Running Head: Confidence Intervals

Confidence Intervals for the Random-Effects Variance Component

Michael T. Brannick

University of South Florida

Steven M. Hall

Embry Riddle Aeronautical University

Paper presented in M. Brannick (Chair) *Advances in meta-analysis*. Symposium presented at the 18th annual conference of the Society for Industrial and Organizational Psychology, Orlando, FL (April 2003).

Abstract

We describe the random-effects variance component (REVC) and its uses in meta-analysis. Because the REVC is an estimate, there is uncertainty about its population value. When interpreting the results of a meta-analysis, it is useful to quantify the uncertainty about the REVC by computing a confidence interval. Three methods for computing such confidence intervals are described and illustrated with data from a meta-analysis of the Pygmalion effect in organizations. We recommend that meta-analysts compute and report confidence intervals for the REVC.

Confidence Intervals for the Random-Effects Variance Component

What is the Random-Effects Variance Component?

When we conduct a meta-analysis, we collect an empirical distribution of effect sizes. This distribution has a mean and a variance. Because the individual studies have finite sample sizes, part of the variance of the distribution will be due to sampling variance. That is, even if all underlying effect sizes were the same, we would still see some variance in our collection of effect size estimates because of sampling error. Fortunately, we can estimate the amount of variance due to sampling error and subtract that, leaving a residual variance. The residual is an estimate of how much variance is due to differences in infinite-sample effect sizes. The variance of infinite-sample effect sizes is called the random-effects variance component. Another way of saying this is to imagine that all our studies were estimated with infinite sample sizes, so that we had parameters for each locality. If we computed the variance of effect sizes for those studies, we would have an estimate of the random-effects variance component. The REVC can either be of the 'bare bones' variety, or it can be estimated by taking into account other artifacts such as range restriction and reliability for the collection of studies.

In the literature on validity generalization (e.g., Hunter and Schmidt, 1990), the random-effects variance component is usually represented by

σ_{ρ}^{2}.

The same random-effects variance component in the writings of Hedges and colleagues (e.g., Hedges & Olkin, 1985; Lipsey & Wilson, 2002; Overton, 1998) is usually represented as either:

σ_τ^2 or as τ^2.

I/O folks think in terms of ρ; others think in terms of τ. There are a couple of different ways of estimating the random-effects variance component, but they all essentially subtract sampling variance from observed variance.

Why is the REVC important?

The REVC is important for several reasons. First, the REVC is an estimate of the variability of infinite-sample effect sizes. That is, it shows the impact of context on the size of the association between two constructs. It is an overall estimate of the impact of moderator variables. As the REVC approaches zero, there is no room for moderators to work and a single number can legitimately summarize the lawful relation between two constructs. On the other hand, as the REVC increases, then the infinite sample effect sizes vary across localities. There are moderators to be discovered and a scientific story waiting to be told. When the REVC is large, the effect size can be large and positive in one context but small or even negative in another context.

Second, for I/O psychologists, the REVC is important because it is used along with the estimate of the population mean to compute a lower bound for a credibility interval in validity generalization studies. This lower bound represents a plausible 'worst case' scenario and drives inferences about the transportability of tests. In other words, if the lower bound value is large, we can be confident that the test will be valid in new contexts. If the REVC is large, it is difficult to establish the transportability of tests.

Third, in maximum likelihood estimation procedures for meta-analysis, the REVC is a weight used in computing the overall mean of the effect sizes in the meta-analysis. The random-effects method developed by Hedges incorporates the REVC in its overall

estimate of the mean, as does the method presented in this symposium by Raju and Drasgow. If the REVC is large, the optimal weights approach unit weights. If the REVC is small, the optimal weights approach a function of N_i, the local sample sizes (technically, the inverse of the expected sampling error for each study).

Fourth, the REVC is used as the weight for the prior distribution in empirical Bayes applications. That is, the relative importance of the meta-analysis and the local study effect size will depend on the size of the REVC. If the REVC is large, the local study will carry most of the influence. If the REVC is small, the meta-analysis will carry most of the influence.

