Journal of Organizational Behavior, J. Organiz. Behav. **32**, 1033–1043 (2011)
Published online 22 July 2010 in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/job.719

Researcher's Notebook

Best-practice recommendations for estimating interaction effects using meta-analysis

HERMAN AGUINIS[1]*, RYAN K. GOTTFREDSON[1] AND THOMAS A. WRIGHT[2]
[1]*Department of Management & Entrepreneurship, Kelley School of Business, Indiana University, 1309 E. 10th Street, Bloomington, Indiana 47405-1701, U.S.A.*
[2]*Department of Management, 215 Calvin Hall, Kansas State University, Manhattan, Kansas 66506, U.S.A.*

Summary One of the key advantages of meta-analysis (i.e., a quantitative literature review) over a narrative literature review is that it allows for formal tests of interaction effects—namely, whether the relationship between two variables is contingent upon the value of another (moderator) variable. Interaction effects play a central role in organizational science research because they highlight boundary conditions of a theory: Conditions under which relationships change in strength and/or direction. This article describes procedures for estimating interaction effects using meta-analysis, distills the technical literature for a general readership of organizational science researchers, and includes specific best-practice recommendations regarding actions researchers can take before and after data collection to improve the accuracy of substantive conclusions regarding interaction effects investigated meta-analytically. Copyright © 2010 John Wiley & Sons, Ltd.

Keywords: meta-analysis; methodology; research synthesis; literature review

Introduction

Interaction, also labeled moderating, effects provide information on whether the relationship between two variables is contingent upon the value of a third variable (Aguinis & Gottfredson, in press). Understanding a theory's moderators means that we understand a theory's boundary conditions. Thus, understanding interaction effects is a key issue for theory development and testing as well as practical applications of a theory. The goal of this article is to provide an overview, together with best-practice recommendations, regarding how to test hypotheses, assess the possible presence and magnitude, and interpret interaction effects using meta-analysis. There is an extensive body of technical literature on meta-analysis. Much of this literature is quite specialized, including analytic work that is mathematically sophisticated as well as Monte Carlo simulations involving lengthy and complex procedures and results. Due to the nature of this research, much of this work is not easily accessible to researchers with the usual methodological and statistical background obtained from doctoral-level training in the organizational sciences. Accordingly, this manuscript distills the technical literature for a general readership and includes specific best-practice recommendations that researchers will be able to implement in their own quest for interaction effects using meta-analysis.

* Correspondence to: Herman Aguinis, Department of Management & Entrepreneurship, Kelley School of Business, Indiana University, 1309 E. 10th Street, Bloomington, IN 47405-1701, U.S.A. E-mail: haguinis@indiana.edu

Received 17 June 2010, Accepted 21 June 2010

Estimating Interaction Effects Meta-analytically

Meta-analysis refers to a family of data collection and analysis techniques whose goal is to produce a quantitative review of a body of literature. The word *meta* indicates that a meta-analysis occurs later than and also transcends the original analysis. A meta-analysis consists of extracting effect-size estimates from studies conducted previously and computing a summary across-study effect size as well as the variance of these estimates. Meta-analyzed effect sizes may be extracted from studies using experimental (i.e., providing evidence regarding causality) or passive observational (i.e., providing evidence regarding covariation) designs. Based on gathering primary-level study effect sizes, meta-analysts have two principal goals: (1) to estimate the overall strength and direction of an effect or relationship and (2) to estimate the across-study variance in the effect or relationship estimates and the factors (i.e., moderator variables) that explain this variance (Aguinis, Pierce, Bosco, Dalton, & Dalton, in press). This second goal involves the investigation of interaction effects.

Meta-analysts assessing interaction effects face numerous choices regarding how to go about doing so. However, meta-analysts seldom explain or justify their choices. As noted by Aguinis, Sturman, and Pierce (2008), ''it seems that meta-analysts often choose a specific meta-analytic technique based on habit, the availability and their familiarity with a specific software package, or usage trends in specific topic areas rather than the relative merits of each of the available approaches'' (p. 10). One of the biggest challenges faced by researchers interested in testing interaction effect hypotheses meta-analytically is what specific steps to take before and after the data are collected and how to choose among the choices available in conducting a meta-analysis given the options available.

The first choice a researcher faces is the type of meta-analytic model to use. The two most established choices are (a) a *random-effects* (RE) model and (b) a *fixed-effect* or *common-effect* (CE) model. These two models have different underlying assumptions and, as such, the choice for one or the other model can affect substantive conclusions regarding the presence of moderating effects. The RE model assumes that the studies in the meta-analysis are a random sample from a superpopulation of possible studies and that the superpopulation of true effect sizes is normally distributed. In the RE model, the variance in the distribution of observed effects is attributed to: (a) within-study variance (i.e., mainly due to sampling error), and (b) between-study variance (i.e., due to differences of true effect sizes in the superpopulation). In contrast, the CE model assumes that there is one common true effect underlying the distribution of observed effects. Hence, in the CE model the variance in the distribution of observed effect sizes is attributed solely to within-study sampling variance. The confidence interval around the mean effect size is wider when using an RE model compared to a CE model because the RE confidence interval is describing a large superpopulation of (most likely heterogeneous) effect sizes while the CE confidence interval is describing only a small subset of homogeneous superpopulation effect sizes.

