Author's response to reviews

Title: The heterogeneity statistic I2 can be biased in small meta-analyses

Authors:

Paul T von Hippel (paulvonhippel.utaustin@gmail.com)

Version: 5 Date: 5 March 2015

Author's response to reviews: see over
March 5, 2015

To the Editor:

Thank you for your suggestions regarding my manuscript entitled “The heterogeneity statistic $I^2$ can be biased in small meta-analyses.” As requested, I have clarified the definitions of fixed and random effects.

The terms are now introduced with the following paragraphs:

Two models can be used in meta-analysis: a fixed-effects model and a random-effects model. Some confusion is possible because the term fixed effects is used in two different senses [15]. In some literature, the term fixed effects means that the $K$ study effects are assumed to be equal—i.e., homogeneous. We use the term fixed effects in its other sense, where it means that we seek only to generalize about the $K$ studies in the meta-analysis. The true effects can be either homogeneous or heterogeneous, but they are regarded as fixed quantities. While the $K$ studies would produce different estimates and if they were repeated, the true effects and true standard errors would not change.

While this definition may be less common, you may recall that it was the definition suggested earlier by Referee 1, Wolfgang Viechtbauer. In fact, Viechtbauer’s suggestion that I calculate the bias under both a fixed and random-effects model hinged on the term fixed effect being used in this way.

The revision continues with this definition of random effects:

Under a random-effects model, by contrast, we assume that the true effects in the meta-analysis were drawn at random from a larger population of effects, and we seek to make inferences about that larger population [16]. So the are not fixed quantities but random variables that would be different if a different sample were drawn from the population of effects.

This revision reflects your request that I follow Higgins et al. in regarding only the study effects as random. The previous version reflected Hedges’ alternative idea that the studies themselves are random.

This change required that a later paragraph be rewritten as follows:
The variance in \( \tau_0^2 \) is partly due to the heterogeneity of the true effects and partly due to estimation error summarized by the standard errors. By the law of total variance we have

\[
\text{(1)}.
\]

where \( \tau_0^2 \) is the heterogeneity variance or between-study variance, and \( \tau_1^2 \) is the average within-study variance. Under a fixed-effects model these variances and expectations refer only to the \( K \) effects and standard errors in the meta-analysis. Under a random effects model \( \tau_0^2 \) refers to the larger population of effects, but still refers only to the \( K \) standard errors in the meta-analysis, unless we are willing to regard the \( \tau_0^2 \) as well as the \( \tau_1^2 \) as samples from a larger population.

I also changed an exclamation mark to a period, as you requested. Aside from that, the article is unchanged. I hope these clarifications are satisfactory to you.

Sincerely yours,

Paul von Hippel
Comments of Referee 1 (Oliver Kuss)
In his short note, the author calculates (by using straightforward elementary methods) the expectation and bias of the popular \( I^2 \) statistic which is the standard tool to report heterogeneity in meta-analysis. He shows that \( I^2 \) is biased especially in situations with a small number of studies and small heterogeneity. The paper is written pleasantly clear and short and I especially liked the author's recommendation of NOT reporting the \( I^2 \) estimate but only its confidence interval in situations where a large bias is to be expected.

*Thank you. I thought the suggestion of omitting the point estimate was fairly radical, so it is nice to know that it appealed to at least one reader.*

I definitely would like to see the paper published, however, I would like to see the author’s comments on the following two issues.

Major compulsory Revisions:
- I feel that the author's evidence on the typical Cochrane meta-analysis (Engels et al. 2000; Ioannidis, Patsopoulos, and Evangelou 2007) is somewhat outdated. A lot happened since 2007 or even 2000, and I wonder if there won't be some more recent evidence. For example, Rebecca Turner (2012) reports the median number of studies in 14,886 meta-analyses to be 3.

