the meta-analysis (Higgins and Thompson 2002; Borenstein et al. 2009). Higgins' I^2 is very popular and has been discussed intensively including critical appraisals (its

dependence on the study-specific precision) given in Rücker et al. (2008) or, more recently, in Borenstein et al. (2017). I^2 has also been recently generalized to the multivariate context, see Jackson et al. (2012). I^2 is indeed a proportion and, if multiplied by 100, a percentage. In addition, I^2 has a variance component interpretation. The variance of the effect measure $Var(\hat{\theta}_i) = \tau^2 + \sigma_i^2$ can be partitioned into the within-study variance and the variance between studies. If all studies would have the same within-study variance $\sigma_i^2 = \sigma^2$, then it would be easy to define with $\tau^2/(\tau^2 + \sigma^2)$ the proportion of variance due to across-study variation, or simply due to heterogeneity. With study-specific variances an average variance needs to be used and, if a specific average is selected, then I^2 can be interpreted as the proportion of variation due to heterogeneity. This might be not obvious from the definition provided in (1) but becomes more evident from the identity (although this can be found elsewhere a proof of this identity is given in the Appendix for completeness)

$$I^2 = \frac{\hat{\tau}^2}{\hat{\tau}^2 + s^2},\tag{2}$$

where $s^2 = (k-1)\sum_{i=1}^{k} w_i / [(\sum_{i=1}^{k} w_i)^2 - \sum_{i=1}^{k} w_i^2]$. If s^2 could be viewed as some sort of average of the study-specific variances $\sigma_1^2, \ldots, \sigma_k^2$, then I^2 could be validly interpreted, as typically done in variance component models, as the proportion of the total variance (variance due to heterogeneity plus within-study variance).

This short note serves two purposes:

- we will show that, under mild regularity assumptions, s^2 is asymptotically identical to the harmonic mean $\bar{\sigma}^2 = \left[\frac{1}{k}\sum_{i=1}^{k} w_i\right]^{-1}$ of the study-specific variances.
- we will show that $s^2 \geq \bar{\sigma}^2$ with the difference $s^2 - \bar{\sigma}^2$ being zero if all study-specific variances are identical, and for the more general case of non-identical study-specific variances, approaching zero for k becoming large.

2 Main results

2.1 The harmonic mean result

We have the following result:

Theorem 1 *If there exist positive constants b and B such that for all $i = 1, 2, \ldots, k$*

$$0 < b \leq \sigma_i^2 \leq B < \infty,$$

then

$$|s^2 - \bar{\sigma}^2| \to 0 \quad for \ k \to \infty.$$

Proof We can write s^2 as

$$s^2 = \frac{(k-1)/k}{\frac{1}{k}\sum_i w_i - [\frac{1}{k}\sum_i w_i^2]/[\sum_i w_i]}.$$

As $(k-1)/k \to 1$ for $k \to \infty$, it is sufficient to show that

$$\frac{\frac{1}{k}\sum_i w_i^2}{\sum_i w_i} \to 0$$

as $k \to \infty$. As σ_i^2 is bounded below by b for all i, we have that

$$w_i^2 \leq 1/b^2$$

for all i, so that

$$\frac{1}{k}\sum_i w_i^2 \leq 1/b^2. \tag{3}$$

In addition, as σ_i^2 is bounded above by B for all i, we have that

$$w_i \geq 1/B$$

so that

$$\frac{1}{\sum_i w_i} \leq B/k. \tag{4}$$

Taking (3) and (4) together yields

$$\frac{\frac{1}{k}\sum_i w_i^2}{\sum_i w_i} \leq \frac{1}{k}\frac{B}{b^2} \to 0$$

for $k \to \infty$. This ends the proof.   □

2.2 The inequality

Further clarification on the relation between s^2 and $\bar{\sigma}^2$ is given by the following inequality. Note that this result does not require any assumption on the variances $\sigma_1^2, \ldots, \sigma_k^2$.