Should a meta-analyst choose an RE or an CE model? In organizational science research, an RE model is almost always the more appropriate choice of the two. Using an RE model allows for the between-study variance to take on any value (including zero) whereas a CE model forces the between-study variance to be zero. Thus, the RE model can be conceptualized as the general case and the CE model as a specific case of the RE model in which the between-study variance in true effect sizes is assumed to be zero. However, to realize the full benefits of an RE meta-analysis, two important assumptions must be satisfied: (a) the observed effect sizes have been randomly sampled from a superpopulation of true effect sizes and (b) the superpopulation of true effect sizes is normally distributed (Bonett, in press). Schulze (2004, p. 41) warned that the random-studies assumption is a potentially serious limitation of the RE model. In addition, the superpopulation normality assumption also should not be taken lightly because RE inferential methods can perform poorly under minor deviations from normality (Bonett, in press).

Alternatively, the CE model would be appropriate when each of the primary-level studies included in a meta-analysis is functionally identical (Borenstein, Hedges, Higgins, & Rothstein, 2009, Chapter 13). This condition rarely exists in organizational research because primary-level studies would have to include samples of participants from exactly the same population, the same team of researchers, the same research design, the same measures for the independent and dependent variables, and the same procedures for all other aspects of the study. This type of

DOI: 10.1002/job

meta-analysis could occur in the biological, medical, and health sciences. For example, the same team of researchers could conduct 20 separate randomized trial studies examining the effect of a drug using samples from the same population of individuals and exactly the same procedures in each study and then conduct a subsequent meta-analysis using the 20 effect-size estimates. But, this situation is almost impossible in the organizational sciences. In short, a CE meta-analysis will be difficult to justify in organizational research.

Alternatives to the more traditional CE and RE methods are the *varying coefficient* (VC) methods proposed by Bonett (2008, 2009, in press). Like the RE methods, the VC methods do not assume equality of effect sizes across studies but, unlike the RE methods, the VC methods do not rely on the random-studies or the normal distribution of superpopulation effect sizes assumptions. However, the VC methods have the same limitation as the CE methods in that they provide inferences only to a subset of the superpopulation. The VC methods will almost always be preferred to the CE methods, but if the random-studies assumption and superpopulation normality assumption can be justified, the RE methods will be preferred to the VC methods. Another advantage of the VC approach is the availability of confidence interval methods for assessing the magnitude of interaction effects (Bonett, 2008, 2009, in press).

Once the meta-analytic model has been chosen, the second important choice that meta-analysts face is what type of effect-size estimate to extract from primary-level studies. Usually, the choice is between an r (i.e., Pearson's correlation coefficient) or a d (i.e., standardized difference between two group means). Once that choice is made, then other types of statistics reported in the primary-level studies (e.g., t, means and standard deviations, F) can be converted to the chosen focal effect size r or d. Those researchers who cumulate rs tend to use the Hunter–Schmidt (2004) meta-analytic approach, whereas those who meta-analyze ds tend to use the procedures developed by Hedges and colleagues (Borenstein et al., 2009; Hedges & Olkin, 1985; Hedges & Vevea, 1998).

The choice regarding focal effect size is also often forced by the software packages available to conduct a meta-analysis: The *Hunter–Schmidt Programs* focus on the Hunter–Schmidt approach whereas *Comprehensive Meta-analysis* focuses on the Hedges et al. approach. It is technically possible to implement the Hunter–Schmidt approach using ds (Hunter & Schmidt, 2004, Chapter 7) and the Hedges et al.'s approach using rs (e.g., Aguinis & Pierce, 1998). However, this is rarely done, probably because the Hunter and Schmidt procedures were developed with one specific research domain in mind: Personnel selection and the situational specificity hypothesis (i.e., the hypothesis that validity coefficients are affected by contextual factors) in which the focal effect size is the validity coefficient expressed as a Pearson's r. Alternatively, the Hedges et al.'s procedures were originally designed to meta-analyze experimental studies, which typically report ds. Another reason for why the Hedges et al.'s procedures are usually implemented using ds and the Hunter–Schmidt procedures using rs is that the former analyzes Fisher transformed correlations (i.e., z_r) whereas the latter analyzes raw correlations. Analyzing z_rs is not a problem if the assumptions of the CE model can be satisfied. However, when the study population correlations are not identical, using z_rs introduces three important challenges regarding the interpretation of results. First, the estimator of the reverse-transformed average will be biased (in both CE and RE models). Second, the parameters of the meta-regression model, which we describe later in our article, will not have a meaningful interpretation because the slope parameters describe changes in z_r values rather than r values. Third, τ^2, which is the variance of the true effect sizes in the superpopulation, refers to the variance of z_rs (instead of raw correlations), which may lead to ambiguous interpretations. Note that the interpretation of slope coefficients and τ^2 does not pose these challenges when modeling ds.

