Beyond Heterogeneity of Effect Sizes

One of the primary benefits of meta-analytic syntheses of research findings is that researchers are provided with an estimate of the heterogeneity of effect sizes. In many instances, estimates of effect size heterogeneity obtained from a meta-analytic review are much smaller than those suggested by a narrative review of the literature because meta-analysis allows researchers to estimate how much of the observed variability in effect sizes across studies is simply an artifact of sampling error. Indeed, the realization that some variability in effect sizes is to be expected purely as a function of sampling error and other artifacts led researchers to shift away from developing situational explanations of disparate research findings, a shift described by Steel and Kammeyer-Mueller as "…among the most significant changes in management research over the past century because it allowed knowledge to accumulate much more rapidly across studies" (2008: p. 55). Modern meta-analytic methods also allow researchers to assess how much of the remaining variability in effect sizes can be attributed to other study artifacts such as unreliability in measurement and range restriction. That is, meta-analyses provide an estimate of the variability in effect sizes that remains after accounting for the variability that is to be expected as a result of sampling error and other study artifacts.

Low values for this estimate are typically interpreted as indicating that the strength of an effect generalizes across situations (Murphy, 2003) – a conclusion that has important implications for researchers and practitioners. DeShon, writing about personnel selection, states that generalizability avoids "the highly undesirable process of conducting a new validation study every time a selection test was used in a new context" (2002: 365). Despite the universal appreciation of generalizability, there is disagreement about how much residual heterogeneity is typically found in meta-analyses and consequently how much generalizability is enabled.

Some have argued that many of the relationships studied in I-O psychology are characterized by little or no residual heterogeneity. Indeed, in a series of articles Schmidt and colleagues (e.g., DeGeest & Schmidt, 2010; Schmidt, 2008; Schmidt & Hunter, 1977; Schmidt, Hunter, & Raju, 1988; Schmidt & Oh, 2010; Schmidt, Ocasio, Hillery, & Hunter, 1985) characterize substantive amounts of residual variance as "the situation specificity hypothesis" and therefore appear to argue that researchers should assume zero heterogeneity (as the effective null hypothesis) unless proven otherwise. For example, Schmidt et al., in discussing meta-analytic reviews of validity coefficients for different jobs write "These studies have shown that most, and in many cases all, of the between-study variance in validity coefficients for similar jobs and a given ability is due to artifacts such as sampling error" (1985, p. 511). And more recently, Schmidt writes "The evidence suggests that moderators are often solipsistic: they exist in the minds of researchers but not in real populations" (2008, p. 111).

On the other hand, as Tett, Hundley, and Christiansen's article emphasizes, there are sound reasons for suspecting that many relationships studied in I-O psychology are, in fact, characterized by non-trivial amounts of residual effect size heterogeneity. First, Baron and Kenny's (1986) classic piece on moderators and mediators in psychological research is approaching 70,000 citations according to Google Scholar, indicating moderators are indeed of common concern if not occurrence. Second, some researchers (e.g., Steel & Kammeyer-Mueller, 2008) have expressed alarm regarding the manner in which estimates of effect size heterogeneity are calculated in commonly used meta-analytic methods such as the Hunter and Schmidt approach. For example, Steel et al. (2010) argued that the near homogeneity professed for job performance validity coefficients is largely the product of methodological and statistical errors, with "plenty of variability" (p. 373) remaining after taking into account sampling error. Third, some researchers formalize the assumption of effect size homogeneity by using fixed-effects meta-analysis such as the Rosenthal-Rubin model or Hedges and Olkin fixed-effects model (Field, 2005; Overton, 1998), although it should be noted that Schmidt and colleagues (e.g., Hunter & Schmidt, 2004) specifically caution against the use of fixed-effects models. Fourth, some meta-analytic practices such as the elimination of outliers can also result in artificially low estimates of effect size heterogeneity (Aguinis, Gottfredson, & Joo, 2013). Other practices, such as the use of Q statistics to test for homogeneity are often similarly misapplied. To repeat what should already be second nature to us, we cannot prove the null or, in other words, the absence of evidence is not the evidence of absence.

