moderators does little in the way of increasing our confidence in the notion that meta-analyses typically isolate those studies producing values sampled from a single population" (2003, p. 426). Thus, we have no way of knowing from meta-analyses whether validity generalization exists or not because we lack the between-study variance in moderators across which validities might (or might not) generalize. Meta-analytic oversampling of studies with extreme values of moderator variables (e.g., primary studies using all-female samples and studies using all-male samples) does not seem feasible. What would help would be for authors of primary studies to include in their correlation matrices products of predictors and common moderators, *even if those interactions are not hypothesized*. This would allow the meta-analysis of relationships between residualized products and outcomes, of regression weights of products, and so on that reflect within-study (rather than between-study) variance in moderators.

References

Cortina, J. M. (2003). Apples and oranges (and pears, oh my!): The search for moderators in meta-analysis. *Organizational Research Methods, 6*(4), 415–439.

Markell, H. M., Lei, X., & Foroughi, C. K. (April 2017). Restricted between-study variance in meta-analytic moderators. Symposium presented at the annual meeting for the Society for Industrial and Organizational Psychology, Orlando, FL.

Mosier, C. I. (1950). Review of difficulty prediction of test items. *Journal of Applied Psychology*, *34*(6), 452–453.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology*, *62*(5), 529–540.

Schmidt, F. L., & Hunter, J. E. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.

Tett, R. P., Hundley, N., & Christiansen, N.D. (2017). Meta-analysis and the myth of generalizability. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 10*(3), 421–456.

Wagner, J. A. (1995). Studies of individualism-collectivism: Effects on cooperation in groups. *Academy of Management Journal*, *38*(1), 152–173.

Don't Get Too Confident: Uncertainty in SD_ρ

Scott B. Morris, Samuel T. McAbee, Ronald S. Landis, and Kristina N. Bauer
Illinois Institute of Technology

Scott B. Morris, Department of Psychology, Illinois Institute of Technology; Samuel T. McAbee, Department of Psychology, Illinois Institute of Technology; Ronald S. Landis, Department of Psychology, Illinois Institute of Technology; Kristina N. Bauer, Department of Psychology, Illinois Institute of Technology.

Correspondence concerning this article should be addressed to Scott B. Morris, Department of Psychology, Illinois Institute of Technology, 3105 South Dearborn, Chicago, IL 60616. E-mail: scott.morris@iit.edu

https://doi.org/10.1017/iop.2017.43 Published online by Cambridge University Press

Tett, Hundley, and Christiansen (2017) raise an important issue related to meta-analysis and our frequent overinterpretation of point estimates to the diminishment of variability of the estimate. We view this as analogous to the situation in which weather forecasters communicate the likely track of hurricanes. Such predictions involve point estimates of where the center of the storm is likely to be at some future time. These point estimates can be connected to identify the most likely path of the storm. In addition to these point estimates, however, forecasters caution that we should also attend to the "cone of uncertainty." That is, we should not focus exclusively on the point estimate to the exclusion of the errors of prediction.

In this commentary, we extend Tett et al.'s (2017) argument by identifying an additional source of variability in meta-analysis: uncertainty with respect to SD_ρ. Greater attention to the between-study variance of effect sizes brings with it a need to consider the precision of our estimates of SD_ρ. Although it is common, if not obligatory, to consider sampling error in estimates of the mean effect size, researchers frequently overlook the degree of uncertainty in our statistics when evaluating between-study heterogeneity. We build confidence intervals and credibility intervals around the mean correlation, but we give little thought to our uncertainty regarding the endpoints of the intervals (Evangelou, Ioannidis, & Patsopoulos, 2007).

