Journal of Management
1998, Vol. 24, No. 5, 577-592

Testing Moderator Variable Hypotheses Meta-Analytically

Herman Aguinis
University of Colorado at Denver
Charles A. Pierce
Montana State University

We propose and illustrate a three-step procedure for testing moderator variable hypotheses meta-analytically. The procedure is based on Hedges and Olkin's (1985) meta-analytic approach, yet it incorporates study-level corrections for methodological and statistical artifacts that are typically advocated and used within psychometric approaches to meta-analysis (e.g., Hunter & Schmidt, 1990). The three-step procedure entails: (a) correcting study-level effect size estimates for across-study variability due to methodological and statistical artifacts, (b) testing the overall homogeneity of study-level effect size estimates after the artifactual sources of variance have been removed, and (c) testing the effects of hypothesized moderator variables.

Quantitative reviews of a research domain (i.e., meta-analysis, MA) are consensually accepted in numerous management subdisciplines, as well as other social sciences (e.g., Aguinis, Pierce, & Quigley, 1993, 1995; Cooper & Hedges, 1994; Cotton & Tuttle, 1986; Dobbins & Platz, 1986; Johnson, 1989). A critical advantage of MA compared to a narrative literature review strategy is that it permits the formal testing of hypotheses regarding the effects of *moderator variables*. Variable Z is defined as a moderator of the relationship between variables X and Y when the nature of this relationship is contingent upon values or levels of Z (Aguinis, 1995; Aguinis, Bommer, & Pierce, 1996; Aguinis & Pierce, 1998; Zedeck, 1971). Similar to primary researchers (e.g., Aguinis, Pierce, & Stone-Romero, 1994; Aguinis & Stone-Romero, 1997; Stone-Romero, Alliger, & Aguinis, 1994), meta-analysts often test theoretically-derived moderator variable hypotheses (Cooper & Lemke, 1991; Mullen, Salas, & Miller, 1991). For example, organizational behavior meta-analysts have long been interested in investigating the moderating effect of pressure for production (Z) on the relationship

Direct all correspondence to: Herman Aguinis, College of Business & Administration, University of Colorado at Denver, Campus Box 165, P.O. Box 173364, Denver, CO 80217-3364; e-mail: <haguinis@castle.cudenver.edu>. World Wide Web addresses: http://www.cudenver.edu/~haguinis and http://www.montana.edu/wwwpy/cppage.html.

between job satisfaction (X) and job performance (Y) (Iaffaldano & Muchinsky, 1985; Petty, McGee, & Cavender, 1984).

The present article addresses the methodological issue of testing moderator variable hypothesis meta-analytically. More specifically, we propose and illustrate a three-step procedure that is based on Hedges and Olkin's (1985) meta-analytic approach, yet it incorporates study-level corrections for methodological and statistical artifacts that are typically advocated and used within psychometric approaches to meta-analysis (e.g., Hunter & Schmidt, 1990). Surprisingly, MA developers and users alike do not seem to recognize that study-level effect sizes can be corrected for methodological and statistical artifacts within the Hedges-Olkin framework.

In the sections that follow, we (a) briefly summarize the Hedges-Olkin approach to meta-analysis and the advantages of using Q homogeneity statistics, (b) briefly summarize the rationale and advantages of implementing study-level corrections for methodological and statistical artifacts in meta-analysis, (c) discuss Hedges and Olkin's consideration of corrections for methodological and statistical artifacts, (d) propose a three-step procedure to implement study-level corrections within the Hedges-Olkin approach, and (e) provide an illustration of implementing the proposed three-step procedure.

Hedges-Olkin (HO) Meta-Analytic Approach

Hedges and Olkin (Hedges, 1982a, 1982b; Hedges & Olkin, 1985) advocate an approach to MA that investigates (a) the magnitude of the relationship between two variables, (b) the variability of this relationship across studies, and (c) the (moderator) variables that determine or predict such variability. The magnitude of the relationship between two variables is investigated by obtaining an unbiased effect size estimate (d) from each study and then computing a mean d based on the study-level estimates.

The degree of variability of ds across studies is assessed with the homogeneity statistic Q. A statistically significant Q indicates that the study-level ds do not estimate a common population effect size and, therefore, the subsequent search for moderating effects is warranted. It has been argued, however, that follow-up statistical tests for moderators can be conducted when Q is statistically significant *or* large (Hedges & Olkin, 1985; Johnson & Turco, 1992). The justification for this argument is the existence of theoretical predictions regarding the effects of hypothesized moderators.

Two sets of statistical procedures are used to test for moderating effects depending upon whether the hypothesized moderator is (a) categorical (e.g., type of job: blue collar vs. white collar; Iaffaldano & Muchinsky, 1985) or (b) continuous (e.g., age; Oliver & Hyde, 1993). In tests for hypothesized *categorical* moderating effects (Hedges, 1982a; Hedges & Olkin, 1985), each study is assigned a numerical value based upon the moderator (e.g., gender, 1 = female, 2 = male) and subgrouped according to this coding scheme. The homogeneity of effect sizes within each subgroup is examined next by computing a within-subgroup homogeneity statistic Q_{W_i}, and the difference between or among *mean* within-subgroup effect sizes is assessed by computing a between-subgroup homogeneity statistic Q_B. The presence of the predicted moderator is indicated by a "complete" meta-analytic model (Johnson & Turco, 1992). That is, nonsignificant Q_{W_i}s (suggesting that all the studies within each subgroup estimate a common population effect size) and a significant Q_B (suggesting a difference between the *mean* effect size estimates across subgroups) indicate the presence of a statistically significant categorical moderator variable. The concurrent presence of a significant Q_B and significant Q_{W_i}s suggests that additional moderators might exist.