Empirically, there have been several quantitative summaries of multiple meta-analyses establishing the average SDρ in a field, which echo Tett et al.'s contribution here, partially summarized by Steel, Kammeyer-Mueller, and Paterson (2015). Focusing on those in the social sciences, these include Carlson and Ji (2011), who found a SDρ of .121, Steel et al. (2013), who found a SDρ of .15, and Paterson et al. (2016), notably based on 258 management studies, who found an average SDρ of 0.14. As can be seen in Figure 1 (from Steel et al., 2015) large amounts of heterogeneity are common and tend to be normally distributed, with virtually no meta-analyses reporting homoegeneity. However, there is a caveat. This does not apply to meta-analyses based on small numbers of study.

As per Figure 2 (from Steel et al., 2014), almost one quarter of all small meta-analyses report zero variance. A number of authors argue that this can be attributed to calculations used for estimating the amount of effect size heterogeneity, which are particularly biased when the number of studies on which the meta-analysis is based is low (Brannick & Hall, 2001; Cornwell & Ladd, 1993; Overton, 1998). This was first highlighted by Spector and Levine (1987) during their Monte Carlo simulations to address meta-analytic variance, where they made a small but substantive change to their formula for variance. Instead of putting *n* in the denominator, as used by Schmidt and Hunter (2003) below, they used *n* – 1. The reason for this adjustment is that using the observed mean systematically underestimates observed variance. Essentially, this is the same issue that arises when standard deviations are calculated at the individual study level. Within studies, *n* – 1 is used in the denominator of the variance equation, unless the true population mean is available instead of the sample mean. Accordingly, Brannick and Hall (2001) recommend multiplying the observed variance by *k*/(*k* – 1) to remove the bias, where *k* represents the number of studies used in the meta-analysis. This is an example of *Bessel's Correction*. The larger the number of studies in a meta-analysis, the less effect this correction will have. Meta-analyses of few studies are significantly impacted; a meta-analysis based on five studies underestimates effect size variance by 25%.

This will impact Tett et al.'s findings considerably. It is worse than they make out. As they note, as meta-analyses become more specific, the observed heterogeneity shrinks, but this is also due to fundamentally biased calculations. You can observe this again in Table 1 (from Paterson et al., 2016), showing a positive correlation between K, the number of studies, and reported SDρ as well as why the common moderator technique of hierarchical subgroup exacerbates matters (Steel & Kammeyer-Mueller, 2002). We recommend they multiply their variance estimates by K/(K-1) to gain a clearer picture of present circumstances.

In addition, we might take issue with the broad use of 80% CI. Though CI are somewhat arbitrary, this choice puts it out of step with the rest of the field. While 95% confidence intervals are the common threshold to indicate statistical significance, we suspect the common use of the 80% CI used in meta-analysis purpose is to exaggerate perceptions of generalizability. As reviewed by Steel and Kammeyer-Mueller (2009), “Because both of these estimates provide information regarding the likelihood that an estimated relationship will hold in future samples of applicants, it is sensible to compare local confidence intervals to meta-analytic credibility intervals (p. 535). We recommend that Tett et al. use 95% CI to portray a more “generalizable” understanding of generalizability.

Finally, we note that generalizability is context dependent, which is not captured by any statistical index. As reviewed by Steel et al. (2015), you can comfortably generalize when extending the findings to the range of samples, settings and measures comprised in the meta-analysis itself. However, as one tries to extend these findings further, to dissimilar measures or participants, confidence in generalizability should wane. The problem is that when we pass from the similar to the dissimilar is not always clear (Matt & Cook, 2009). Meta-analysis perpetually struggles with the issue of external validity, often termed the commensurability or the “apples and oranges” issue (Cortina, 2003; Steel, Schmidt & Shultz, 2008). Murphy (2009) notes that studies in most meta-analyses do not represent independent random samples from a well-defined population. Given that most studies in management science don’t have a clear description of what the population of interest is, it becomes rather difficult to know if a new instance is represented. Since single studies must deal with external validity for conditions outside its scope, a meta-analytic aggregation of several studies from this exact same population or method does not make this concern disappear. To stress that this is a fundamental that cannot be readily resolved by any statistical technique, it considered an example of Hume’s “problem of induction,” one of the most pernicious problems in all of science.