Precision of SD_ρ Estimates

Two approaches exist for evaluating sampling error in between-study heterogeneity statistics. First, we can test the between-study variance for statistical significance, using Hedges *Q*-test (Hedges & Olkin, 1985). As a significance test, this approach indicates that SD_ρ is greater than zero (if significant) or that we cannot rule out zero variance as a possibility (if nonsignificant). Unfortunately, neither of these conclusions is entirely satisfactory. If significant, we still do not know how much faith to place in our estimate of SD_ρ. If nonsignificant, the result is inherently ambiguous in situations with small *K*, because the test tends to have lower power unless the number of studies is large (Hedges & Vevea, 1998).

The second approach is to build a confidence interval around the variance estimate. Confidence intervals are generally favored over significance tests because (a) statistical significance can be inferred when the confidence interval does not include zero, and (b) confidence intervals appropriately present heterogeneity as ranging on a continuum rather than simply being present or absent (i.e., *dichotomous thinking*; Cumming & Fidler, 2009). Several methods for constructing confidence intervals on between-study variance have been developed (Borenstein, Hedges, Higgins, & Rothstein, 2009; Brannick & Hall, 2003; Viechtbauer, 2005, 2007). These methods build a confidence interval around the between-study variance rather than SD_ρ;

https://doi.org/10.1017/iop.2017.43 Published online by Cambridge University Press

however, the square root of the resulting endpoints can be used to represent an interval around SD_ρ (Borenstein et al., 2009). Confidence intervals can also be constructed around the I^2 statistic (Higgins & Thompson, 2002), which quantifies the degree of heterogeneity as the proportion of between-study variance to total variance. Pragmatically, the R package *metafor* (Viechtbauer, 2010) provides confidence intervals around VAR_ρ, SD_ρ, and I^2.

In many situations, examining confidence intervals around SD_ρ is likely to reveal substantial uncertainty in the degree of between-study variance. Credibility intervals can be quite large, even when the true SD_ρ equals zero (Oswald & Johnson, 1998). This is particularly true when examining moderator subgroups, which often rely on a fairly small number of studies.

To illustrate, Morris, Daisley, Wheeler, and Boyer (2015) reported a random-effects meta-analysis on the validity of individual psychological assessments, with an average uncorrected validity of .27 predicting job performance ratings. SD_ρ was estimated at .12, indicating substantial variability across studies (i.e., an 80% credibility interval of .12 to .42). Using *metafor*, the current authors estimated a 95% confidence interval for SD_ρ ranging from .078 to .192. Further, the confidence interval on the I^2 statistic indicated that between 41% and 81% of the observed variance was due to true between-study differences.

The impact of the uncertainty in the SD_ρ estimate can be seen by constructing alternate credibility intervals using the endpoints of the SD_ρ confidence interval. At the lower end ($SD_\rho = .078$), the 80% credibility interval would be $.27 \pm 1.28 * .078$, ranging from .17 to .37, and suggests relatively minor differences in validity between studies. Using the upper bound estimate, however, resulted in an 80% credibility interval ($.27 \pm 1.28 * .192$) that ranged from .02 to .52, implying strikingly different levels of validity across different settings.

So, what can we conclude from these results? Validity is not homogenous, and there is some degree of situation specificity present in these data. However, the extent of this variability is quite uncertain. We simply do not have enough data with 37 samples to accurately estimate the likely range of validities.

Effects of *K* on the Precision of SD_ρ

In their focal article, Tett et al. (2017) draw attention to the paradoxical association between increased situational specificity and the corresponding decrease in certainty around estimates of ρ and VAR_ρ (i.e., SD_ρ^2) as a function of *K*. It is similarly important to recognize how *K* influences the precision of confidence intervals around our estimates of heterogeneity.

https://doi.org/10.1017/iop.2017.43 Published online by Cambridge University Press