In tests for hypothesized *continuous* moderating effects, weighted least squares (WLS) regression is used (Hedges, 1982b; Hedges & Olkin, 1985). The use of a WLS model prevents information loss due to falsely polychotomizing a truly continuous variable (Cohen, 1983). When using WLS regression, the effect size estimate (*d*) is regressed onto a continuous moderator Z (i.e., $\mathbf{d} = \mathbf{Z}\beta + \varepsilon$, where $\mathbf{d}$ represents a vector of d values, $\mathbf{Z}$ represents a vector of values for the hypothesized moderator Z, β represents a vector of regression coefficients, and ε represents a vector of residuals). The Q_R and Q_E statistics are then computed, where Q_R tests the null hypothesis that the vector of regression coefficients for the moderator equals zero (H_0: $\beta = 0$) and Q_E assesses the overall regression model fit (H_0: $\varepsilon = 0$). A statistically significant Q_R indicates variable Z is a significant moderator and a nonsignificant Q_E suggests favorable model fit.

Support for the Use of Q Homogeneity Statistics

Recent Monte Carlo investigations have concluded that moderator variable hypotheses examined using the Q statistics described above are tested without noticeable deviations from nominal Type I error rates and with adequate statistical power. For example, a Monte Carlo study indicated that (a) Type I error rates ranged from .07 to .10, and (b) power rates ranged from .63 to .91, across several values for sample size and number of studies included in a meta-analysis (Sagie & Koslowsky, 1993). Results of another Monte Carlo study, specifically addressing small sample situations, also supported the accuracy of Q statistics: (a) Type I error rates ranged from .041 to .045, and (b) power rates were acceptable (e.g., when a meta-analysis was based on five studies, power rates ranged from .709 to 1.000) (Alliger, 1995).

In sum, the HO approach allows meta-analysts to test the statistical significance of overall study-level effect size variability across studies and formally test for the presence of hypothesized moderator variables. Furthermore, recent Monte Carlo investigations have concluded that hypotheses examined via Q statistics seem to be tested without noticeable deviations from nominal Type I error rates and with adequate statistical power. Finally, an additional advantage of the HO approach is its ease of testing for both categorical and continuous moderating effects.

Implementing Study-Level Corrections in Meta-Analysis

Proponents of psychometric approaches to meta-analysis extend arguments from measurement theory to MA and contend that a substantial portion of the variability observed in an X-Y relationship across studies is the result of artifactual sources of variance (Hunter & Schmidt, 1990; Raju, Burke, Normand, & Langlois, 1991; Schmidt, 1992; Schmidt, Law, Hunter, Rothstein, Pearlman, & McDaniel, 1993). Stated differently, across-study variability in effect size estimates may be due to (a) methodological and statistical artifacts, and/or (b) moderating effects. Consequently, in order to better estimate an X-Y relationship in the population, researchers should (a) attempt to control the impact of artifacts by implementing sound research designs, and (b) correct for artifactual across-study variability by subtracting it from the total observed variance in study-level effect size estimates. The goal of implementing the corrections is not to eliminate *all* kinds of variability, but rather only the across-study variability that is caused by methodological and statistical artifacts (e.g., sampling error, measurement error in the dependent variable, range restriction; Aguinis & Whitehead, 1997).

Researchers who champion psychometric approaches to meta-analysis contend that quantitative reviews of nonexperimental (typically based on *r*s as the estimate of effect size) and experimental (typically based on *d*s as the estimate of effect size) research suffer from the same problem: across-study effect size variability is not only the result of true differences caused by moderators, but to methodological and statistical artifacts. Thus, unless the artifactual variability is removed, across-study variability may be attributed to "false" moderating effects by committing a Type I statistical error (see Hunter & Schmidt, 1990: 23-29 for an illustration).

Some researchers, however, have pointed to controversial issues regarding the implementation of corrections for artifacts. First, James and his colleagues (James, Demaree, & Mulaik, 1986; James, Demaree, Mulaik, & Ladd, 1992) have suggested that some artifacts may be correlated with substantive situational moderators (e.g., organizational climate). Thus, James and colleagues argued that by correcting for methodological artifacts, a meta-analyst may also be correcting for (i.e., partialling out) a substantive moderator variable. In such situations, researchers would incorrectly conclude that the magnitude of a situational moderating effect is zero (or near zero).

A second controversial issue regarding the corrections was presented by Murphy (1993), who demonstrated that although the implementation of corrections typically leads to smaller across-study variability, the reverse can also happen (see Johnson, Mullen, & Salas, 1995: 100, Table 5 for an illustration). There are several situations that may lead to a distribution of corrected effect sizes that has a larger variance than a distribution of uncorrected effect sizes (see Murphy, 1993, for a detailed analysis of this issue). However, the conclusion reached by Murphy should not necessarily be interpreted as being inconsistent with psychometric approaches to meta-analysis: across-study variability in effect sizes is not undesirable *per se*, it is the variability caused by *artifacts* that meta-analysts should attempt to minimize.

A third controversial issue regarding corrections for artifacts is that the application of these corrections may lead to an effect size with a negative variance. In other words, the across-study variance accounted for by methodological and statistical artifacts may be larger than 100% (e.g., Rothstein, Schmidt, Erwin, Owens, & Sparks, 1990: 179, Table 5). This phenomenon may be interpreted as evidence that (a) the methodological and statistical corrections are intercorrelated, or (b) there exists a statistical anomaly. Both of these interpretations would suggest a weakness of meta-analytic procedures incorporating corrections for artifacts. However, several researchers have argued that meta-analytic results in which artifacts explain more than 100% of the across-study variance is a common and expected phenomenon. If the across-study variance is entirely artifactual in nature, then 100% of the across-study observed variance is accounted for by artifacts. In such situations, values for artifactual variance are expected to be larger than 100% approximately 50% of the time because of second-order sampling error (Burke, 1996; Callender & Osburn, 1988; Rothstein et al., 1990).

