January 2007

I2 Is Subject to the same statistical power problems as Cochran's Q

Tania Huedo-Medina
University of Connecticut, tania.huedo-medina@uconn.edu

Blair T. Johnson
University of Connecticut, blair.t.johnson@uconn.edu

I^2 Is Subject to the Same Statistical Power Problems as Cochran's Q

Tania Huedo-Medina, *post-doctoral fellow*[1,2], **Blair T Johnson**, *professor of psychology*[1]

[1]Center for Health, Intervention, and Prevention (CHIP), University of Connecticut, 2006 Hillside Road, Unit 1248, Storrs, CT 06269-1248 USA. [2]Correspondence: tania.huedo-medina{at}uconn.edu.

In their popular article, Higgins and colleagues (2003) provided a valuable explanation about the importance of assessing heterogeneity in overall meta-analytic findings and how their new, I^2, index helps scholars to attain this goal. As these authors review, there are three general ways to assess heterogeneity in meta-analysis, but each has a liability for interpretation. First, one can assess the between-studies variance, τ^2, but its values depend on the particular effect size metric used, along with other factors. The second is Cochran's Q, which follows a chi-square distribution to make inferences about the null hypothesis of homogeneity. (It is actually not a test of *heterogeneity*, as Higgins and colleagues assert, but of the hypothesis of homogeneity.) The problem with Q is that it has a poor power to detect true heterogeneity when the number of studies is small. Because neither of these first two methods has a standardized scale, they are poorly equipped to make comparisons of the degree of homogeneity across meta-analyses.

The third and final way to assess the heterogeneity is calculating a scale-free index of variability. The Birge ratio, originated in 1932, has been the most commonly used scale-free index to quantify the consistency of study findings; it is defined as the ratio of a chi-square to its degrees of freedom. Because the degrees of freedom are the expected value of each chi-square, when the chi-square shows only random variation, the Birge ratio is close to 1.00. Thus, to the extent that the Birge ratio exceeds 1.00, results of a set of studies lack homogeneity. That is, they are more varied than one can expect based merely on sampling error.

Higgins and Thompson (2002; Higgins et al., 2003) extended the Birge ratio to the I^2 index in an effort to overcome the shortcomings of Q and τ^2. Like the Birge ratio, the I^2 index is a scale-free index of variability in defining the ratio of Q in relation to its degrees of freedom. The advantage of this new index is its easier interpretation because it defines variability along a scale-free range as a percentage from 0 to 100%. Although Higgins et al. claimed that an advantage of the I^2 index is that it "does not inherently depend on the number of studies in the meta-analysis" (p. 559), they provided no evidence to support this claim.

Direct comparisons of I^2 to Q are difficult because only the second index has a known sampling distribution theory that can be used to estimate the probability of a particular value's appearance. To counter this problem with I^2, Higgins and Thompson (2002) developed approximate confidence intervals for I^2 based on the Birge ratio (which they termed the *H* index). Huedo-Medina et al. (2006) used these confidence intervals in order to compare the performance of I^2 to Q in a Monte-Carlo simulation across a wide variety of potential meta-analytic conditions. Huedo-Medina and colleagues' results demonstrated that like Q, I^2 suffers from the same problem of low statistical power with small numbers of studies. Specifically, the confidence intervals around I^2 behave very similarly to tests of Q in terms of Type I error and statistical power. Readers can examine this conclusion for themselves: In each of the 14 examples that Higgins et al. (2003) provided, the inference about consistency reached from the I^2 index is identical to that reached by the Q.

We concur with Higgins and colleagues that (1) in reporting Q (with its associated p value) and I^2 (with its confidence intervals), it is easier to interpret the degree of consistency in a set of study outcomes; (2) using I^2 greatly facilitates comparisons across meta-analyses; and (3) the values of I^2 themselves do not depend on the number of studies. Nonetheless, inferences from both Q and I^2 can be misleading when the number of studies is small. Under such circumstances, analysts should still interpret results with caution.

References

Birge, R. T. (1932). The calculation of errors by the method of least squares. *Reviews of Modern Physics, 40,* 207-227.

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21*, 1539-1558.

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal, 327,* 557–560.

Huedo-Medina, T. B., Sánchez-Meca, F., Marín-Martínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: I^2 or Q statistic? *Psychological Methods, 11*, 193-206.