Theorem 2

$$s^2 \geq \bar{\sigma}^2,$$

with equality if

$$\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2.$$

Proof We need to show that

$$s^2 = (k-1) \sum_{i=1}^{k} w_i \Bigg/ \left[\left(\sum_{i=1}^{k} w_i \right)^2 - \sum_{i=1}^{k} w_i^2 \right] \geq k / \sum_i w_i.$$

This is equivalent to showing that

$$\frac{k-1}{1 - \dfrac{\sum_i w_i^2}{\left(\sum_i w_i\right)^2}} \geq k,$$

or

$$\frac{\sum_i w_i^2}{\left(\sum_i w_i\right)^2} \geq 1/k.$$

This is equivalent to

$$E(V^2)/[E(V)^2] \geq 1,$$

for a random variable V giving equal weights to $w_1, \ldots, w_k$. This inequality holds as

$$Var(V) = E(V^2) - E(V)^2 \geq 0,$$

with equality if $w_1 = w_2 = \cdots = w_k$, and this ends the proof. □

3 Empirical illustrations

Here, we illustrate these relationships on the basis of 15 meta-analyses. Details on these are given in Table 1. These meta-analyses were not selected in any particular way, they were simply collected from the literature while teaching a course on statistical methods for meta-analysis.

It is clear from Theorem 1 that the difference between s^2 and $\bar{\sigma}^2$ should become smaller with increasing number of studies k. In Fig. 1, we examine $\log(s^2/\bar{\sigma}^2)$ in dependence on k. Here we are taking the log-ratio of s^2 and $\bar{\sigma}^2$ to remove any scaling factor variation across different meta-analyses. There is a clear decreasing trend for $\log(s^2/\bar{\sigma}^2)$ with increasing k.

In Fig. 2, we examine $\log(s^2/\bar{\sigma}^2)$ vs. $\sqrt{\frac{1}{k-1} \sum_i (\sigma_i^2 - \tilde{\sigma}^2)^2}/\tilde{\sigma}^2$ where $\tilde{\sigma}^2 = \frac{1}{k} \sum_i \sigma_i^2$. Here we are considering coefficient of variation as measure of variability of the within-study variances. Again we need to take the coefficient of variation (in contrast to the standard deviation) to remove scale variation across different meta-analyses. There is a clear increasing trend for $\log(s^2/\bar{\sigma}^2)$ with increasing variability of the study-specific variances involved in the meta-analysis.

Table 1 Large details of the 15 meta-analyses used in illustrating the asymptotic result of Theorem 1

ID	Name	k	References
1	Antihistamine	9	Clinic. Pharmacol. Therapeut. 58, 605–616
2	bcgtrial	13	JAMA, 271, 698–702
3	Caesarean	60	Cochrane Review 2001, 3
4	CALGB	21	Journ. Clinical Epidem. 11, 155–150
5	Catheter	17	Intens. Care Med. 33, 2058–2068
6	Cholesterol	28	British Medical Journal, 306, 1367–1373
7	GSTT1_prostate	39	Americ. Journ. Epidem. 180, 1–10
8	Lidocaine	6	Arch. Intern. Medicine, 149, 2694–2698
9	Ectiopic_pregnancy	27	Human Reproduction, 13, 3220–3227
10	ma_ps_lc	21	Int Journ. Epidem. 36, 1048–1059
11	Endometrial_cancer	56	Acta Obstetr. Gynecol. Scand. 81, 799–816
12	Multiplesclerosis	12	Neurol. Neurosurg. Psychiatry 87, 324–331
13	Smoking	26	BMC Public Health, 6, 300
14	Strepto_cma	22	New England Journ. Med., 327, 248–254
15	xrcise4deprsn	10	British Medical Journal, 322, 763
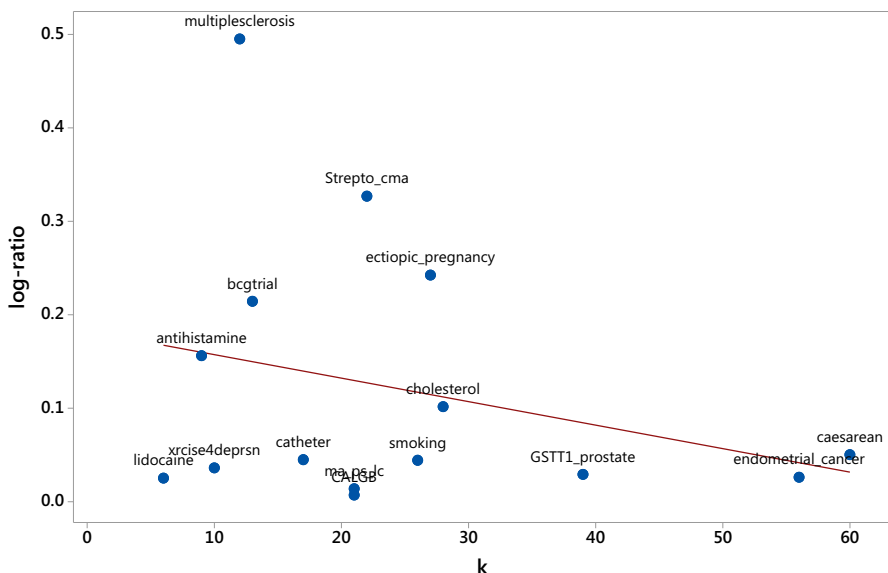
Fig. 1 Scatter plot with regression line of $\log(s^2/\bar{\sigma}^2)$ versus k for 15 different meta-analyses; the correlation coefficient is -0.29