Despite all this, we believe some generalizability is possible, particularly directional generalizability, that is a relationship will stay either negative or positive under a majority of circumstance. Consider again the distribution of SDρ from Figure 1 (from Steel et al., 2015). Its 95th percentile of SDρ is equal to .282. Multiplying this by 1.645, gives the one tailed credibility interval of .464. Consequently, any effect size greater than .464 is expected to generalize at least 95% of the time. That is, the heterogeneity of 95% of relationships is sufficiently small that effects will generalize if they are greater than .46. Consequently, we can argue for generalizability even in the absence of a meta-analysis, though meta-analyses are useful for further narrowing the range of possible values.

To more adeptly and precisely argue for generalizability, it would help if we accepted that moderators and heterogeneity are the rule, not the exception. It is important to reflect what an incredibly bold statement it is to profess perfect generalizability. It is essentially putting a social science relationship on par with the speed of light, as both would be considered constants of the universe. And yet we do this often and with a straight face. For example, in our field the position that validity coefficients do not have substantive moderators is a defendable position by many (e.g., Schmidt & Oh, 2010). This false assumption of homogeneity can seriously stifle scientific progress by preventing researchers from exploring possible moderators of relationships (Sackett, Harris, & Orr, 1986; Steel & Kammeyer-Mueller, 2008). Aguinis, Sturman, and Pierce warn "The failure to detect population moderating effects is detrimental to the advancement of management …If the meta-analysis incorrectly reported the null result, this could have serious negative effects, particularly because meta-analytic reviews are often more influential than primary-level studies" (2008, p. 28).

Once we accept the inherent complexity of the world, that findings are not hermetically sealed from outside influences, our science can progress. We encourage researchers to be more thorough in their exploration of potential moderating variables and to properly contextualize their studies. Too often, when looking for potential moderators, we are limited to common demographic variables, such as sex or age, as this is the only characteristics that are recording. If we could describe our studies more completely, what strides we could make in understanding variation. At a minimum, we should be commonly describing job's with ONET codes and industries with SEC codes. There are several sources of information to aid researchers in exploring potential moderators to consider in or contextualize their research (e.g., Cohen, Cohen, West & Aiken, 2003; Cortina, 2003; Gardner, Li & Harris, 2013) and evidence presented here along with numerous meta-analyses that have found sizeable moderation effects (e.g., Dalal, 2005; Harms & Credé, 2010) suggest that this is worth exploring.

The benefits of taking moderators seriously are worth the effort. Aside from increasing our understanding of the world, we improve our predictions. We will understand when an effect holds and when it fails, allowing application with confidence. So far, the only topic that has seriously grappled with potential moderators in synthetic validity, that is being able to predict entire selection system from a job analysis alone (Steel et al., 2010). But imagine if our every topic could mature to the point of confidently going from diagnosis to customized treatment that maximizes effectiveness. IO psychology's star would surely rise.

References

Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods, 16*(2), 270-301.

Aguinis, H., Sturman, M. C., & Pierce, C. A. (2008). Comparison of three meta-analytic procedures for estimating moderating effects of categorical variables. *Organizational Research Methods*, *11*(1), 9-34.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*(6), 1173-1182.

Brannick, M. T. & Hall, S. M. (2001). *Reducing bias in the Schmidt-Hunter meta-analysis.* Poster session presented at the annual conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

Carlson, K. D., & Ji, F. X. (2011). Citing and building on meta-analytic findings. *Organizational Research Methods, 14*(4), 696-717.

Cohen, J., Cohen, P., West, S. G., Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.) Mahwah, NJ: Erlbaum.

Cornwell, J. M., & Ladd, R. T. (1993). Power and accuracy of the Schmidt and Hunter meta-analytic procedures. *Educational and Psychological Measurement, 53*(4), 877-895.

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.

Dalal, R. S. (2005). A meta-analysis of the relationship between organizational citizenship behavior and counterproductive work behavior. *Journal of Applied Psychology, 90*(6), 1241-1255.

DeGeest, D. S., & Schmidt, F. L. (2010). The impact of research synthesis methods on industrial–organizational psychology: The road from pessimism to optimism about cumulative knowledge. *Research Synthesis Methods, 1*(3-4), 185-197.

DeShon, R. (2002). A generalizability theory perspective on measurement error corrections in validity generalization. In K.R. Murphy (Ed.)*, Validity generalization: A critical review* (pp. 365-402). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Field, A. P. (2005). Is the meta-analysis of correlation coefficients accurate when population correlations vary? *Psychological Methods, 10*(4), 444-467.