For all these reasons, the REVC is important. The REVC has a major impact on the conclusions or inferences that flow from the meta-analysis. For the implications of the meta-analysis to be correct, the REVC needs to be accurate. That is, the REVC must be well estimated for us to have much confidence in certain conclusions drawn from the meta-analysis.

How well is the REVC Estimated?

The short answer to this question is that typically, we don't know. There is reason to believe that often times the REVC is not well estimated. Suppose, just for the sake of argument, that we found a collection of *k* studies that had no sampling error because each was computed based on an infinite *N*, that isfinite *k*, infinite *N*. The distribution would have some mean and variance, and that variance would be a direct estimate of the REVC. There would be no need to subtract sampling variance. The sampling distribution of the variance is known to follow the chi-square distribution, and chi-square can be used to construct a confidence interval for the variance. If we had

infinite-sample effect sizes, we could use chi-square to find confidence intervals. In the situation just described, the size of the confidence interval would depend on *k*, the number of studies and not on *N*, the study sample sizes. If *k* is small, as it often is in meta-analyses, then the variance will not be estimated accurately, or to put it more technically, the variance of the sampling distribution of the estimated REVC will be large and so will the size of the confidence interval around the estimated REVC. The effect of having finite, rather than infinite, sample sizes is to make the estimation of the variance even more difficult. Because of sampling error, we still have *k* studies, but we are not sure about the real values of the effect sizes, because of sampling error.

Therefore, we argue that researchers should describe the precision of their estimates. We should describe the probable error of our results. We should at least calculate confidence intervals for the REVC both to help us as researchers to understand the limits of our knowledge and also to communicate the uncertainty of our findings to others. Next we present and illustrate three ways to estimate confidence intervals for the REVC.

The data that we will use for illustration are taken from the meta-analysis of 17 studies by McNatt (2000) on the Pygmalion effect in organizations. In a typical study, managers are told that some employees are expected to be especially productive. Other employees serve as the control participants. At a later time, the job performance of all employees is measured to see whether the 'exceptional' employees perform better than the controls. The 'exceptional' employees are chosen at random, and so are not really any different than the controls. The effect size is *d*, the standardized mean difference, computed as the exceptional mean minus the control mean divided by the pooled

standard deviation. McNatt used the Hunter and Schmidt (1990) method of analysis and adjusted *d* for sampling error and unreliability in the measure of job performance. He found a mean *d* of 1.13 and a random-effects variance component equal to .60. How much uncertainty is associated with this value? To answer this question, we need to put confidence intervals around the .60 value. If we had infinite-sample studies, we could put confidence intervals about the REVC using the Chi-square distribution:

$$p\left[\frac{(k-1)\hat{\tau}^2}{\chi^2_{(k-1;.025)}} \le \tau^2 \le \frac{(k-1)\hat{\tau}^2}{\chi^2_{(k-1;.975)}}\right] = .95 .$$

In our example, the values would be

$$p\left[\frac{16(.60)}{28.845} \le \tau^2 \le \frac{16(.60)}{6.908}\right] = .95$$; the resulting interval is .33 to 1.39.

This interval is certainly too small as it fails to consider sampling error at the study level.

Method 1: Approximate Distribution Method

Biggerstaff and Tweedie (1997) developed three methods of computing confidence intervals for the random-effects variance components. Two of the methods are asymptotic and appropriate for meta-analyses based on large numbers of studies; those methods are not described here. The third method is appropriate for small *k*, and so is appropriate for most situations encountered by I/O psychologists. The method is based on an estimator of the random-effects variance component developed by DerSimonian and Laird (1986). Like the Hunter and Schmidt method, the DerSimonian and Laird method begins by finding a weighted mean and sum of squared deviations from the mean. The weights are the inverse of the sampling variance for each study rather than the sample size.