4 A simulation study

To further investigate these findings, we have undertaken the following simulation work. We assume that study-specific variances differ only by the study size. Hence
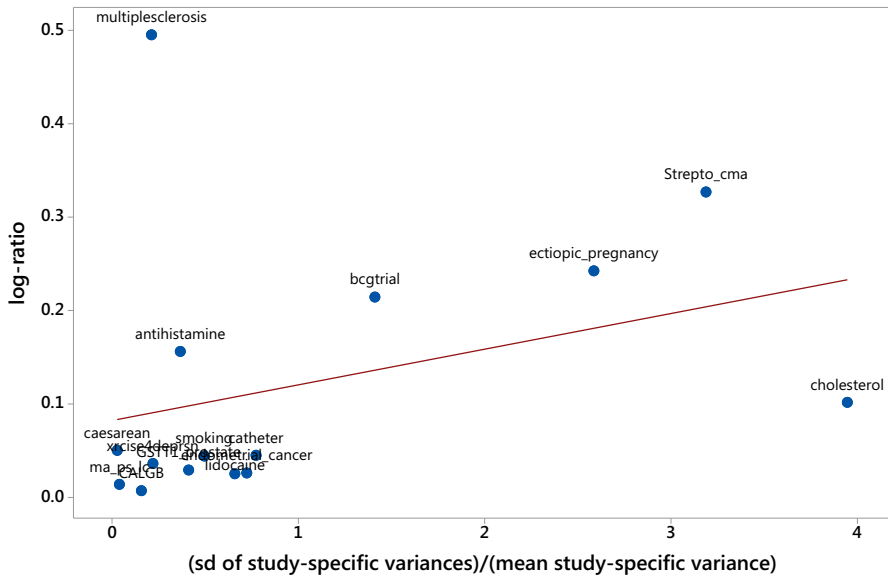
Fig. 2 Scatter plot with regression line of $\log(s^2/\bar{\sigma}^2)$ versus standard deviation of $\sigma_1^2, \ldots, \sigma_k^2$ divided by their mean for 15 different meta-analyses; the correlation coefficient is 0.33

$\sigma_i^2 = \sigma^2/n_i$ where n_i is the sample size of study i. We take $\sigma^2 = 1$ so that $w_i = \frac{1}{\sigma_i^2} = n_i$. We consider three settings in which $w_i = n_i$ is sampled from a uniform with common mean 55 but different range:

- setting 1: $w_i = n_i \sim Uniform(10, 100)$ for $i = 1, \ldots, k$ is arising from a uniform distribution on [10, 100];
- setting 2: $w_i = n_i \sim Uniform(30, 80)$ for $i = 1, \ldots, k$ is arising from a uniform distribution on [30, 80];
- setting 3: $w_i = n_i \sim Uniform(45, 65)$ for $i = 1, \ldots, k$ is arising from a uniform distribution on [45, 65].