Gardner, R. G., Li, N., & Harris, B. (2013). *Moderation in all things: Interaction effects in management research.* Unpublished paper presented at the 28th Annual Conference of the Society for Industrial and Organizational Psychology, Houston, TX.

Harms, P. D., & Credé, M. (2010). Emotional intelligence and transformational and transactional leadership: A meta-analysis. *Journal of Leadership & Organizational Studies*, 17(1), 5-17.

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

Matt, G. E., & Cook, T. D. (2009). Threats to the validity of generalized inferences. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis*: 537-560. New York: Russell Sage Foundation.

Murphy, K. R. (2003). The logic of validity generalization. In K.R. Murphy (Ed.), *Validity generalization: A critical review* (1-29). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Murphy, K. R. (2009). Validity, validation and values. *Academy of Management Annals, 3*, 421-461.

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*(3), 354-379.

Paterson, T.A., Harms, P.D., Steel, P., & Credé, M. (2016). An Assessment of the Magnitude of Effect Sizes: Evidence from 30 Years of Meta-Analysis in Management. *Journal of Leadership & Organizational Studies*, *23*(1), 66-81.

Sackett, P. R., Harris, M. M., & Orr, J. M. (1986). On seeking moderator variables in the meta-analysis of correlational data: A Monte Carlo investigation of statistical power and resistance to Type I error. *Journal of Applied Psychology, 71*, 302-310.

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. (2003). History, development, evolution, and impact of validity generalization and meta-analysis methods, 1975−2001. In K. R. Murphy (Ed.), Validity generalization: A critical review (pp. 31-65). Mahwah, NJ: Lawrence Erlbaum.

Schmidt, F. L., & Oh, I. S. (2010). Can synthetic validity methods achieve discriminant validity? *Industrial and Organizational Psychology, 3*(3), 344-350.

Schmidt, F. L., Hunter, J. E., & Raju, N. S. (1988). Validity generalization and situational specificity: A second look at the 75% rule and Fisher's *z* transformation. *Journal of Applied Psychology, 73*(4), 665-672.

Schmidt, F. L., Ocasio, B. P., Hillery, J. M., & Hunter, J. E. (1985). Further within-setting empirical tests of the situational specificity hypothesis in personnel selection. *Personnel Psychology, 38*(3), 509-524.

Schmidt, F.L. (2008). Meta-Analysis: A constantly evolving research integration tool. *Organizational Research Methods*, *11*(1), 96-113.

Spector, P. E., & Levine, E. L. (1987). Meta-analysis for integrating study outcomes: A Monte Carlo study of its susceptibility to Type I and Type II errors. *Journal of Applied Psychology, 72*(1), 3-9.

Steel, P., Schmidt, J., & Shultz, J. (2008). Refining the relationship between personality and subjective well-being. *Psychological Bulletin*, *134*(1), 138.

Steel, P., Johnson, J. W., Jeanneret, P. R., Scherbaum, C. A., Hoffman, C. C., & Foster, J. (2010). At sea with synthetic validity. *Industrial and Organizational Psychology*, *3*(3), 371-383.

Steel, P. D., & Kammeyer-Mueller, J. D. (2002). Comparing meta-analytic moderator estimation techniques under realistic conditions. *Journal of Applied Psychology*, *87*(1), 96.

Steel, P., & Kammeyer-Mueller, J. (2008). Bayesian variance estimation for meta-analysis: Quantifying our uncertainty. *Organizational Research Methods, 11*(1), 54-78.

Steel, P., & Kammeyer-Mueller, J. (2009). Using a meta-analytic perspective to enhance job component validation. *Personnel Psychology*, *62*(3), 533-552.

Steel, P., Kammeyer-Mueller, J., & Paterson, T. A. (2015). Improving the meta-analytic assessment of effect size variance with an informed Bayesian prior. *Journal of Management*, *41*(2), 718-743.

Steel, P., Nguyen, B., & Kammeyer-Mueller, J. (2013). *Removing meta-analytic bias: Bayesian variance estimation with an informed prior.* Poster session presented at the annual meeting of the Society for Industrial Organizational Psychology, Houston, TX.