There seems to be a clear divide in terms of researchers who cumulate rs versus ds and the subsequent choices guided by the initial decision regarding focal effect-size estimate. Specifically, Aguinis, Dalton, Bosco, Pierce, and Dalton (in press) reported that 90 per cent of meta-analyses published in *Academy of Management Journal* (AMJ), *Journal of Applied Psychology* (JAP), *Journal of Management* (JOM), *Personnel Psychology* (PPsych), and *Strategic Management Journal* (SMJ) from January 1982 through August 2009 cumulated rs. Showing the precise opposite pattern or results, Schmidt, Oh, and Hayes (2009) reported that more than 90 per cent of meta-analyses published in *Psychological Bulletin* over the past 20 years cumulated ds. Most likely, this difference is due to the passive observational nature of the primary-level studies meta-analyzed by articles published in AMJ, JAP, JOM, PPsych, and SMJ reporting rs and the experimental nature of primary-level studies meta-analyzed by articles

DOI: 10.1002/job

published in *Psychological Bulletin* reporting *d*s. Note that the choice of *r* or *d* has implications for understanding the practical significance of the effect: r^2 refers to proportion of variance explained in an outcome, whereas *d* refers to standardized differences between groups. Thus, discussing practical significance based on r^2 involves discussing the relative usefulness of a (predictor) variable in terms of explaining fluctuations in another (criterion) variable, whereas discussing practical significance based on *d* involves a discussion of how an intervention may have differential effects across groups. Although other meta-analytic approaches have been proposed over the years, the Hedges et al. and Hunter–Schmidt approaches remain by far the most popular to estimate interaction effects using meta-analysis in the organizational sciences (Aguinis, Dalton, et al., in press). Accordingly, we discuss each of these approaches next.

Hedges et al. approach

Using the Hedges et al. (Borenstein et al., 2009; Hedges & Olkin, 1985; Hedges & Vevea, 1998) approach to estimate interaction effects involves computing an unbiased effect-size estimate from each primary-level study and then computing a summary effect size, usually a mean, based on the study-level estimates. To compute the summary effect size, each study-level effect size is weighted by the inverse of its variance, which in an RE model includes both a within-study and a between-study component and in a CE model includes a within-study component only.

Note that some meta-analysts assess the extent to which there is heterogeneity (i.e., variance in true effect sizes) by computing the statistic *Q*, which follows a χ^2 distribution with $df = k - 1$ (k is the number of studies included in a meta-analysis). However, similar to all other tests of significance, not rejecting a null hypothesis of homogeneity may be caused by insufficient statistical power and not lack of heterogeneity (i.e., the statistical power of the *Q* statistic is affected by total sample size and also the number of primary-level studies included in the meta-analysis) (Aguinis & Pierce, 1998). Thus, *Q*'s statistical significance level should *not* be used as a decision-making tool for choosing among RE, CE, and VC models. As noted earlier, this decision should be made based on the goals of a meta-analysis and the nature of the primary-level studies included in the review. Rather than test for exact homogeneity of effect sizes using a *Q* statistic, one could follow the general recommendations of Bonett and Wright (2007) and assess the *degree* of effect-size heterogeneity in the superpopulation by constructing a confidence interval for τ, which is the standard deviation of the true effect sizes in the superpopulation. Methods for computing a confidence interval for τ are also described by Borenstein et al. (2009, pp. 122–124).

Two sets of statistical procedures are used to test for the presence of particular moderating effects depending upon whether the hypothesized moderator is categorical (i.e., variables for which values represent discrete categories—e.g., gender, ethnicity) or continuous (i.e., variables for which, within the limits of the variable's range, any value is possible—e.g., job satisfaction, work motivation). In tests for hypothesized *categorical* moderating effects, each primary-level study is assigned a numerical value based upon the moderator (e.g., gender: 1 = female, 2 = male) and subgrouped according to this coding scheme. In this so-called *subgroup analysis*, the goal is to examine whether effect sizes differ across the subgroups while it is assumed that they do not differ within each subgroup. Although there are three types of subgroup analysis, they are algebraically equivalent and yield the same *p* value.