As noted, there are several methods for constructing confidence intervals around SD_ρ (see Viechtbauer, 2007). For ease of demonstration, we focus on the Wald-type confidence interval, for which the role of K is most transparent. First, we estimate the sampling variance of SD_ρ^2,

$$VAR\left[SD_\rho^2\right] = 2\left(\sum w_{j^2}\right)^{-1} = 2\left(\sum \frac{1}{\left(SD_\rho^2 + \sigma_j^2\right)^2}\right)^{-1} \quad \textbf{(1)}$$

where SD_ρ^2 is the between-studies variance estimate, and σ_j^2 is the within-study variance estimate. Once this estimate is obtained, a confidence interval for SD_ρ^2 can be constructed as:

$$SD_\rho^2 \pm 1.96\sqrt{VAR\left[SD_\rho^2\right]} \quad \textbf{(2)}$$

A 95% confidence interval for SD_ρ is then obtained by calculating the square root of the upper and lower confidence limits for SD_ρ^2. Typically, lower confidence limits that fall below zero are truncated to zero (Viechtbauer, 2007).

To see the effect of K, assume that the within-study variance is constant such that $\sigma_j^2 = \sigma_e^2$. Once we make this substitution, Equation (1) simplifies to:

$$VAR\left[SD_\rho^2\right] = 2\left(\frac{K}{\left(SD_\rho^2 + \sigma_e^2\right)^2}\right)^{-1} = \frac{2\left(SD_\rho^2 + \sigma_e^2\right)^2}{K} \quad \textbf{(3)}$$

Thus, for given values of SD_ρ and σ_e^2, as K decreases, the width of the confidence interval around SD_ρ increases.

Table 1 presents a brief demonstration of the effects of K on the confidence interval for SD_ρ for two levels of heterogeneity. On the left, we present the results for the estimates of SD_ρ and the average estimate of σ_j^2 (referred to below as σ_e^2) obtained from Morris et al. (2015). The degree of heterogeneity in the Morris et al. study is fairly large, and as Tett et al. note, the magnitude of SD_ρ is likely to decrease as the analysis focuses on more specific situations (i.e., moderator subgroups). Therefore, on the right side of Table 1, we present the findings for smaller estimates of SD_ρ and σ_e^2, representing likely values for more narrowly focused situations (Tett et al., 2017). As shown in Table 1, for a given level of SD_ρ and σ_e^2, the width of the confidence interval decreases as K increases. Notice, however, that this pattern is stronger for the larger estimates of SD_ρ and σ_e^2. As such, the influence of K on the confidence interval for SD_ρ is stronger where there is greater heterogeneity around the effects.

https://doi.org/10.1017/iop.2017.43 Published online by Cambridge University Press

Table 1. Effect of *K* on 95% Confidence Intervals for SD_ρ

	[a]$SD_\rho = .118,\ \sigma_e^2 = .008$					[b]$SD_\rho = .050,\ \sigma_e^2 = .002$				
		SD_ρ^2		SD_ρ			SD_ρ^2		SD_ρ	
K	$\sqrt{VAR[SD_\rho^2]}$	LCL	UCL	LCL	UCL	$\sqrt{VAR[SD_\rho^2]}$	LCL	UCL	LCL	UCL
4	.016	−.016	.044	.000	.210	.003	−.004	.009	.000	.093
5	.014	−.013	.041	.000	.203	.003	−.003	.008	.000	.090
6	.013	−.011	.039	.000	.197	.003	−.003	.008	.000	.087
7	.012	−.009	.037	.000	.192	.002	−.002	.007	.000	.085
8	.011	−.008	.035	.000	.188	.002	−.002	.007	.000	.083
9	.010	−.006	.034	.000	.185	.002	−.002	.007	.000	.082
10	.010	−.005	.033	.000	.182	.002	−.001	.006	.000	.080
15	.008	−.002	.030	.000	.172	.002	−.001	.006	.000	.076
20	.007	.000	.028	.018	.166	.001	.000	.005	.000	.073
25	.006	.002	.026	.042	.161	.001	.000	.005	.002	.071
30	.006	.003	.025	.053	.158	.001	.000	.005	.015	.069
35	.005	.004	.024	.060	.156	.001	.000	.005	.020	.068
40	.005	.004	.024	.066	.153	.001	.001	.004	.023	.067
50	.004	.005	.023	.073	.150	.001	.001	.004	.027	.065

Note. LCL = lower 95% confidence limit, UCL = upper 95% confidence limit. [a]Values from Morris et al. (2015). [b]Likely values for more narrow situations, as represented in Tett et al. (2017).