Setting 3 has the smallest range whereas setting 1 has the largest, and setting 2 is in between. For $i = 1, \ldots, k$ the following measures are calculated:

1. the arithmetic mean $\tilde{\sigma}^2 = \frac{1}{k} \sum_i \sigma_i^2 = \frac{1}{k} \sum_{i=1}^{k} w_i^{-1}$ of the study-specific variances;
2. the harmonic mean $\bar{\sigma}^2 = [\frac{1}{k} \sum_{i=1}^{k} w_i]^{-1}$ of the study-specific variances;
3. $s^2 = (k-1) \sum_{i=1}^{k} w_i / [(\sum_{i=1}^{k} w_i)^2 - \sum_{i=1}^{k} w_i^2]$;
4. $\log(s^2/\bar{\sigma}^2)$;
5. the coefficient of variation $CV = \sqrt{\frac{1}{k-1} \sum_i (\sigma_i^2 - \tilde{\sigma}^2)^2} / \tilde{\sigma}^2$ where $\tilde{\sigma}^2 = \frac{1}{k} \sum_i \sigma_i^2$.

This process has been repeated 10,000 times and the mean of the above performance measures calculated. Figure 3 shows a scatter plot of $\log(s^2/\bar{\sigma}^2)$ vs. the number of studies k involved in the meta-analysis. It can be seen that for k larger than 20 the difference between s^2 and the harmonic mean becomes rather small. Of course, this occurs much earlier for the small-range setting 3 and the moderate-range setting 3.
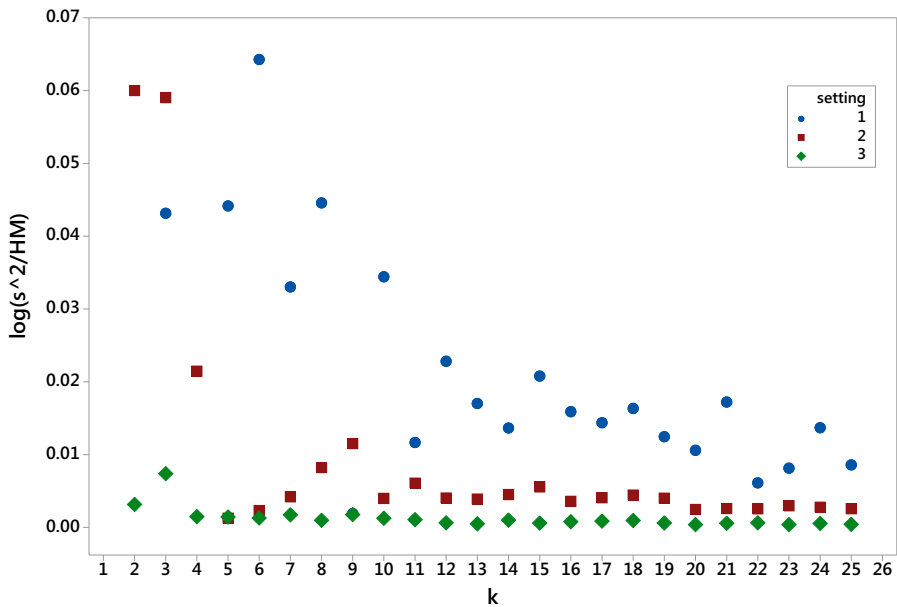
Fig. 3 Scatter plot of $\log(s^2/\bar{\sigma}^2)$ versus the number of studies k involved in the meta-analysis for the three settings of the simulation study as described in Sect. 4

This illustrates the result of Theorem 1 and also indicates when the limit is approached with acceptable approximation. Figure 4 shows that $\log(s^2/\bar{\sigma}^2)$ also depends strongly on the variability of the study-specific variances involved in the meta-analysis. Clearly, the smaller the coefficient of variation the closer to one is also the ratio of s^2 to $\bar{\sigma}^2$. This illustrates Theorem 2.

5 Discussion and conclusion

It was seen that s^2 approximates the harmonic mean $[\frac{1}{k} \sum_{i=1}^{k} w_i]^{-1}$. This asymptotic result has been achieved under fairly mild assumptions (study-specific variances need to be bounded). As a referee gratefully points out, Jackson and Bowden (2009, Sect. 3.1) also mention this result without proof in a simulation study and it remains unclear, in the work of Jackson and Bowden, how general the result is meant to hold. Some general conditions are given here, in Theorem 1, under which the asymptotic result is valid.