Support for the Implementation of Corrections

In spite of the aforementioned three controversial issues, corrections for methodological and statistical artifacts such as measurement error and range restriction are rooted in a long tradition in social science methodology indicating that corrected effect sizes provide better population estimates than observed (i.e., uncorrected) effect sizes (Spearman, 1904; Thorndike, 1949: 104-105). First, regarding measurement error, it is well known that this artifact reduces the observed relationship between two variables as compared to the relationship between the two constructs underlying these measures. The goal of meta-analysis is to better understand relationships between or among constructs, and not merely relationships between or among fallible measures of such constructs (Schmidt & Hunter, 1996). If effect sizes are not corrected for measurement error, the meta-analyzed effect sizes have a systematic downward bias. In addition, differential levels of measurement error across studies artificially increase the across-study variance in effect size estimates. This variability, caused by differential measurement error and not by theoretically meaningful moderator variables, can lead researchers to falsely conclude that effect sizes vary across studies and that a moderator exists. Hunter and Schmidt (1990: 117-125) provided a detailed discussion and review of the advantages of implementing the measurement error correction in meta-analytic investigations (see also Muchinsky, 1996, for a review on the measurement error correction).

There is also a long tradition in the social sciences measurement literature regarding the need for correcting for range restriction. As early as 1903, Pearson argued that effect size estimates computed from censored or range-restricted data underestimate population effects. Meta-analysts often encounter study-level effect sizes computed from scores that comprise a smaller range than that of population scores (e.g., in personnel selection research). The effects of range restriction are twofold: (a) range restriction produces a downward bias in study-level effect sizes, and (b) differential levels of range restriction across studies increases the across-study variability in effect size estimates. Consequently, researchers are advised to implement range restriction corrections (Thorndike, 1949: 169-176; see also Ree, Carretta, Earles, & Albert, 1994). Hunter and Schmidt (1990: 125-133) described the range restriction correction in detail and provided illustrative examples.

In sum, a clear advantage of performing the corrections is that they control artifactual variability in a set of study-level population effect size estimates. More precisely, artifacts that spuriously increase across-study variability can be controlled for through statistical corrections. Consequently, the probability of committing a Type I error by attributing artifactual across-study variance to "false" moderating effects is maintained at the nominal level.

Hedges and Olkin's Consideration of Study-Level Corrections for Methodological and Statistical Artifacts

The HO meta-analytic approach allows for the correction of some types of methodological artifacts (i.e., sampling error via the use of Q statistics and measurement error in the dependent variable). More precisely, on page 135 of Hedges and Olkin (1985), Equation 39 shows that study-level *d*s can be individually corrected for measurement error in the dependent variable. Also, Equation 40 (p. 136) shows how to compute a weighted estimate of the corrected mean d (i.e., d^R), and Equation 43 (p. 137) shows how to compute the overall Q^R homogeneity statistic. However, Hedges and Olkin's (1985) equations allow *only* for the correction of *d*s based on *measurement error*. There are additional artifacts, such as range restriction and dichotomization of continuous variables, that are pervasive in research in several management fields such as human resources and organizational behavior.

Despite the fact that Hedges and Olkin (1985) suggested that study-level effect sizes can be individually corrected for measurement error, and that the HO meta-analytic framework is not opposed to the use of corrections, meta-analysts using the HO approach are seemingly unaware of this possibility. We reviewed all the meta-analyses published in *Psychological Bulletin* and *Journal of Applied Psychology* between January 1991 and January 1996. Thirty of the 55 articles in which meta-analysis was used implemented the HO approach. However, in *none* of these articles did the authors implement corrections for measurement error. Thus, it seems that although the HO approach allows for corrections based on measurement error in the dependent variable, users of the HO approach typically choose not to implement this correction or are unaware of the possibility.

In sum, the HO approach does consider correcting study-level effects sizes for artifacts. However, in addition to considering sampling error via the use of Q homogeneity statistics, this approach allows only for the correction of measurement error. Finally, although the HO framework is not averse to effect size corrections, meta-analysts using this approach typically do not implement them.

Incorporating Study-Level Corrections within the Hedges-Olkin Meta-Analytic Framework: A Three-Step Procedure

The HO approach provides Q statistics for estimating the effects of moderator variables, but provides equations for *Q*s based only on the measurement error correction. Also, users of the HO approach usually do not implement corrections for methodological and statistical artifacts. We advance a three-step procedure in order to (a) correct study-level effect size estimates for across-study variability that is due to methodological and statistical artifacts, (b) test the overall homogeneity of study-level effect size estimates after artifactual sources of variance have been removed, and (c) test the effects of hypothesized moderator variables. Each of these three steps is described next.

Step One

The across-study variance that is due to artifactual sources is removed from the total observed across-study variability. Artifactual variability can be removed by individually correcting each effect size estimate for artifactual sources of variance (Hunter & Schmidt, 1990: Chapter 3). Information regarding artifact correction values can be (a) obtained from the primary study in question, (b) estimated from previous research, (c) computed as a mean correction factor from the subset of studies in which this information is reported, or (d) computed based on regression or maximum likelihood-based techniques. Roth (1994) provided a detailed discussion and recommendations regarding techniques to overcome the problem of missing data that are readily applicable to the estimation of missing artifact correction factors.[1]

The first step in our proposed procedure consists of individually correcting each effect size estimate for artifacts that have been identified in the meta-analytic literature (Hunter & Schmidt, 1990). As an illustration of correcting a correlation coefficient for *just one* of these artifacts, a study-level effect size estimate r can be corrected for range restriction, and the unrestricted population correlation $\hat{\rho}$ can be obtained by dividing r / a, where:

$$a = \frac{u}{\sqrt{(u^2 - 1)r^2 + 1}}, \tag{1}$$

r is the observed effect size estimate, and u is the ratio of restricted (sample) to unrestricted (population) standard deviations (Hunter & Schmidt, 1990: 48).