The estimated sampling variance of *d* is

$$Var(d_i) = \frac{1}{w_i} = \frac{n_{E_i} + n_{C_i}}{n_{E_i} n_{C_i}} + \frac{d_i^2}{2(n_{E_i} + n_{C_i} - 2)}$$

We find a weighted mean by:

$$\hat{\mu} = \frac{\sum w_i d_i}{\sum w_i}$$

and a weighted sum of squares by

$$Q = \sum w_i (d_i - \hat{\mu})^2 \ .$$

Then the DerSimonian and Laird estimator of the REVC is

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

It may not look like it, but what is happening is that the sum of squares, Q, is basically chi-square, and its expected value is basically (*k*-1), so we are subtracting expected sampling variance from observed variance. The weights (w_i) affect the observed variance, so the denominator adjusts the result for the influence of the weights. If (*k*-1) is greater than Q, the result will be negative and set to zero. What Biggerstaff and Tweedie did was to figure the approximate distribution of $\hat{\tau}^2_{DL}$. They then found upper and lower bounds of the distribution, plus provided some handy code in SAS for calculating the bounds. Most psychologists will not find the equations meaningful, but here they are (the interested reader is referred to Biggerstaff and Tweedie, 1997, for further detail).

$$L(\tau^2) = \int_{\lambda(\tau^2)[c\hat{\tau}^2_m + k - 1]}^{\infty} \frac{1}{\Gamma(r(\tau^2))} u^{r(t^2)-1} e^{-u} \mathrm{d}u$$

$$U(\tau^2) = \int_0^{\lambda(\tau^2)[c\hat{\tau}_m^2+k-1]} \frac{1}{\Gamma(r(\tau^2))} u^{r(t^2)-1} e^{-u} \mathrm{d}u$$

A SAS program with all the analyses reported here can be found on Brannick's website under 'Software' (so you don't have to understand the math to calculate the confidence intervals using their method). For the McNatt data, using the DerSimonian and Laird method and no corrections for unreliability (i.e., the bare bones method), the estimates of the mean and REVC are 1.09 and .46, respectively. If the 17 studies had no sampling error and provided a variance estimate of .46, we could use chi-square to calculate a confidence interval. The interval would be .26 to 1.08 in this case. The Biggerstaff and Tweedie estimated confidence interval is .18 to 2.59, clearly a large interval, and larger than the chi-square estimate, as it should be. As the study sample sizes increase, the Biggerstaff confidence intervals approach the chi-square values.

Because it is based on statistical theory, this method has much to recommend it. However, it cannot be applied to the Hunter and Schmidt estimates because of the difference in weights. It also does not incorporate multiple corrections (reliability, range restriction). It is based on the assumption that both the distribution of sampling error and the distribution of delta (infinite-sample effect sizes) are normal.

Method 2: Bootstrap Estimates

Bootstrap estimates of confidence intervals in meta-analysis were introduced by Switzer, Paese and Drasgow (1992). What you do with the bootstrap is to take repeated samples with replacement from your dataset. You compute estimates from each sample to create a sampling distribution. Then you look at the tails of the sampling distribution to estimate confidence intervals. We wrote a SAS program to do this for the McNatt data and generated 5000 trials. The confidence interval ranges from .16 to 1.29 (see Table 1).

These numbers are smaller than the estimates provided by Biggerstaff and Tweedie, particularly on the upper tail. However, these estimates still show a rather large interval.

An advantage of the bootstrap method is that it makes no assumption about the distribution of the data other than that the data in the meta-analysis are representative of the population of studies to which we wish to generalize. On the other hand, the bootstrap estimates can be erroneous if the small sample of data is not representative because of outliers, bias in the selection of studies, and so forth.

Empirical Method with Assumed Distributions

A third method is similar to the bootstrap in that it also makes use of repeated sampling. The difference is that in this method, assumptions are made about the underlying distribution of effect sizes and reliability coefficients as is typically done in Monte Carlo studies, e.g., Hall and Brannick (2002). Instead of repeatedly sampling from the raw data, the raw data are used to set parameters of the underlying distribution of effect sizes and reliability coefficients. The underlying distributions are used to simulate studies, which become the basis for each meta-analysis. Like the bootstrap, thousands of iterations are used to compute an empirical distribution of REVCs. The tails of the distribution are then used to construct the confidence intervals, just as in the bootstrap method.