A type of subgroup analysis which resonates with most organizational science researchers is analogous to ANOVA in primary-level research. This analysis involves computing a between-subgroup statistic Q_B to test that the difference between or among mean within-subgroup effect sizes is zero and a within-subgroup homogeneity statistic Q_W, which is obtained using an underlying CE model within each subgroup (i.e., assuming that all effect sizes within each group share a common population effect). A statistically significant Q_B, which follows a χ^2 distribution with $df = p - 1$ (i.e., p is the number of subgroups), suggests that the subgrouping variable is indeed a moderator. As a follow-up analysis, it is possible to use an RE model for each Q_W (i.e., within each subgroup) to test for further heterogeneity and subdivide subgroups further based on additional hypothesized moderators. In terms of the size of the moderating effect, meta-analysts can compute an R^2 value, which indicates the proportion of variance of the total between-group variance (τ^2) that is explained by each moderator variable.

DOI: 10.1002/job

If a meta-analyst is interested in a hypothesis regarding at least one continuous moderator variable, then the procedure involves conducting a *meta-regression*. Similar to the more familiar multiple regression, meta-regression consists of a regression model involving one or more predictors (i.e., potential moderators) and a criterion (i.e., weighted observed effect sizes). In other words, the primary-level study effect sizes are regressed onto one or more moderators, which can be continuous or categorical. Assuming a situation with *d* as the focal effect size metric and two continuous moderators, the meta-regression model is $\mathbf{d} = \mathbf{X_1}\boldsymbol{\beta_1} + \mathbf{X_2}\boldsymbol{\beta_2} + \boldsymbol{\varepsilon}$, where $\mathbf{d}$ represents a vector of effect sizes, $\mathbf{X_1}$ represents a vector of values for the hypothesized moderator X_1, $\mathbf{X_2}$ represents a vector of values for the hypothesized moderator X_2, and $\boldsymbol{\varepsilon}$ represents a vector of residuals. The statistical significance of each individual moderator (i.e., predictor) can be assessed by computing a *Z*-statistic for each coefficient in the meta-regression model (i.e., $Z_1 = B_1/SE_{B_1}$ and $Z_2 = B_2/SE_{B_2}$). In addition, confidence intervals can be created around each slope coefficient. Regarding the size of a moderating effect, similar to subgroup analysis for categorical moderators, meta-analysts can compute an R^2 value, which indicates the proportion of variance of the total between-group variance (τ^2) that is explained by the moderator variables. Also, an examination of the individual regression coefficients provides an estimate of the strength of the relationship between each moderator and the effect sizes.

Finally, note that moderators could be substantive or methodological in nature. For example, a meta-analyst can test the hypothesis that measurement error is a continuous moderator and, hence, assess whether the degree of measurement error reported in each primary-level study is a predictor of effect size in a meta-regression model. The effects of measurement error, range restriction, and other methodological and statistical artifacts on the variance of across-study effect sizes receive a different treatment in the Hunter–Schmidt approach, as described next.

Hunter–Schmidt approach

Similar to the Hedges et al.'s approach, the goal of the Hunter–Schmidt approach is to estimate a summary effect size across primary-level studies, the variance of these effects, and the factors that explain this variability (i.e., moderating effects). Although the Hedges et al. approach can be implemented using an RE or CE model, the Hunter–Schmidt approach is exclusively an RE model. The Hunter–Schmidt approach has also been labeled validity generalization or psychometric meta-analysis. Proponents of psychometric approaches to meta-analysis extend arguments from measurement theory suggesting that a substantial portion of the variability observed in the relationship between two variables across primary-level studies is not substantive but the result of methodological and statistical artifacts.

Psychometric meta-analysis posits that if these artifacts are not controlled for via research design prior to data collection or via corrective measures after the data have been collected, researchers may conclude that there is an interaction effect whereas, in fact, the source of variance across primary-level studies may be artifactual and not substantive in nature (Aguinis, 2001; Aguinis & Whitehead, 1997). These artifacts include sampling error, error of measurement in the independent and dependent variables, dichotomization of a continuous dependent variable, dichotomization of a continuous independent variable, range variation in the independent variable, range variation in the dependent variable due to attrition artifacts, deviation from perfect construct validity in the independent variable, deviation from perfect construct validity in the dependent variable, reporting or transcriptional errors, variance due to extraneous factors, and scale coarseness (Aguinis, Pierce, & Culpepper, 2009; Hunter & Schmidt, 2004; Le, Schmidt, & Putka, 2009).