The harmonic mean appears to be a more reasonable summary measure of the study-specific variance than the arithmetic mean. To make this point more clear consider that situation that $\sigma_i^2 = \sigma^2/n_i$ with $\sigma^2 = 1$ for simplicity. Let us further assume that σ_i^2 is estimated by s_i^2/n_i where s_i^2 is the study-specific variance of study i. Then, under normality, $Var(s_i^2)$ is proportional to $1/n_i$, in other words, $Var(s_i^2/n_i)$ is proportional to $1/n_i^3$. As an optimal meta-analytic weighting scheme uses inverse variances the optimal summary measure would be
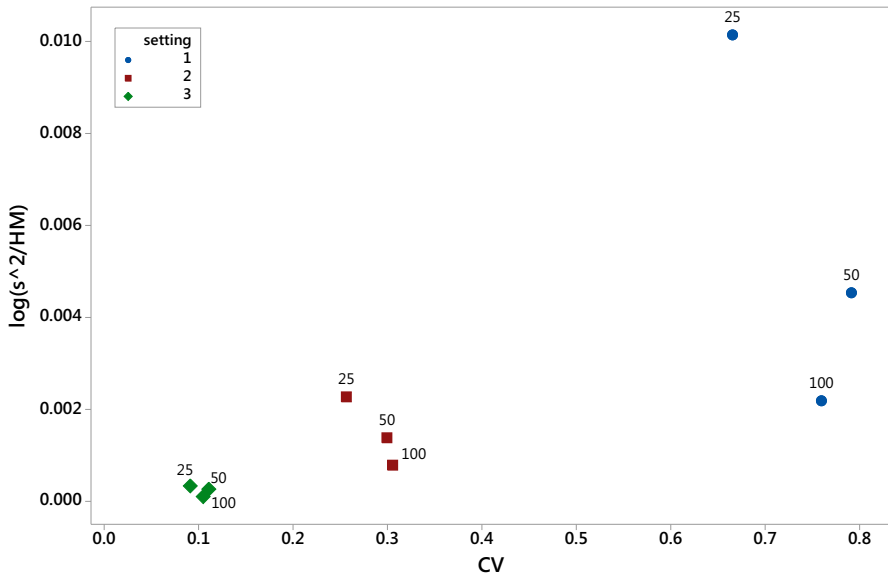
Fig. 4 Scatter plot of $\log(s^2/\bar{\sigma}^2)$ versus the coefficient of variation (CV) of the study-specific variances involved in the meta-analysis for the three settings of the simulation study as described in Sect. 4; the numbers next to the symbols in the graph indicate the number of studies involved in the meta-analysis

$$\frac{\sum_i n_i^3 s_i^2/n_i}{\sum_i n_i^3} \approx \frac{\sum_i n_i^2}{\sum_i n_i^3}. \tag{5}$$

The harmonic mean $[\frac{1}{k}\sum_{i=1}^{k} n_i]^{-1}$ is a sharp upper bound for the RHS in (5), with the bound being sharp if all sample sizes agree. Hence we can understand the harmonic mean as an approximation of the optimal weighted summary of the study-specific variances.

Higgins' I^2 can be validly interpreted as the proportion of variance due to heterogeneity if the variance of the study-specific variances are small and/or the number of component studies is moderately large. This provides a valuable way of interpreting any value of I^2. If, however, the number of component studies are small with considerable variation in the variances of the component studies it may be advisable to consider a more direct measure for the proportion due to heterogeneity.

Evidently, there is interest in developing a measure of heterogeneity which has an interpretation as a 'proprotion of total variation due to heterogeneity'. In this line, Wetterslev et al. (2009) suggest a measure of heterogeneity, denoted by D^2, involving, besides τ^2, a new measure of average sampling error. The basis of D^2 is the ratio of the harmonic mean of the study-specific variance increased by the heterogeneity variance to the harmonic mean of the study-specific variances. This new measure of heterogeneity seems to have a lot of rationale in its construction: however, it remains to be seen if it establishes itself as an alternative to Higgins' I^2.

In conclusion, a measure of heterogeneity of the form $\tau^2/(\tau^2 + \bar{\sigma}^2)$, with $\bar{\sigma}^2$ being the harmonic mean of the study-specific variances, seems to be a reasonable good