  *AUTHOR: Thank you for this reference. I cite Turner’s (2012) article in the revision, as well as Davey’s (2011) equally recent summary. Turner reports that the median number of studies per meta-analysis is 3; Davey reports that it is 6. Naturally the biases that I report are worse when the number of studies per meta-analysis is smaller.*

- I wondered if the two values of \( I^2 \) where there is no bias (one around 0.2 and one around 0.8) can be given exactly.

  *AUTHOR: Unfortunately, no. The calculations require numerical integration and, although the results can be made arbitrarily precise, they can never be exact.*

Is there a heuristic (or even an exact proof) that there must be exactly two values of \( I^2 \) where the bias disappears?

  *AUTHOR: The revision offers an informal explanation in which there are two sources of bias which are opposite in sign and cancel out under some circumstances.*

Comments of Referee 2 (Wolfgang Viechtbauer)
This is a short but interesting communication regarding the \( I^2 \) statistic that is commonly reported in meta-analyses. With the help of Mathematica, the author derives/computes the expected value of \( I^2 \) under homogeneity and heterogeneity and shows it to be biased. For reasons described in the paper, this is hardly surprising under homogeneity (due to the truncation of negative values).

  *AUTHOR: It may be unsurprising that \( I^2 \) is biased due to truncation. However, the size of the bias is not obvious and required some calculation. The size of the bias may be surprising to some readers. It is large enough to raise concerns about the routine use of \( I^2 \) in small meta-analyses.*

The results under heterogeneity are, as far as I can tell, novel (although some references to relevant related work are missing -- see comments below).

  *AUTHOR: Thank you for the references which I have incorporated into the revision. Some of the ideas in the referenced works were very helpful in extending and clarifying the results.*
Just as a side-note: I am glad to see a clear distinction being made between the estimator $I^2$ and the underlying true value (denoted as $I^2$ by the author).

**AUTHOR:** Thank you. Yes, I view that distinction as an important conceptual contribution.

**Major compulsory Revisions:**
1) Under heterogeneity, $Q$ is only non-central chi-squared under a fixed-effects model (note that I am talking about a model with potentially different but fixed study-specific true effects, not the fixed-effect -- without 's' -- model that assumes that the true effects are homogeneous; for the distinction, see Hedges & Vevea (1998). Under a random-effects model, the non-null distribution is different. Hedges and Pigott (2001) only describe the non-null distribution for this case when all of the sampling variances are equal to each other. The exact non-null distribution in the more general case (without assuming homoscedasticity) was derived by Biggerstaff and Jackson (2008). They also describe the exact distribution of $I^2$ and its expectation. Some earlier work in this regard was done by Mittlböck and Heinzl (2006). Reference should be made to this work. Ideally, the case of the non-null distribution under a random-effects model should also be covered in the present paper.

**AUTHOR:** Thank you for these references, which I have incorporated into the revision. The revision covers both fixed and random effects models.

2) In the section on "Notation, assumptions, and distributions", the author writes that "$Q$ has approximately a central chi-square distribution". Why only approximately? Please be more specific about this.

3) Similarly, in the section on "Notation, assumptions, and distributions", the author writes that $Q$ has asymptotically a non-central chi-square distribution. In what sense "asymptotically"? Again, please be more specific about this.

**AUTHOR:** This is an application of the central limit theorem, which we make more explicit in the text. The statistic $Q = \sum_{k=1}^{K} Z_k^2$, where $Z_k = \left( \hat{\beta}_k - \hat{\beta} \right) / \hat{\sigma}_k$, would have exactly a chi-square distribution if the summand $Z_k$ were standard normal. $Z_k$ would be standard normal if the point estimate $\hat{\beta}_k$ were normally distributed and the standard errors $\hat{\sigma}_k$ were known with certainty. In fact, the standard errors are only estimates, and for some estimands the estimate $\hat{\beta}_k$ may not be normally distributed if the study sample size $n_k$ is small. So the exact distribution of $Z_k$ is not standard normal, and the exact distribution of $Q$ is not chi-square. But as $n_k$ gets larger, the estimate $\hat{\beta}_k$ approaches normality and the error in the estimate $\hat{\sigma}_k^2$ shrinks toward zero, so $Z_k$ approaches standard normality and $Q$ approaches chi-square.