The goal of the corrections is not to eliminate any kind of source of variance in the distribution of primary-level effect-size estimates, but rather to minimize the across-study variability that is caused by methodological and statistical artifacts. So, the first step involves estimating the overall variance of effect sizes across studies. The second step involves computing the variance attributed to methodological and statistical artifacts. The variance attributed to artifacts can be computed based on information provided by each study (e.g., reliability estimates and range restriction values reported in the primary-level studies included in the meta-analysis) or by using an artifact distribution approach. The third step includes subtracting the variance in effect sizes due to artifacts from the total

DOI: 10.1002/job

variance. The resulting variance due to actual differences in true effect sizes and not due to sampling error and other artifacts, which is σ_ρ^2 for the case of correlation coefficients, is conceptually similar to τ^2 in the Hedges et al. approach. An important difference between these two variance estimates is that σ_ρ^2 excludes variance attributed to sampling error and also methodological and statistical artifacts, whereas τ^2 only excludes variance attributed to sampling error.

There are several suggestions regarding how to determine whether a hypothesis about an interaction effect has received support. The most frequently used is the 75 per cent heuristic: If less than 75 per cent of variance in effect-size estimates across primary-level studies is due to artifacts, it is likely that there is substantive variance and the search for moderators is warranted (Hunter & Schmidt, 2004). Although there is no formal statistical justification for this heuristic, the argument is that meta-analysts can never correct for all artifacts that cause effect-size estimates to vary across studies because they may not have sufficient information to implement a correction and also because some factors are simply uncorrectable (e.g., deviation from perfect construct validity in the independent variable, transcriptional errors, variance due to extraneous factors). The Hunter–Schmidt approach assumes that if correctable factors account for at least 75 per cent of the across-study variance, then it is likely that the remaining variance is accounted for by uncorrectable factors and is also artifactual in nature. Also, reporting a credibility interval around the mean across-study effect (which is obtained by weighting each primary-level study by its sample size) provides useful information regarding the heterogeneity of effects. For example, the lower bound of a 90 per cent credibility interval is the value above which 90 per cent of the true effect sizes are expected to lie. Finally, in tests of categorical moderator variables, similar to the Hedges et al. approach, studies are assigned a numerical value based on the moderator and subgrouped accordingly.

Best-practice Design, Measurement, and Analysis Recommendations for Estimating Moderating Effects Using Meta-Analysis

Meta-analysis is often referred to and treated as a data-analytic technique. In fact, its very name implies that it is about ''analysis.'' However, it is useful to think of meta-analysis not only in terms of data analysis, but also in terms of research design and measurement. In fact, if there are problems related to design and measurement, it is unlikely that meta-analysis will be used effectively or lead to important theoretical advancements. Also, choices that primary-level researchers make about design and measurement will place constraints on any subsequent meta-analyses. For example, if primary-level studies do not provide information on measurement error for each of the variables, correcting for unreliability in a meta-analysis may turn into a guessing game. Results of such meta-analysis may be contested on the grounds that correction factors for methodological and statistical artifacts were not appropriate. Similarly, if there is not a clear correspondence between the constructs a meta-analysis is attempting to study and the constructs measured in the primary-level studies, then the resulting analysis is likely to lead to inconclusive results.

Before discussing specific best-practice recommendations for testing interaction effect hypotheses meta-analytically, we consider three overarching and fundamental recommendations. Each of these recommendations stem from an important and fairly unique challenge due to the nature of meta-analysis: Meta-analysts use data that have been collected in studies designed by other researchers and variables that have also been measured by other researchers. So, meta-analysts have to rely on choices regarding research design and measurement that were made not by themselves, but by others, and on the primary-level studies available in a particular research domain.

The first overarching best-practice recommendation is to choose the most appropriate meta-analytic model. Given our earlier discussion regarding RE, CE, and VC methods, the recommendation is that in almost all organizational science research domains, the RE and VC methods will almost always be preferred to CE methods.

DOI: 10.1002/job

A second overarching recommendation is to understand whether a meta-analysis aimed at testing a particular interaction effect is possible given the available data. A fundamental issue to address is whether to conduct the meta-analysis to test an interaction hypothesis in the first place. An advantage of the CE and VC methods is that they can be used with as few as 2 studies (Borenstein et al., 2009, p. 363; Bonett, 2008, 2009, in press). In contrast, the RE methods require at least 20 primary-level studies in order to obtain a properly performing CI for the mean effect size assuming approximate normality of the superpopulation of effect sizes (Field, 2005). So, meta-analysts need to answer the following questions: Are there studies available that assessed the relationship to be estimated? Have those studies implemented research designs conducive to valid conclusions in terms of construct, internal, external, and statistical conclusion validity? Will a meta-analysis have sufficient statistical power to detect an interaction effect if it exists (Borenstein et al., 2009, Chapter 29; Hedges & Pigott, 2001, 2004)? Only if these questions are answered in the affirmative will meta-analysis be likely to lead to valid conclusions. If these questions are not answered in the affirmative, then meta-analysis may not be a good approach for testing a particular interaction effect hypothesis and, instead, conducting a primary-level study in which a researcher makes his or her own choices about design and measurement may be a better alternative.