Tett, R.P., Hundley, N., & Christiansen, N.D. Forthcoming. Meta-Analysis and the Myth of Generalizability. *Industrial and Organizational Psychology: Perspectives on Science and Practice.*

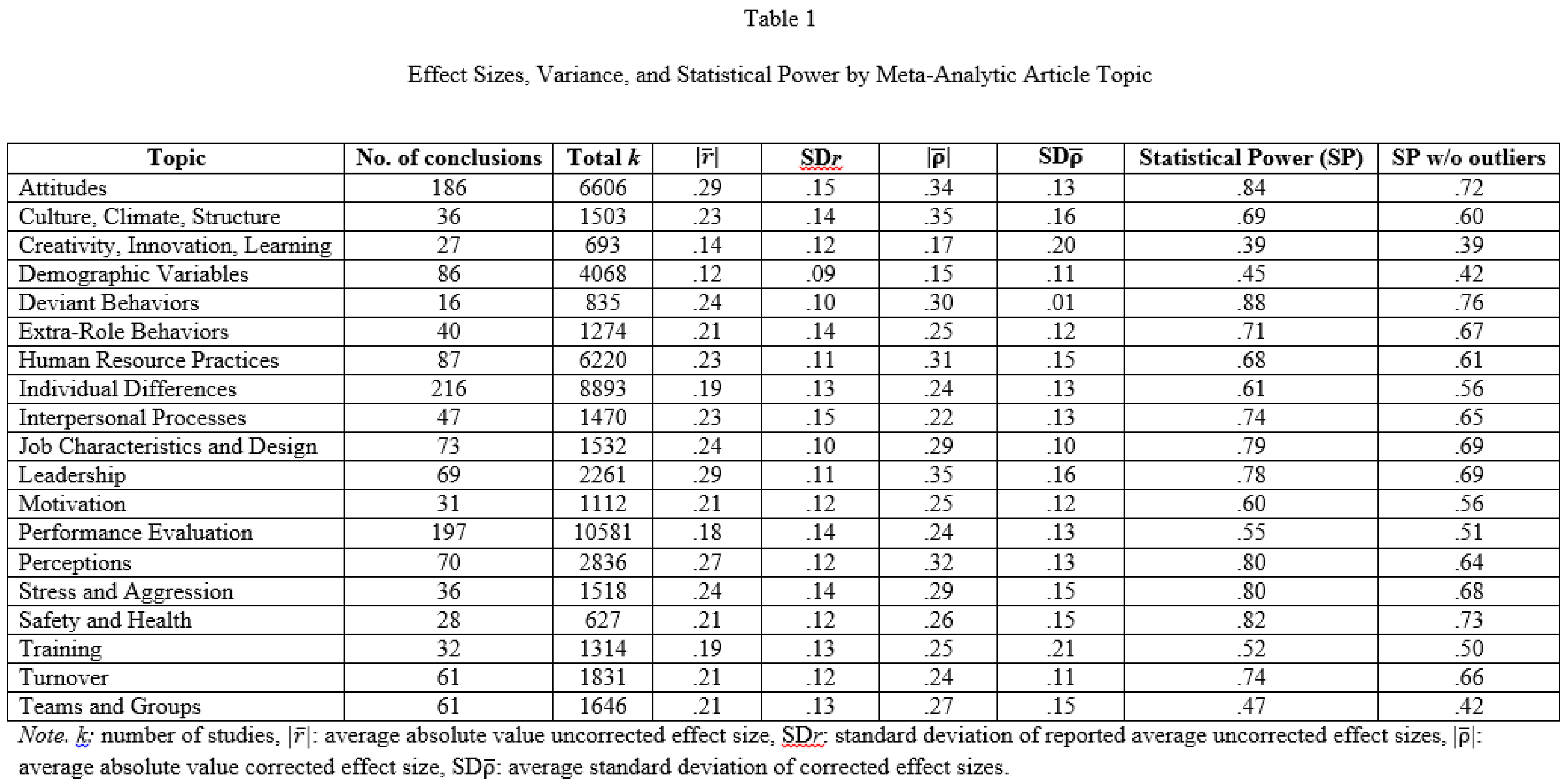


Table 1

Effect Sizes, Variance, and Statistical Power by Meta-Analytic Article Topic

Topic	No. of conclusions	Total *k*	$\lvert\bar{r}\rvert$	SD*r*	$\lvert\bar{\rho}\rvert$	SD$\bar{\rho}$	Statistical Power (SP)	SP w/o outliers
Attitudes	186	6606	.29	.15	.34	.13	.84	.72
Culture, Climate, Structure	36	1503	.23	.14	.35	.16	.69	.60
Creativity, Innovation, Learning	27	693	.14	.12	.17	.20	.39	.39
Demographic Variables	86	4068	.12	.09	.15	.11	.45	.42
Deviant Behaviors	16	835	.24	.10	.30	.01	.88	.76
Extra-Role Behaviors	40	1274	.21	.14	.25	.12	.71	.67
Human Resource Practices	87	6220	.23	.11	.31	.15	.68	.61
Individual Differences	216	8893	.19	.13	.24	.13	.61	.56
Interpersonal Processes	47	1470	.23	.15	.22	.13	.74	.65
Job Characteristics and Design	73	1532	.24	.10	.29	.10	.79	.69
Leadership	69	2261	.29	.11	.35	.16	.78	.69
Motivation	31	1112	.21	.12	.25	.12	.60	.56