In the revision, we don’t go into this much detail, which would be distracting. But we do say that the distribution of $Q$ is chi-square “if $n_k$ is large.”

**Minor essential revisions:**
4) It is worth noting that under homogeneity, $1/Q$ follows an inverse chi-squared distribution. Since the moments of the inverse chi-squared distribution can be given in closed form, some basic properties of $I^2$ can be easily derived (truncation of negative values introduces an additional level of complexity here). See also Mittlböck and Heinzl (2006).
AUTHOR: Yes, this is noted in the revision. It is a starting point that helps to build intuition, though things get more complicated when I introduce heterogeneity, truncation, and fixed effects.

5) Why was .05 chosen as the value of $\tau^2$ in Figure 2 and values of .1 to .9 in Figure 3? Do these values reflect actual $\tau^2$ values as seen in practice? And if yes, for which outcome measures?

AUTHOR: This is a misunderstanding. The values given in Figures 2-3 refer to $i^2$ not $\tau^2$. As $i^2$ can only take values between 0 and 1, the range of values considered in the paper ($i^2 = 0,.05,.1,.2,.3,.4,.5,.6,.7,.8,.9$) is fairly comprehensive.

6) In the conclusions, the author suggests to report CIs around $I^2$ and refers to the 'best methods'. Unfortunately, no reference is made to the two methods that are actually best, in the sense of providing exact CIs under the random-effects model. The Q-profile method (Hartung and Knapp 2005; Viechtbauer 2007) provides an exact CI for $\tau^2$, whose bounds can be easily plugged into the equation for computing $I^2$, which then provides an exact CI for $i^2$. A different but also exact CI for $\tau^2$ can be obtained with the method by Biggerstaff and Jackson (2008) and Jackson (2013), which again leads to an exact CI for $i^2$. See also Jackson et al. (2014), which generalizes these ideas to mixed-effects meta-regression models (with the random-effects model just being a special case).

AUTHOR: Thank you for these very helpful references, which I have added to the revision. I'm not sure that an exact CI for $\tau^2$ is sufficient to get an exact CI for $i^2$. Uncertainty about $i^2$ depends not just on uncertainty about the between-study variance $\tau^2$ but also on uncertainty about the within-study variances $\sigma_k^2$.

7) Some additional words of caution regarding the interpretation of $I^2$ should be added. See Rücker et al. (2008).

AUTHOR: Reference to Rücker et al. (2008) has been added.

Discretionary revisions
There is nothing wrong with the present notation, but my personal preference (and what I see most commonly used in publications) is to use "i" as the index for studies and k (not capitalized) as the number of studies.

AUTHOR: Thank you. If it's all right with you I will leave the notation as it stands. If I changed it at this point, I would probably miss something and introduce error.

Comments of the Associate Editor
1. Please have a look at another paper with a similar focus (Melsen et al. 2014).

AUTHOR: Thank you for this suggestion. The main finding of Melsen's article is that, as $I^2$ increases, meta-analyses are less accurate at estimating the true average effect size or predicting the result of a new intervention. These results highlight the importance of heterogeneity. The article is cited in the revision.

2. In addition to a recommendation by reviewer Oliver Kuss, other more recent empirical studies on properties of Cochrane reviews worth having a look at are (Davey et al. 2011; Turner, Bird, and Higgins 2013).

AUTHOR: Thank you for these references, which are cited in the revision.