The third overarching best-practice recommendation also stemming from the secondary nature of all meta-analytic databases is that, once a decision has been made to conduct a meta-analysis, there is a need to spell out the procedures implemented and choices made at each stage of the meta-analysis including design, measurement, and analysis. These include a clear definition of the criteria for study inclusion, how studies were searched and selected, how data were extracted from the studies, how coding for moderators took place, and how the data were analyzed (including decision rules) for testing interaction effect hypotheses (APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008). Replicability is a fundamental principle in any scientific endeavor and it is particularly relevant in the context of testing interaction effect hypotheses meta-analytically (Eden, 2002). The choices made in conducting a meta-analysis must be described explicitly and in great detail so that obtained results about interaction effects can be replicated in the future. Having described three fundamental and critical best-practice recommendations, we now turn to more specific recommendations about design, measurement, and analysis. Table 1 includes a summary of these recommendations, which were derived from several sources published in and outside of the organizational sciences (e.g., Borenstein et al., 2009; Cooper, 2010; Cooper, Hedges, & Valentine, 2009; Hunter & Schmidt, 2004).

Regarding design issues, the first recommendation is to establish a clear purpose for the meta-analysis so that a meta-analyst can select appropriate studies to include in the review. In other words, a clear definition of the research domain will lead to clear criteria for study inclusion. As noted by Bobko and Roth (2008), there is a need for ''many more systematic sets of articles focused on up-front thinking about each meta-analytic study's purpose and how this reflection relates to the choice of primary studies, analytic method, coding, etc.'' (p. 124). Second, the search for relevant studies should be conducted using keywords with electronic databases, but also using other procedures such as ancestry searching (also known as citation chasing). Ancestry searching consists of working from the references list from more contemporary references and tracking prior work on which they relied. A third best-practice recommendation in terms of design is to avoid including duplicate studies in the meta-analysis (Wood, 2008). The inclusion of duplicate studies can produce substantial bias in the meta-analytic results. Wood (2008) provided a detailed description of procedures for detecting and avoiding the inclusion of both covert and overt duplicate studies in a meta-analytic database.

Regarding measurement issues, as noted earlier regarding one of the three overarching best-practice recommendations, there should be a clear description of any coding that takes place. Meta-analysts can rely on excellent advice deriving from the content analysis literature in terms of how to create coding protocols (Duriau, Reger, & Pfarrer, 2007) as well as how to gather evidence regarding the validity and reliability of the coding process (LeBreton & Senter, 2008). Second, measures used in the primary-level studies are never perfectly reliable. The decision to correct for measurement error and other artifacts in the meta-analysis depends on whether one is interested in understanding construct-level (i.e., measurement error free) relationships or operational-level (i.e., constructs as usually assessed using fallible measures) relationships. For example, one meta-analyst may have a

DOI: 10.1002/job

Table 1. Summary of best-practice recommendations for estimating interaction effects using meta-analysis

Pre data collection recommendations
1. Choose to implement a random-effects (RE), varying-coefficient (VC), or common-effect (CE) model. In almost all situations, RE and VC models are more appropriate than a CE model. RE methods are preferred over the VC methods if the random-studies and superpopulation normality assumptions can be justified.
2. Assess whether there are sufficient primary-level studies available to conduct a meta-analysis in a specific research domain and determine the statistical power to detect hypothesized interaction effects.
3. Create a thorough and detailed description of all procedures implemented and choices made at each stage of the meta-analysis ranging from design to measurement and analysis.
4. Define the research domain clearly so that the criteria for study inclusion are also clearly defined.
5. Search for relevant studies using keywords with electronic databases, but also use other procedures (e.g., ancestry searching).
6. Avoid including duplicate studies.
7. Create coding protocols following best-practice recommendations derived from the content analysis literature.
Post data collection recommendations
8. Gather and report evidence regarding the reliability and validity of the coding protocols.
9. Implement corrections for statistical and methodological artifacts (e.g., measurement error, range restriction) only if there is an interest in understanding construct-level interaction effects. Do not implement corrections if there is an interest in observed (i.e., operational) relationships.
10. Use *r*s and the Hunter–Schmidt (2004) approach if primary-level studies report *r*s and there is an interest in applying corrections for methodological and statistical artifacts.
11. Use meta-regression methods to assess theory-based interaction hypotheses.
12. Assess the degree of effect-size heterogeneity in the superpopulation of true effect sizes by constructing a confidence interval for τ (i.e., *SD* of the true effect size superpopulation).
13. Use the trim-and-fill method to estimate extent of possible publication bias.