Performance Evaluation	197	10581	.18	.14	.24	.13	.55	.51
Perceptions	70	2836	.27	.12	.32	.13	.80	.64
Stress and Aggression	36	1518	.24	.14	.29	.15	.80	.68
Safety and Health	28	627	.21	.12	.26	.15	.82	.73
Training	32	1314	.19	.13	.25	.21	.52	.50
Turnover	61	1831	.21	.12	.24	.11	.74	.66
Teams and Groups	61	1646	.21	.13	.27	.15	.47	.42

Note. *k*: number of studies, $\lvert\bar{r}\rvert$: average absolute value uncorrected effect size, SD*r*: standard deviation of reported average uncorrected effect sizes, $\lvert\bar{\rho}\rvert$: average absolute value corrected effect size, SD$\bar{\rho}$: average standard deviation of corrected effect sizes.

*Table 1 was previously published in Paterson, Harms, Steel, & Crede, 2016. Used with permission.

Table 2
Correlations among Coded Variables

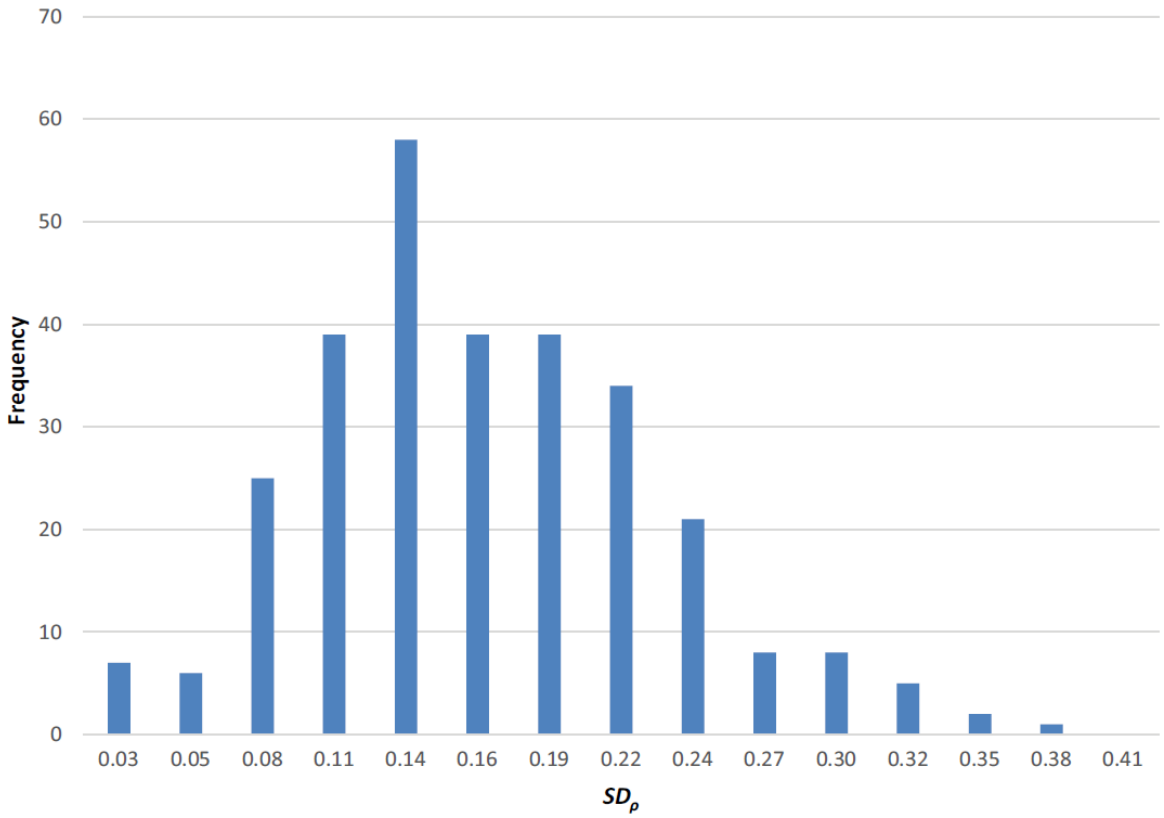
Variable	1	2	3	4	5	6
1. *k*	*r*(*N*)					
2. *N*	.10* (762)	—				
3. *ρ*	.01 (689)	-.03 (688)	—			
4. SD*ρ*	.11* (560)	.10* (559)	.26** (537)	—		
5. Year pub.	-.05 (776)	-.01 (762)	-.08* (690)	.07 (562)	—	
6. % Unpub.	.02 (376)	.14** (366)	-.07 (330)	.04 (287)	.25** (456)	—