3. Related to point 7 by reviewer Wolfgang Viechtbauer, I have a more fundamental problem with the approach in the manuscript. It is assumed that in the population of studies there is
some parameter $\tau^2$ that is estimated by $I^2$. The assumption that $\tau^2$ is fixed means that $\tau^2/(\tau^2+\sigma^2)$ is fixed (eq. (4) of the manuscript), where $\sigma^2$ is the ‘typical’ sampling variance of a study in the meta-analysis (Higgins and Thompson 2002). This would be ideally the case if $\sigma^2$ were constant across studies, that is, if all studies had the same standard error (e.g., the same standard deviation and the same size). In other words, given $\tau^2$, the perception of $\tau^2$ as a population parameter (instead of $\tau^2$) puts an unnecessary restriction on the sample size of the trials. $I^2$ is straightforwardly derived from Q which is a test statistic, and so is $I^2$, like all statistics dependent on sample sizes (numbers of patients within studies). For this reason, I don’t understand what the meaning of a ‘true’ $I^2$ is, as little as I know what a true Q or a true p-value is.

**AUTHOR:** An estimate such as $l^2$ only has meaning if it refers to some well-defined estimand. In this article, we define the estimand as $l^2 = \tau^2/(\tau^2 + \sigma^2)$, which is interpreted as the fraction of variance in the estimates that is due to heterogeneity rather than sampling error. Reviewer 2 (Viechtbauer) was “glad to see a clear distinction being made between the estimator $I^2$ and the underlying true value (denoted as $l^2$ by the author).” However, your comment indicates that some clarification of $l^2$ is needed. We offer clarification in the revision.

Under a fixed-effects model it is not necessary to interpret $l^2$ as a parameter referring to a larger population. Instead, $l^2$ is simply the fraction of variance that is due to heterogeneity among the estimates from the K studies in the meta-analysis. Under a random-effects model, we regard the K studies as sampled from a larger population of studies that have a heterogeneity variance of $\tau^2$ and an average sampling variance of $\sigma^2$. Each study can have its own sampling variance $\sigma_k^2$ so no restriction is put on the sample sizes $n_k$.

It is common to suppose that $l^2$ is only defined if every study has the same sampling variance $\sigma^2$. Indeed, this is how Higgins and Thompson (2002) originally defined $l^2$. In the revision, however, we show that $l^2$ is still defined in the situation where each of the K studies has its own sampling variance $\sigma_k^2$, k=1,…K. In that case, the definition is still $l^2 = \tau^2/(\tau^2 + \sigma^2)$ but $\sigma^2$ is defined as $\sigma^2 = E(\sigma_k^2)$. The interpretation of $l^2$ is the same. It is the fraction of variance in the estimates that is due to heterogeneity rather than sampling error.

See reference Rücker et al. (2008) given by Viechtbauer and also a letter by the same authors (Rücker et al. 2009). The heterogeneity parameter I would prefer to consider is $\tau^2$ that does not depend on the number of studies, nor the number of patients in studies. Please discuss this issue.

**AUTHOR:** $\tau^2$ and $l^2$ are different quantities. $\tau^2$ is the heterogeneity variance and $l^2$ is the fraction of total variance that is due to heterogeneity. Both quantities are useful for different purposes. Rücker et al. (2008, 2009) are not arguing that $l^2$ should never be used; they are arguing that $l^2$ should not be used to support decisions and interpretations for which $\tau^2$ would be more appropriate. As discussed in the revision, estimates of both $\tau^2$ and $l^2$ continue to be reported, not just in meta-analyses, but in other studies where it is important to know how much of the variance in a set of estimates is due to heterogeneity (von Hippel et al. 2014; Koedel 2009; Koedel et al. 2012). It is important to understand the statistical properties of those estimates.
4. Cochrane reviews are repeatedly mentioned in the manuscript, sometimes as `Cochrane reviews', sometimes (correctly) as Cochrane Library. This should be unified, the Cochrane Library should be cited and the url given as http://www.thecochranelibrary.com.

**AUTHOR:** The revision does this.

**References**


Jackson, Dan, Rebecca Turner, Kirsty Rhodes, and Wolfgang Viechtbauer. 2014. “Methods for Calculating Confidence and Credible Intervals for the Residual between-Study Variance