theoretical interest in understanding whether the relationship between job satisfaction and organizational citizenship behavior (OCB) depends on an individual's identification with organizational goals. In such a situation, implementation of corrections would be appropriate and this meta-analysis would answer the question of whether, at the construct level, there is an interaction effect between job satisfaction and goal identification on OCB. However, if a meta-analyst is interested in predicting the specific level of OCB that would be observed for employees with various levels of job satisfaction and goal identification, the measurement error correction would not be appropriate because predictions will be made with the (less than perfectly reliable) measures available. In short, implementing measurement error corrections, as well as other corrections to address statistical and methodological artifacts (e.g., range restriction) is dictated by the meta-analyst's purpose: Corrections are appropriate if the goal is to understand construct-level relationships (What would results look like if all primary-level studies had been conducted without any methodological imperfections?), but not appropriate if the goal is to understand and predict actual (imperfect) observed scores.

Regarding analysis issues, as noted earlier, meta-analysts can choose to use *r*s or *d*s. Although corrections for methodological and statistical artifacts can be applied to either type of effect size (Aguinis & Pierce, 1998), the Hunter–Schmidt approach provides a more detailed analysis of how this is done using *r*s. In addition, if the primary-level studies report *d*s, they can be converted to *r*s (i.e., point–biserial correlations). However, if the primary-level studies report *r*s, converting them to *d*s would require artificially crippling the data (Schmidt, 2008). In short, the recommendation is to use *r*s and the Hunter–Schmidt procedures if primary-level studies report *r*s and there is an interest in applying corrections for methodological and statistical artifacts.

Second, in terms of testing the interaction effect hypothesis, researchers can use the meta-regression procedures described earlier and also compute confidence intervals to assess the size of the effects. Finally, another analysis issue is the need to consider possible publication bias (McDaniel, Rothstein, & Whetzel, 2006). Publication bias occurs when ''the research that appears in the published literature is systematically unrepresentative of the population of completed studies'' (Rothstein, Sutton, & Borenstein, 2005, p. 1). For example, there is publication bias when studies that show a strong and positive relationship between two variables are available, but those that do

DOI: 10.1002/job

not show that relationship are not (e.g., because they did not survive the peer-review process due to reporting statistically nonsignificant results). There are other reasons for the possible presence of publication bias such as when studies contravene financial, political, ideological, professional, or other interests of investigators, research sponsors, and journal editors. Publication bias means that the sample of studies is not random and, hence, that an assumption of the RE methods has been violated. The trim-and-fill technique allows meta-analysts to understand whether there is publication bias and its potential impact (McDaniel et al., 2006). The main goal of the trim-and-fill method is not to find the values of missing studies but, rather, to assess how much the value of the estimated summary effect size might change if there are missing studies. Essentially, the trim-and-fill technique involves first inspecting a funnel plot of study-level effect-size estimates to determine whether there is (funnel plot is asymmetrical) or is not (funnel plot is symmetrical) possible publication bias. Second, if the funnel plot is asymmetrical, then the trim-and-fill method imputes the missing study-level effect-size estimates needed to make the funnel plot symmetrical, adds them to the meta-analysis, and calculates a trim-and-fill adjusted mean effect-size estimate. Third, the meta-analyst examines the size of the difference between the value of the observed mean effect-size estimate and the value of the trim-and-fill adjusted mean effect-size estimate. Note, however, that although the trim-and-fill method seems to be the best current technique for detecting potential publication bias, it does have some limitations and the most important one is that it can confuse true heterogeneity with bias.

Concluding Comments

As noted by Eden (2002), meta-analysis ''has won widespread recognition as an indispensable tool... and has replaced the traditional narrative assessment of a body of research as a better way of conducting a literature review'' (p. 841). Meta-analysis has also become the ''gold standard'' in terms of methodological procedures available to summarize empirical evidence available to guide management practice. In this article, we offered best-practice recommendations to implement before and after data are collected with the goal of estimating interaction effects meta-analytically. An overarching recommendation is that procedures must be spelled out in detail so that meta-analyses are fully replicable. After all, the impetus for the development of meta-analytic methods was that traditional narrative reviews are subjective, idiosyncratic, and difficult to replicate. It would be terribly ironic if meta-analysis falls prey to the same criticism that served as a major impetus for its own foundation and development.

Author biographies

Herman Aguinis is the Dean's Research Professor, Professor of Organizational Behavior and Human Resources, and the Director of the Institute for Global Organizational Effectiveness in the Kelley School of Business, Indiana University. His current research interests span several organizational behavior, human resource management, and research methods and analysis topics.

Ryan K. Gottfredson is a doctoral student in organizational behavior in the Kelley School of Business, Indiana University. He received his Bachelor's degree from Brigham Young University in Business Administration with a major in Finance. His current research interests includes research methods and analysis, trust violations, and personnel selection.