If, instead, a meta-analyst investigates an experimental or dichotomous effect and thus cumulates ds, the artifactual variability across studies can be quantified by directly correcting the ds. Alternatively, ds can be easily converted to Pearson's product-moment correlation coefficients (rs). Thus, the study-level corrections are applied to rs. Some meta-analysis developers (Hunter & Schmidt, 1990; Rosenthal, 1991) recommend the transformation of ds to rs for ease of interpretation and because additional multivariate techniques such as partial correlation analysis, path analysis, and multiple regression can be more easily utilized if needed.

Converting ds to rs requires that ds first be converted to point-biserial correlation coefficients (r_{pb}s), and that r_{pb}s then be converted to rs because r_{pb}s underestimate rs when subgroup sample sizes are unequal. Equation 2 (Wolf, 1986: 35) can be used for transforming ds to r_{pb}s:

$$r_{pb} = \frac{d}{\sqrt{d^2 + 4}}. \tag{2}$$

Glass and Stanley (1970: 171) have suggested the following formula to convert r_{pb}s to rs:

$$r = \frac{r_{pb}\sqrt{n_1 n_2}}{uN}, \tag{3}$$

where n_1 is the sample size in subgroup 1, n_2 is the sample size in subgroup 2, $N = n_1 + n_2$, and u is the ordinate (i.e., height) of the unit normal distribution at the point above which lies $100 \cdot (n_1/N)\%$ of the area under the curve.

As the last part of this step, each corrected study-level r needs to be converted back to the original d metric using Equation 4 (Wolf, 1986: 35):

$$d = \frac{2r}{\sqrt{1 - r^2}}. \tag{4}$$

At this point, all study-level effect size estimates (rs or ds) are now expressed in the d metric and have been individually corrected for methodological and statistical artifacts (e.g., range restriction, measurement error, dichotomization of continuous variables). It is these corrected study-level ds that are used in Step Two.

Step Two

The corrected study-level ds are tested for homogeneity using a *modified Q* statistic (the original equation is presented by Hedges & Olkin, 1985: 123). Q approximates a chi-square distribution with k-1 degrees of freedom, where k is the number of studies:

$$Q = \sum \frac{(d_i - d_+)^2}{\hat{\sigma}'^2(d_i)}, \tag{5}$$

where d_i is the corrected (for artifacts) and adjusted (for sample size bias by multiplying the estimate by 1 - [3 / {4N - 9}]; Hedges & Olkin, 1985: 81) study-level effect size estimate for study i, d_+ is the overall corrected and adjusted mean effect size estimate, and $\hat{\sigma}'^2(d_i)$ is the estimated *corrected* effect size variance for d_i. Adjusted, or unbiased (Johnson, 1989), effect sizes are used because they are more precise estimates of the population effect size. Furthermore, adjusted estimates minimize the across-study effect size variance (Hedges & Olkin, 1985).

Note that Equation 5's $\hat{\sigma}'^2(d_i)$ differs from $\hat{\sigma}^2(d_i)$, the *uncorrected* variance typically used in meta-analyses adopting the HO approach. The uncorrected sampling error variance is (cf. Hedges & Olkin, 1985: 86):

$$\hat{\sigma}^2(d_i) = \frac{n_1 + n_2}{n_1 n_2} + \frac{d_i^2}{2(n_1 + n_2)}, \tag{6}$$

where n_1 is the sample size in the experimental group and n_2 is the sample size in the control group. (When a d is transformed from a Pearson r and, therefore, there is only one sample with size N, it is assumed that $N = n_1 + n_2$ and $n_1 = n_2$; Johnson, 1989).

Once the effect size estimates are corrected for artifacts, their sampling error variances change. To continue with the example set forth in Step One, let $u_1,\ldots, u_k$ be the range restriction correction factors (Hunter & Schmidt, 1990: 254; see also Equation 1) for each of k studies. The *corrected* sampling error variances are (Hedges, personal communication, February 14, 1995):

$$\hat{\sigma}'^2_1 = u_1^2\hat{\sigma}_1^2, \quad \hat{\sigma}'^2_2 = u_2^2\hat{\sigma}_2^2, \quad \ldots, \quad \hat{\sigma}'^2_k = u_k^2\,\hat{\sigma}_k^2 . \tag{7}$$

If needed, Equation 7 can be expanded to correct sampling error variances for additional artifacts such as measurement error in the dependent variable (Hedges, personal communication, February 14, 1995):

$$\hat{\sigma}'^2_1 = u_1^2 w_1^2\hat{\sigma}_1^2, \quad \hat{\sigma}'^2_2 = u_2^2 w_2^2\hat{\sigma}_2^2, \quad \ldots, \quad \hat{\sigma}'^2_k = u_k^2 w_k^2\hat{\sigma}_k^2 , \tag{8}$$

where w is the measurement error correction factor ($w = 1 / [r_{yy}]^{1/2}$).

Once the ds are corrected in Step One, and Equation 7 or 8 is used to compute the corrected sampling error variances, all the usual HO meta-analytic procedures using homogeneity statistics and regression models are valid (Hedges, personal communication, February 14, 1995). The significance and magnitude of Q computed using Equation 5 is evaluated in order to determine whether the overall mean d estimates a common population effect size. A significant Q suggests the presence of unexplained effect size variability and, therefore, the testing of theoretically relevant moderator variables is warranted. Moreover, if theory predicts a particular moderating effect, it has been suggested that a statistically nonsignificant yet large Q can be followed up by moderator variable testing (Hedges & Olkin, 1985; Johnson & Turco, 1992) as described next in Step Three.