Recommendations

Our primary motivation in offering this commentary is to reify the point made by Tett et al. (2017) regarding the need for greater attention to the variability of our estimates. In the context of meta-analysis, we recommend that researchers report confidence intervals around SD_ρ and incorporate this into their discussion of results. More broadly, we encourage researchers to comprehensively and accurately report and interpret the degree of (un)certainty in results across all analyses.

References

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. (2009). *Introduction to meta-analysis*. Chichester, UK: Wiley.

Brannick, M. T., & Hall, S. M. (2003). Confidence intervals for the random-effects variance component. In M. T. Brannick (Chair), *Advances in meta-analysis*. Symposium conducted at the 18th Annual Conference of the Society for Industrial and Organizational Psychology, Orlando, FL.

Cumming, G., & Fidler, F. (2009). Confidence intervals: Better answers to better questions. *Zeitschrift für Psychologie/Journal of Psychology, 217*, 15–26.

Evangelou, E., Ioannidis, J. P., & Patsopoulos, N. A. (2007). Uncertainty in heterogeneity estimates in meta-analyses. *BMJ: British Medical Journal*, *335*(7626), 914–916.

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.

https://doi.org/10.1017/iop.2017.43 Published online by Cambridge University Press

Higgins, J., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*, 1539–1558.
Morris, S. B., Daisley, R. L., Wheeler, M., & Boyer, P. (2015). A meta-analysis of the relationship between individual assessments and job performance. *Journal of Applied Psychology, 100*, 5–20.
Oswald, F. L., & Johnson, J. W. (1998). On the robustness, bias, and stability of statistics from meta-analysis of correlation coefficients: Some initial Monte Carlo findings. *Journal of Applied Psychology*, *83*, 164–178.
Tett, R. P., Hundley, N. A., & Christiansen, N. D. (2017). Meta-analysis and the myth of generalizability. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 10*(3), 421–456.
Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics, 30*, 261–293.
Viechtbauer, W. (2007). Confidence intervals for the amount of heterogeneity in meta-analysis. *Statistics in Medicine*, *26*, 37–52.
Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*, 1–48.

Empirical Benchmarks for Interpreting Effect Size Variability in Meta-Analysis

Brenton M. Wiernik
Ghent University

Jack W. Kostal and Michael P. Wilmot
University of Minnesota

Stephan Dilchert
Baruch College, CUNY

Deniz S. Ones
University of Minnesota

Generalization in meta-analyses is not a dichotomous decision (typically encountered in papers using the *Q* test for homogeneity, the 75% rule, or null hypothesis tests). Inattention to effect size variability in meta-analyses may stem from a lack of guidelines for interpreting credibility intervals. In this commentary, we describe two methods for making practical interpretations and determining whether a particular SD_ρ represents a meaningful level of variability.

Brenton M. Wiernik, Department of Developmental, Personality and Social Psychology, Ghent University; Jack W. Kostal, Department of Psychology, University of Minnesota; Michael P. Wilmot, Department of Psychology, University of Minnesota; Stephan Dilchert, Narendra Paul Loomba Department of Management, Baruch College, CUNY; Deniz S. Ones, Department of Psychology, University of Minnesota.

Correspondence concerning this article should be addressed to Brenton M. Wiernik, Department of Developmental, Personality and Social Psychology, Ghent University, Henri Dunantlaan 2, 9000 Gent, Belgium. Email: wiernik@workpsy.ch

https://doi.org/10.1017/iop.2017.43 Published online by Cambridge University Press