In the McNatt data, the estimate of delta is 1.13 and the REVC is .60. We assumed that the distribution of effect sizes was normal, with a mean of 1.13 and a variance of .60. We sampled from this distribution to find local population values. The McNatt reliability values ranged from .65 to .95. We assumed that reliability is uniformly distributed in this interval, and sampled from this distribution to simulate a

local reliability. We attenuated the local population effect sizes for unreliability, and then took samples of sizes that McNatt found from these local values. We then computed meta-analyses on the simulated data and found an estimated REVC. After doing so 5000 times, we had an empirical distribution of REVC that we used to find confidence intervals. The resulting interval ranges from .13 to 1.28 (see Table 1).

The empirical method should work well so long as the assumed distributions are accurate. If the data are representative of the population of studies, the bootstrap method might be preferred; if they are not, then the empirical method may provide better estimates. Of course we do not know the underlying distributions of effect sizes and reliabilities, so results of the empirical method may be viewed with some suspicion. On the other hand, the empirical method can be used with unlimited numbers of distributions, so that the researcher can see what effect changes in the underlying distributions has on the confidence interval. If even pathological distributions fail to affect the confidence intervals much, then one can bolster confidence in the estimates.

Overall Results

All of the methods showed several common points:

1. The estimated REVC for the Pygmalion studies is large; in all cases, the bottom of the confidence intervals are well above zero. Therefore, the overall mean Pygmalion effect provided by the meta-analysis is only an approximate value and may not apply to the local context of interest. That is, the Pygmalion effect at work appears to be quite variable. Further research is needed to describe and understand the sources of variance in outcomes across studies.

2. The confidence intervals are all large, indicating that the data are consistent with a wide range of population values. In other words, there is a good deal of uncertainty remaining about the magnitude of the REVC in studies of the Pygmalion effect at work.

There were also differences among the methods. The Biggerstaff and Tweedie method produced estimates that were larger on both tails than the estimates provided by the other two methods. Estimates from the two Monte Carlo methods were quite similar. On the upper tail, both Monte Carlo methods produced estimates that were more narrow than the chi-square method. The chi-square method should produce the narrowest confidence intervals about REVC. Thus it appears that the Monte Carlo methods may be overly optimistic with regard to the size of interval. Further research is warranted on the behavior of each of the techniques.

Conclusions

Random-effects meta-analyses produce estimates of the Random Effects Variance Component. The REVC is important for several reasons, including its role in the search for moderators, the computation of the credibility value, the computation of the overall mean effect size, and the computation of empirical Bayes estimates. There is uncertainty about the actual value of the REVC; a meta-analysis only produces an estimate of the REVC. We showed 3 methods of constructing confidence intervals for the REVC, and we recommend that researchers routinely compute and report the confidence intervals of their choice in future meta-analyses.

References

Biggerstaff, B. J., & Tweedie, R. L. (1997). Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. *Statistics in Medicine, 16*, 753-768.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trails. *Controlled Clinical Trials, 7*, 177-188.

Hall, S. M., & Brannick, M. T. (2002). Comparison of two random-effects methods of meta-analysis. *Journal of Applied Psychology, 87*, 377-389.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Hedges, L. V. & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3*, 486-504.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

McNatt, D. B. (2000). Ancient Pygmalion joins contemporary management: A meta-analysis of the result. *Journal of Applied Psychology, 85*, 314-322.

Overton, R. C. (1998). A comparison of fixed-effects and mixed (random-effects) models for meta-analysis tests of moderator variable effects. *Psychological Methods, 3*, 354-379.

Switzer, F. S., Paese, P. W., & Drasgow, F. (1992). Bootstrap estimates of standard errors in validity generalization. *Journal of Applied Psychology*, *77*, 123-129.

Table 1

Summary of Results

Method	REVC	Confidence Interval
Chi-square (H&S)	.60	.33 to 1.39 (hypothetical)
Chi-square (D&L)	.46	.26 to 1.08 (hypothetical)
Biggerstaff & Tweedie	.46	.18 to 2.59
Bootstrap	.60	.16 to 1.29
Empirical Assumed	.60	.13 to 1.29