Step Three

HO's fixed effects models approach is implemented to test for the presence of hypothesized moderator variables using the corrected and adjusted study-level ds and the corrected sampling error variances. As described next, categorical or continuous models are fitted to the effect size data depending upon the nature of the hypothesized moderator(s).

For tests of categorical models, Equations 9 and 11 show modified formulae for computing Q_B and Q_{Wi} (the original equations are presented by Hedges & Olkin, 1985: 154-155). Q_B approximates a chi-square distribution with p-1 degrees of freedom, where p is the number of classes or levels of the hypothesized moderator:

$$Q_B = \sum \frac{(d_{i+} - d_{++})^2}{\hat{\sigma}'^2(d_{i+})} , \tag{9}$$

where d_{i+} is the corrected and adjusted mean effect size estimate for the *i*th class of the hypothesized moderator, d_{++} is the adjusted grand mean of all *k* corrected study-level effect size estimates, and $\hat{\sigma}'^2(d_{i+})$ is the corrected effect size variance for d_{i+}. Note that this corrected within-class variance differs from the uncorrected variance (provided by Hedges & Olkin, 1985: 152) because it cumulates corrected within-class study-level variances:

$$\hat{\sigma}'^2(d_{i+}) = \left(\sum \frac{1}{\hat{\sigma}'^2(d_{ij})}\right)^{-1} \tag{10}$$

Q_{W_i} approximates a chi-square distribution with *m*-1 degrees of freedom, where *m* is the number of effect sizes within a given class or level of the hypothesized moderator:

$$Qw_i = \sum \frac{(d_{ij} - d_{i+})^2}{\hat{\sigma}'^2(d_{ij})}, \tag{11}$$

where d_{ij} is the corrected study-level effect size estimate for the *i*th class of the hypothesized moderator and the *j*th study, d_{i+} is the corrected and adjusted mean effect size estimate for the *i*th class of the hypothesized moderator, and $\hat{\sigma}'^2(d_{ij})$ is the corrected effect size variance for d_{ij}. Equation 11 is used for each class, level, or category of the hypothesized moderator.

For tests of continuous models, Equations 12 and 13 show the formulae for Q_R and Q_E (cf. Hedges & Olkin, 1985: 169-172). Q_R approximates a chi-square distribution with *l* degrees of freedom, where *l* is the number of regression coefficients in the vector β:

$$Q_R = \hat{\beta}' \hat{\Sigma}_{\beta}^{-1} \hat{\beta}, \tag{12}$$

where $\hat{\beta}'$ is the transpose of the modified generalized least squares estimator of the vector β and $\hat{\Sigma}_{\beta}$ is an estimate of the covariance matrix Σ_{β}.

Q_E approximates a chi-square distribution with *k* - *p* - 1 degrees of freedom, where *k* is the number of studies and *p* is the number of predictor variables:

$$Q_E = d' \hat{\Sigma}_d^{-1} d - Q_R, \tag{13}$$

where $\mathbf{d}'$ is the transpose of a vector of corrected study-level *d* values, $\hat{\Sigma}_d$ is an estimate of the diagonal covariance matrix Σ_d, and Q_R is the value obtained using Equation 12.

An Illustration of Implementing the Three-Step Procedure

To illustrate how the proposed three-step meta-analytic procedure can be easily implemented, we arbitrarily selected a data set from Hunter and Schmidt

Table 1. An Illustration of the Proposed Three-Step Meta-Analytic Procedure Using Data from Hunter and Schmidt (1990: 24, Table 1.1)

Study	*N*	*r*	*Sex*	*a*	$\hat{\rho}$	*d*	d_i	$(d_i-d_+)^2$	$\hat{\sigma}^2(d_i)$	$\hat{\sigma}'^2(d_i)$	$(d_{ij}-d_{i+})^2$
(1)	20	.46	F	.65	.71	2.03	1.95	.0529	.29478	.10612	.0036
(2)	72	.32	M	.62	.52	1.20	1.19	.2809	.06540	.02354	.0064
(3)	29	.10	M	.60	.17	.34	.33	1.9321	.13978	.05032	.8836
(4)	30	.45	M	.64	.70	1.96	1.91	.0361	.19387	.06979	.4096
(5)	71	.18	F	.61	.30	.62	.61	1.2321	.05900	.02124	1.9600
(6)	62	.45	F	.64	.70	1.96	1.93	.0441	.09469	.03409	.0064
(7)	25	.56	F	.67	.83	3.03	2.93	1.4641	.33150	.11934	.8464
(8)	46	.41	M	.64	.65	1.69	1.66	.0036	.11696	.04211	.1521
(9)	22	.55	F	.67	.82	2.90	2.79	1.1449	.35865	.12911	.6084
(10)	69	.44	F	.64	.69	1.89	1.87	.0225	.08322	.02996	.0196

Notes:
N = total sample size
r = Pearson's product-moment correlation coefficient between organizational commitment and job satisfaction
F = female
M = male
a = range restriction correction factor computed using Equation 1
$\hat{\rho}$ = r / a, corrected (for range restriction) correlation coefficient
d = transformation of $\hat{\rho}$ to d using Equation 4
d_i = corrected (for artifacts) and adjusted $d = d \cdot (1 - [3 / \{4N - 9\}])$
d_+ = grand mean = mean of all d_is = 1.72 (r = .65) in this example
$(d_i-d_+)^2$ = squared deviations between each study-level effect size and the grand mean
$\hat{\sigma}^2(d_i)$ = uncorrected sampling error variance computed using Equation 6
$\hat{\sigma}'^2(d_i)$ = corrected sampling error variance computed using Equation 7
$(d_{ij} - d_{i+})^2$ = squared deviations between each study-level effect size and its group mean

(1990: 24, Table 1.1, Studies 1-10). Table 1 shows a hypothetical set of 10 studies that independently assessed the relationship between organizational commitment and job satisfaction. For each of the ten studies, information is available regarding (a) total sample size (N), (b) Pearson correlation coefficient (r) between organizational commitment and job satisfaction, and (c) sex of the study participants (male vs. female). Assume that the goal of this research is to test a theory-based prediction that sex is a moderator of the organizational commitment-job satisfaction link, such that this relationship is stronger for women than for men. Next, we turn to the implementation of our three-step meta-analytic procedure. These calculations can be performed using a spreadsheet computer program or even a pocket calculator if the number of studies is manageable.