choice. It seems also possible to extend this most straight forward definition of I^2 to meta-regressive contexts. Clearly, estimating I^2 as $\frac{Q-(k-1)}{Q}$ (and not as $\hat{\tau}^2/(\hat{\tau}^2 + \bar{\sigma}^2)$) has considerable benefits, in particular as it is easily carried forward to meta-analytic regression approaches. The results given here justify that $\frac{Q-(k-1)}{Q}$ can still be interpreted as proportion of explained variance due to heterogeneity. In fact, it is the latter way of defining I^2 which allow Jackson et al. (2012) to extend I^2 to a multivariate meta-analytic context. Here, multivariate meta-analysis is understood in the dense of having several outcome measures or endpoints simultaneously available for meta-analysis. In the approach of Jackson et al. (2012), Q is first extended to a quadratic form incorporating all outcome measures, and then, in a second step, I^2 is then defined as $\frac{Q-\nu}{Q}$ where ν then corresponds to the degrees of freedom involved in the quadratic form. The concept of 'explained variance due to heterogeneity' is evidently more difficult to generalize as within-study and between-study variation involve covariance matrices due to the multivariate nature of the meta-analysis. This area is clearly of great interest for future work.

Appendix

Theorem 3 *Let I^2 be defined as $I^2 = \frac{Q-(k-1)}{Q}$. Then*

$$I^2 = \frac{\hat{\tau}^2}{\hat{\tau}^2 + s^2}$$

where $s^2 = (k-1) \sum_{i=1}^{k} w_i / [(\sum_{i=1}^{k} w_i)^2 - \sum_{i=1}^{k} w_i^2]$.

Proof We have that

$$
\begin{aligned}
I^2 &= \frac{Q-(k-1)}{Q} = \frac{Q-(k-1)}{Q-(k-1)+(k-1)} \\
&= \frac{Q-(k-1)}{Q-(k-1)+(k-1)} \frac{\sum_i w_i - [\sum_i w_i^2]/\sum_i w_i}{\sum_i w_i - [\sum_i w_i^2]/\sum_i w_i} \\
&= \frac{Q-(k-1)}{\sum_i w_i - [\sum_i w_i^2]/\sum_i w_i} \frac{\sum_i w_i - [\sum_i w_i^2]/\sum_i w_i}{Q-(k-1)+(k-1)} \\
&= \frac{\hat{\tau}^2}{\hat{\tau}^2 + \frac{k-1}{\sum_i w_i - [\sum_i w_i^2]/\sum_i w_i}} \\
&= \frac{\hat{\tau}^2}{\hat{\tau}^2 + s^2}
\end{aligned}
$$

which ends the proof. $\qquad\square$

References

Biggerstaff BJ, Jackson D (2008) The exact distribution of Cochran's heterogeneity statistic in one-way random-effects meta-analysis. Stat Med 27:6093–6110

Böhning D, Kuhnert R, Rattanasiri S (2008) Meta-analysis of binary data using profile likelihood. Chapman & Hall, London

Borenstein M, Hedges LV, Higgins JPT, Rothstein HR (2009) Introduction to meta-analysis. John Wiley & Sons, Chichester

Borenstein M, Higgins JPT, Hedges LV, Rothstein HR (2017) Basics of meta-analysis: I^2 is not an absolute measure of heterogeneity. Res Synth Methods 8:5–18

Egger M, Smith GD, Aliman DG (2001) Systematic reviews in health care: meta-analysis in context. BMJ Publishing Group, London

Hoaglin DC (2016) Misunderstandings about Q and 'Cochran's Q test' inn meta-analysis. Stat Med 35:485–495

Higgins JPT, Thompson SG (2002) Quantifying heterogeneity in meta-analysis. Stat Med 21:1539–1558

Jackson D, Bowden J (2009) A re-evaluation of the 'quantile approximation method' for random effects meta-analysis. Stat Med 28:338–348

Jackson D, White IR, Riley RD (2012) Quantifying the impact of between-study heterogeneity in multi-variate meta-analysis. Stat Med 31:3805–3820

Petitti DB (1994) Meta-analysis decision analysis and cost-effectiveness analysis. Oxford University Press, New York

Rücker G, Schwarzer G, Carpenter JR, Schumacher M (2008) Undue reliance on I^2 in assessing hetero-geneity may mislead. BMC Med Res Methodol 8:79

Schulze R, Holling H, Böhning D (2003) Meta-analysis: new developments and applications in medical and social sciences. Göttingen, Hogrefe & Huber

Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F (2000) Methods for meta-analysis in medical research. Wiley, New York

Wetterslev J, Thorlund K, Brok J, Gluud C (2009) Estimating required information size by quantifying diversity in random-effects model meta-analyses. BMC Med Res Methodol 9:86