*p < .05,

**p < .01

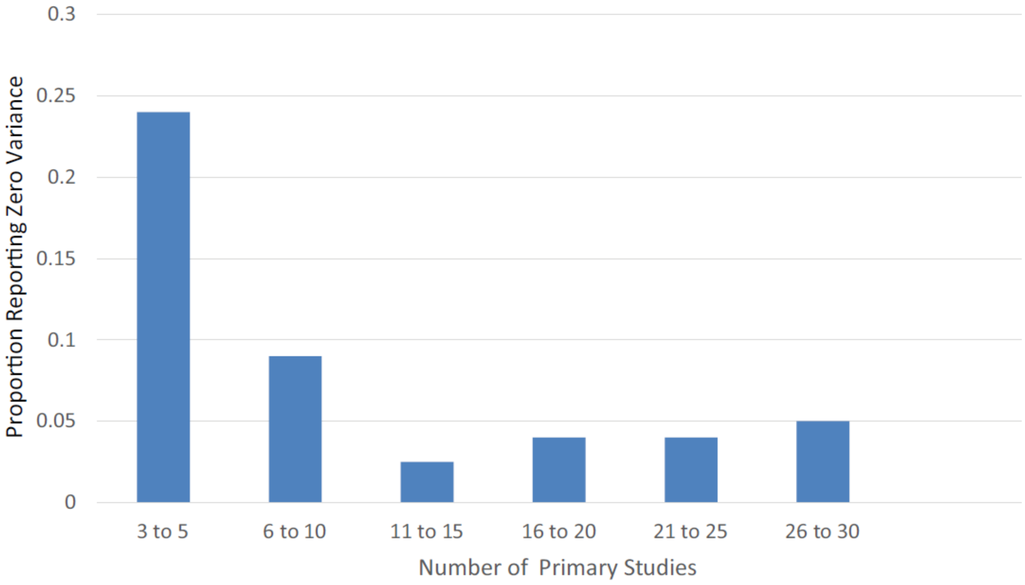
*Note: Table 2 was previously published in Paterson, Harms, Steel, & Crede, 2016. Used with permission.

Figure 1
Standard Deviation of the Effect Size (SD_ρ) Distribution of 292 Management Related Meta-Analyses Based on 25 or More Studies



Note: Figure 1 was previously published in Steel, Kammeyer-Mueller, and Paterson, 2015. Used with permission.

Figure 2
Frequency of Zero Variance Reported in 292 Management Meta-Analyses With 30 or Fewer Studies



Note: Figure 2 was previously published in Steel, Kammeyer-Mueller, and Paterson, 2015. Used with permission.