Thomas A. Wright earned his Ph.D. degree at the University of California, Berkeley and is the Jon Wefald Leadership Chair in the College of Business at Kansas State University. His current research interests include

DOI: 10.1002/job

optimizing peak performance, character-based leadership, finding innovative ways to enhance employee health and well-being, and such methodological topics as sample size determination and interval estimation.

References

Aguinis, H. (2001). Estimation of sampling variance of correlations in meta-analysis. *Personnel Psychology*, *54*, 569–590.

Aguinis, H., Dalton, D. R., Bosco, F. A., Pierce, C. A., & Dalton, C. M. (in press) Meta-analytic choices and judgment calls: Implications for theory building and testing, obtained effect sizes, and scholarly impact. *Journal of Management.*

Aguinis, H., & Gottfredson, R. K. (in press) Best-practice recommendations for estimating interaction effects using moderated multiple regression. *Journal of Organizational Behavior.*

Aguinis, H., & Pierce, C. A. (1998). Testing moderator variable hypotheses meta-analytically. *Journal of Management*, *24*, 577–592.

Aguinis, H., Pierce, C. A., Bosco, F. A., Dalton, D. R., & Dalton, C. M. (in press). Debunking myths and urban legends about meta-analysis. *Organizational Research Methods.*

Aguinis, H., Pierce, C. A., & Culpepper, S. A. (2009). Scale coarseness as a methodological artifact: Correcting correlation coefficients attenuated from using coarse scales. *Organizational Research Methods*, *12*, 623–652.

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*, 9–34.

Aguinis, H., & Whitehead, R. (1997). Sampling variance in the correlation coefficient under indirect range restriction: Implications for validity generalization. *Journal of Applied Psychology*, *82*, 528–538.

APA Publications and Communications Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist*, *63*, 839–851.

Bobko, P., & Roth, P. L. (2008). Psychometric accuracy and (the continuing need for) quality thinking in meta-analysis. *Organizational Research Methods*, *11*, 114–126.

Bonett, D. G. (2008). Meta-analytic interval estimation for bivariate correlations. *Psychological Methods*, *13*, 173–189.

Bonett, D. G. (2009). Meta-analytic interval estimation for standardized and unstandardized mean differences. *Psychological Methods*, *14*, 225–238.

Bonett, D. G. (in press). Varying coefficient meta-analysis methods for alpha reliability. *Psychological Methods.*

Bonett, D. G., & Wright, T. A. (2007). Comments and recommendations regarding the hypothesis testing controversy. *Journal of Organizational Behavior*, *28*, 647–659.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. West Sussex, England: John Wiley & Sons.

Cooper, H. M. (2010). *Research synthesis and meta-analysis: A step-by-step approach* (4th edition). Los Angeles, CA: Sage.

Cooper, H. M., Hedges, L. V. & Valentine, J. C. (Eds.). (2009). *The handbook of research synthesis and meta-analysis* (2nd edition). New York, NY: Russell Sage Foundation.

Duriau, V. J., Reger, R. K., & Pfarrer, M. D. (2007). A content analysis of the content analysis literature in organization studies: Research themes, data sources, and methodological refinements. *Organizational Research Methods*, *10*, 5–34.

Eden, D. (2002). Replication, meta-analysis, scientific progress, and AMJ's publication policy. *Academy of Management Journal*, *45*, 841–846.

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

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

Hedges, L. V., & Pigott, T. D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods*, *6*, 203–217.

Hedges, L. V., & Pigott, T. D. (2004). The power of statistical tests for moderators in meta-analysis. *Psychological Methods*, *9*, 426–445.

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. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd edition). Thousand Oaks, CA: Sage.

Le, H., Schmidt, F. L., & Putka, D. J. (2009). The multifaceted nature of measurement artifacts and its implications for estimating construct-level relationships. *Organizational Research Methods*, *12*, 165–200.

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*, *11*, 815–852.

DOI: 10.1002/job

McDaniel, M. A., Rothstein, H. R., & Whetzel, D. L. (2006). Publication bias: A case study of four test vendors. *Personnel Psychology*, *59*, 927–953.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). Publication bias in meta-analysis. In: H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment, and adjustments* (pp. 1–7). West Sussex, England: John Wiley & Sons.

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

Schmidt, F. L., Oh, I., & Hayes, T. L. (2009). Fixed- versus random-effects models in meta-analysis: Model properties and an empirical comparison of differences in results. *British Journal of Mathematical and Statistical Psychology*, *62*, 97–128.

Schulze, R. (2004). *Meta-analysis: A comparison of approaches*. Cambridge, MA: Hogrefe & Humber.

Wood, J. A. (2008). Methodology for dealing with duplicate study effects in a meta-analysis. *Organizational Research Methods*, *11*, 79–95.

DOI: 10.1002/job