Step One

The first step is to correct the study-level effect size estimates for artifactual sources of variance. To be consistent with the example provided in the conceptual description of the three steps above, we corrected the study-level rs for range restriction so that an unrestricted population correlation $\hat{\rho}$ could be obtained for each study ($\hat{\rho} = r / a$, where a is defined in Equation 1). For the sake of simplicity, assume that u (ratio of restricted to unrestricted standard deviations) is .60 for all ten studies.[2] Table 1 shows the values for each study's a, computed using Equation 1, and the values for each study's $\hat{\rho}$, the corrected (for range restriction) correlation coefficient. Next, we transformed each corrected correlation coefficient into a d using Equation 4. Note that the conversion to the d metric is necessary so as to compute the Q statistics in Step Two and Step Three. Nevertheless, to be consistent with the original metric and to ease the interpretation of results, ds can be converted back to rs at any point during the procedures. At the end of Step One, the effect size estimates are expressed in the d metric and have been individually corrected for methodological and statistical artifacts. (We illustrate the procedures correcting only for range restriction, but the corrections can be extended to include additional artifacts such as measurement error, dichotomization of continuous variables, and so forth).

Step Two

As suggested by the HO approach, the corrected ds are tested for overall homogeneity using the modified Q statistic shown in Equation 5. This step requires that each corrected d first be converted to a d_i by multiplying each d by $1 - [3 / (4N - 9)]$ (i.e., $d_i = d \cdot \{1 - [3 / (4N - 9)]\}$). As expected, greater differences between d (corrected and unadjusted) and d_i (corrected and adjusted) are observed for studies having smaller sample sizes.

Following the computation of d_is, we obtained the corrected adjusted mean d_+, which is simply the mean of all the d_is (i.e., the corrected and adjusted ds; $d_+ = 1.72$, or converted to r metric, using Eq. 2, $r = .65$). Next, we computed the numerator for Equation 5; namely the squared deviations between each study-level d_i and the mean d_+ ($[d_i - d_+]^2$). We then obtained values for the denominator in Equation 5; namely the corrected sampling error variances ($\hat{\sigma}'(d_i)$). We obtained these values by first computing the uncorrected variances using Equation 6 for each d_i and then using Equation 7 to convert each uncorrected variance $\hat{\sigma}(d_i)$ to a corrected variance $\hat{\sigma}'(d_i)$. In this particular example, the original effect size estimates were rs, and therefore there is only one sample with size N. Thus, consistent with the HO approach, in Equation 6 we assumed that $N = n_1 + n_2$ and $n_1 = n_2$ (Johnson, 1989; Johnson & Eagly, in press).

Finally, the Q statistic is computed as indicated in Equation 5, $Q(9) = 132.15$, $p < .01$. Thus, the conclusion from this analysis is that the corrected effect size estimates for the organizational commitment-job satisfaction relationship are heterogeneous across the 10 studies. Thus, we proceed to Step Three where we test the hypothesis that the across-study variability in the organizational commitment-job satisfaction link can be explained by the moderating effect of sex.

Step Three

Step Three consists of computing the Q_B and Q_{Wi} statistics using Equations 9 and 11, respectively. All the information needed to compute these statistics is displayed in Table 1.

Computation of Q_B. We used Equation 9 to compute Q_B and, thus, first obtained a corrected and adjusted mean effect size (d_{i+}) based on the d_is within each of the two classes of the moderator (i.e., males, females; there are six female and four male samples). For males $d_{1+} = 1.27$ ($r = .54$, cf. Eq. 2) and for females $d_{2+} = 2.01$ ($r = .71$, cf. Eq. 2). We then obtained the squared deviations between the mean effect size in each class and the grand mean of all k corrected study-level effect size estimates ($[d_{i+} - d_{++}]^2$, where $d_{++} = 1.72$). This resulted in a squared deviation of .2025 for males and .0841 for females and provides the information needed for the numerator in Equation 9.

For the denominator in Equation 9, we computed the corrected effect size variance for the male ($\hat{\sigma}'^2 d_{1+}$) and female ($\hat{\sigma}'^2 d_{2+}$) classes using Equation 10. This is easily achieved because each within-class corrected effect size variance consists of an aggregation of the variances for the effect size estimates within each class ($\hat{\sigma}'^2(d_i)$), which is information obtained during Step One. The variance for males is .00996 and the variance for females is .00739. Then, given the values for $[d_{i+} - d_{++}]^2$ and $\hat{\sigma}'^2 d_{i+}$, $Q_B(1) = 32.54$, $p < .01$. Consequently, we conclude that there is a statistically significant moderating effect of sex on the organizational commitment-job satisfaction relationship such that the relationship is stronger for females ($r = .71$) than for males ($r = .54$).

Computation of Q_{W_i}. To compute Q_{W_i} for each of the two classes, we first obtained the squared deviations between the study-level estimates in each class and the class mean ($[d_{ij} - d_{i+}]^2$). The individual effect size estimates are displayed in Table 1 (i.e., d_is within each class), and the class means were computed above (for males $d_{1+} = 1.27$ and for females $d_{2+} = 2.01$). The information needed for the denominator is also displayed in Table 1: $\hat{\sigma}'^2(d_{ij})$ is the corrected effect size variance for the studies included in each class (i.e., $\hat{\sigma}'^2(d_i)$s). Then, all that remains is to compute Q_{W_i} based on $[d_{ij} - d_{i+}]^2$ and $\hat{\sigma}'^2(d_{ij})$ for each study. For the male class $Q_{W_1}(3) = 59.82$, $p < .01$, and for the female class $Q_{W_2}(5) = 80.58$, $p < .01$.

In sum, based on these results, we conclude that there is overall heterogeneity across the ten hypothetical studies that examined the relationship between organizational commitment and job satisfaction. Thus, moderators seem to be affecting this relationship. Moreover, we determined that there was heterogeneity *between* classes, as well as *within* each class of the hypothesized moderator sex. These results suggest that (a) there are moderators of the examined organizational commitment-job satisfaction relationship, (b) sex of the sample is one of these moderators, and (c) because of the presence of within-class heterogeneity, there may be other variables, in addition to sex, that are related to the magnitude of the organizational commitment-job satisfaction relationship.

Summary and Conclusions

As stated by Hall and Rosenthal (1991: 447), "the search for moderator variables is not only an exciting intellectual enterprise but, indeed,...it is at the very heart of the scientific enterprise." Given the advancement in management theory over the past few decades, together with the increasingly pervasive postulation of theoretical models including complex moderated relationships, it is not surprising that the interest in testing moderating effects meta-analytically has increased substantially.

The present article favors the implementation of study-level corrections for methodological and statistical artifacts within the HO meta-analytic approach. Future Monte Carlo research is needed to ascertain the impact of incorporating such corrections within the HO approach and, specifically, to investigate the statistical properties (i.e., empirically-derived Type I error and power rates) of the resulting modified Q statistics. Nevertheless, there is analytic (e.g., Pearson, 1903) and Monte Carlo (e.g., Alliger, 1995; Sagie & Koslowsky, 1993) empirical evidence regarding (a) the advantages of implementing corrections of study-level effect sizes, and (b) the accuracy of Q homogeneity statistics based on uncorrected effect sizes. Consequently, we recommend that corrections for artifactual across-study variability be implemented. We also recommend that the proposed modified Q statistics, which incorporate *corrected* sampling error variances, be computed. Using Q statistics enables researchers to explain across-study effect-size variability by formally testing moderator variable hypotheses.

Acknowledgment: A previous version of this article was presented at the meeting of the American Psychological Society, Washington, DC, July 1994. Both authors contributed equally to this research. We thank Larry V. Hedges for his suggestions and encouragement regarding this article, and two *Journal of Management* reviewers for providing helpful comments on previous drafts.

Notes

1. Hunter and Schmidt have suggested additional algorithms, which do not have an analogue in the HO approach, to generate across-study variance due to assumed differences in artifact values (e.g., measurement error and range restriction, Hunter & Schmidt, 1990: 158-198). Although these procedures can be used when it is not possible to obtain or estimate information regarding study-level artifact values, they do *not* lead to individually corrected study-level effect sizes. Instead, these algorithms result in an aggregate-level variance-due-to-artifact estimate. Thus, these aggregate-level procedures are not useful in the present context because individually corrected effect sizes are needed to compute Q statistics as described in Steps Two and Three.
2. By assuming the same degree of range restriction across the ten studies, the across-study variability does not decrease by implementing the range restriction correction. Moreover, because the magnitude of the corrected effect sizes does increase, the variability of corrected effect sizes may be larger than the variability of their uncorrected counterparts (Murphy, 1993). Nevertheless, we chose to use a constant value for range restriction across studies to ease the interpretation of the computations involved in the illustration.

References

Aguinis, H. 1995. Statistical power problems with moderated multiple regression in management research. *Journal of Management*, 21: 1141–1158.

Aguinis, H., Bommer, W. H., & Pierce, C. A. 1996. Improving the estimation of moderating effects by using computer-administered questionnaires. *Educational and Psychological Measurement*, 56: 1043–1047.

Aguinis, H., Pierce, C. A. 1998. Heterogeneity of error variance and the assessment of moderating effects of categorical variables: A conceptual review. *Organizational Research Methods*, 1: 296–314.

Aguinis, H., Pierce, C. A., & Quigley, B. M. 1993. Conditions under which a bogus pipeline procedure enhances the validity of self-reported cigarette smoking: A meta-analytic review. *Journal of Applied Social Psychology*, 23: 352–373.

Aguinis, H., Pierce, C. A., & Quigley, B. M. 1995. Enhancing the validity of self-reported alcohol and marijuana consumption using a bogus pipeline procedure: A meta-analytic review. *Basic and Applied Social Psychology*, 16: 515–527.

Aguinis, H., Pierce, C. A., & Stone-Romero, E. F. 1994. Estimating the power to detect dichotomous moderators with moderated multiple regression. *Educational and Psychological Measurement*, 54: 690–692.

Aguinis, H., & Stone-Romero, E. F. 1997. Methodological artifacts in moderated multiple regression and their effects on statistical power. *Journal of Applied Psychology*, 82: 192–206.

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.

Alliger, G. M. 1995. The small sample performance of four tests of the difference between pairs of meta-analytically derived effect sizes. *Journal of Management*, 21: 789–799.

Burke, M. J. 1996. An empirical examination of the effect of second-order sampling error on selection test validity coefficients. In M. J. Burke (Chair), *Quantitative issues in personnel selection research and practice*. Symposium conducted at the meeting of the Society for Industrial and Organizational Psychology, San Diego, CA.

Callender, J. C., & Osburn, H. G. 1988. Unbiased estimation of sampling variance of correlations. *Journal of Applied Psychology*, 73: 312–315.

Cohen, J. 1983. The cost of dichotomization. *Applied Psychological Measurement*, 7: 249–253.

Cooper H. M., & Hedges, L. V. (Eds.) 1994. *The handbook of research synthesis*. New York: Russell Sage Foundation.

Cooper, H. M., & Lemke, K. M. 1991. On the role of meta-analysis in personality and social psychology. *Personality and Social Psychology Bulletin*, 17: 245–251.

Cotton, J. L., & Tuttle, J. M. 1986. Employee turnover: A meta-analysis and review with implications for research. *Academy of Management Review*, 11: 55–70.

Dobbins, G. H., & Platz, S. J. 1986. Sex differences in leadership: How real are they? *Academy of Management Review*, 11: 118–127.

Glass, G. V., & Stanley, J. C. 1970. *Statistical methods in education and psychology*. Englewood Cliffs, NJ: Prentice-Hall.

Hall, J. A., & Rosenthal, R. 1991. Testing for moderator variables in meta-analysis: Issues and methods. *Communication Monographs*, 58: 437–448.

Hedges, L. V. 1982a. Fitting categorical models to effect sizes from a series of experiments. *Journal of Educational Statistics*, 7: 119–137.

Hedges, L. V. 1982b. Fitting continuous models to effect size data. *Journal of Educational Statistics*, 7: 245–270.

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

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

Iaffaldano, M. T., & Muchinsky, P. M. 1985. Job satisfaction and job performance: A meta-analysis. *Psychological Bulletin*, 97: 251–273.

James, L. R., Demaree, R. G., & Mulaik, S. A. 1986. A note on validity generalization procedures. *Journal of Applied Psychology*, 71: 440–450.

James, L. R., Demaree, R. G., Mulaik, S. A, & Ladd, R. T. 1992. Validity generalization in the context of situational models. *Journal of Applied Psychology*, 77: 3–14.

Johnson, B. T. 1989. *DSTAT: Software for the meta-analytic review of research literatures*. Hillsdale, NJ: Erlbaum.

Johnson, B. T., & Eagly, A. H. in press. Quantitative synthesis of social psychological research. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social psychology*. London: Cambridge University Press.

Johnson, B. T., Mullen, B., & Salas, E. 1995. Comparison of three major meta-analytic approaches. *Journal of Applied Psychology*, 80: 94–106.

Johnson, B. T., & Turco, R. M. 1992. The value of goodness-of-fit indices in meta-analysis: A comment on Hall and Rosenthal. *Communication Monographs*, 59: 388–396.

Muchinsky, P. M. 1996. The correction for attenuation. *Educational and Psychological Measurement*, 56: 63–75.

Mullen, B., Salas, E., & Miller, N. 1991. Using meta-analysis to test theoretical hypotheses in social psychology. *Personality and Social Psychology Bulletin*, 17: 258–264.

Murphy, K. R. 1993. The situational specificity of validities: Correcting for statistical artifacts does not always reduce the trans-situational variability of correlation coefficients. *International Journal of Selection and Assessment*, 1: 158–162.

Oliver, M. B., & Hyde, J. S. 1993. Gender differences in sexuality: A meta-analysis. *Psychological Bulletin*, 114: 29–51.

Pearson, K. 1903. Mathematical contributions to the theory of evolution: II. On the influence of natural selection on the variability and correlation of organs. *Royal Society Philosophical Transactions*, 200 (Series A): 1–66.

Petty, M. M., McGee, G. W., & Cavender, J. W. 1984. A meta-analysis of the relationships between individual job satisfaction and individual performance. *Academy of Management Review*. 9: 712–721.

Raju, N. S., Burke, M. J., Normand, J., & Langlois, G. M. 1991. A new meta-analytic approach. *Journal of Applied Psychology,* 76: 432–446.

Ree, M. J., Carretta, T. R., Earles, J. A., & Albert, W. 1994. Sign changes when correcting for range restriction: A note on Pearson's and Lawley's selection formulas. *Journal of Applied Psychology,* 79: 298–301.

Rosenthal, R. 1991. *Meta-analytic procedures for social research* (rev. ed.). Beverly Hills, CA: Sage.

Roth, P. L. 1994. Missing data: A conceptual review for applied psychologists. *Personnel Psychology,* 47: 537–560.

Rothstein, H. R., Schmidt, F. L., Erwin, F. W., Owens, W. A., & Sparks, C. P. 1990. Biographical data in employment selection: Can validities be made generalizable? *Journal of Applied Psychology,* 75: 175–184.

Sagie, A., & Koslowsky, M. 1993. Detecting moderators with meta-analysis: An evaluation and comparison of techniques. *Personnel Psychology,* 46: 629–640.

Schmidt, F. L. 1992. What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist,* 47: 1173–1181.

Schmidt, F. L., & Hunter, J. E. 1996. Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods,* 1: 199–223.

Schmidt, F. L., Law, K., Hunter, J. E., Rothstein, H. R., Pearlman, K., & McDaniel, M. 1993. Refinements in validity generalization methods: Implications for the situational specificity hypothesis. *Journal of Applied Psychology,* 78: 3–12.

Spearman, C. 1904. The proof and measurement of association between two things. *American Journal of Psychology,* 15: 72–101.

Stone-Romero, E. F., Alliger, G. M., & Aguinis, H. 1994. Type II error problems in the use of moderated multiple regression for the detection of moderating effects of dichotomous variables. *Journal of Management,* 20: 167–178.

Thorndike, R. L. 1949. *Personnel selection.* New York: Wiley.

Wolf, F. M. 1986. *Meta-analysis: Quantitative methods for research synthesis.* Newbury Park, CA: Sage.

Zedeck, S. 1971. Problems with the use of "moderator" variables. *Psychological Bulletin,* 76: 295–310.