Fitting Parametric Fixed Effect Categorical Models to Effect Sizes: A Neglected Meta-Analytic Approach in Organizational Studies

ALEXANDER D. STAJKOVIC
University of California, Irvine

Hedges and Olkin's meta-analytic procedures for fitting parametric fixed effect categorical models to effect sizes have been widely used in research in experimental, social, and educational psychology. However, to date, this meta-analytic approach has been largely absent from the management literature in general and organizational behavior literature in particular. This article summarizes Hedges and Olkin's meta-analytic method by (a) outlining the specific step-by-step analytical procedures, (b) using a simplified notational system, and (c) providing methodological and conceptual rationale for each analysis. The use of this approach can add considerable methodological power and provide a relevant meta-analytic alternative to the study of organizational behavior and management.

About 20 years ago, Gene Glass introduced the term *meta-analysis* for the first time. He defined meta-analysis as "the statistical analysis of a large collection of analysis results from individual studies for the purposes of integrating the findings" (1976a). Since 1976, meta-analysis has been widely used as a new approach to research review that involves explicit use of quantitative methods to combine the results of individual studies (Hedges, 1986). Although methodologists have outlined a number of specific methods for conducting meta-analysis (e.g., Cooper, 1984; Friske & Treines, 1985; Light, 1983; Light & Pillemer, 1984; Rosenthal, 1980, 1984; Rosenthal & Rubin, 1978, 1988; Yeaton & Wortman, 1984), the most widely recognized meta-analytic procedures have been offered by Rosenthal (1991a, 1991b); Glass, McGaw, and Smith (1981); Hunter, Schmidt, and Jackson (1982); Hunter and Schmidt (1990, 1995); Hedges & Olkin (1985); and Hedges (1986).

Being the oldest of what we now define as meta-analytic techniques (Rosenthal, 1961), Rosenthal's (1991a, 1991b) approach provides procedures that further extend and enhance some of the basic statistical methods previously developed (e.g., Fisher, 1932; Pearson, 1933). The Glass et al. (1981) meta-analytic method was also introduced relatively early; however, given the lack of subsequent methodological and conceptual enhancements, the use of this approach seems to have decreased in the

Organizational Research Methods, Vol. 2 No. 1, January 1999 88-102

recent years. This currently leaves the Hunter et al. (1982) and the Hedges & Olkin (1985) procedures as the two most comprehensive and commonly used approaches for meta-analytic research (Johnson, Mullen, & Salas, 1995). Although both are widely used, the frequency of application of these two meta-analytic methods differs depending on the field of inquiry. Since its introduction, Hedges and Olkin's (1985) method has almost exclusively been used in experimental, social-psychological, and educational research in general but has largely been absent as the meta-analytic approach used in organizational research.

For example, an examination of the 30 most recent meta-analyses published in journals commonly used in organizational behavior and management research (*Academy of Management Journal, Journal of Applied Psychology*) found that 21 (70%) used the Hunter et al. (1982) meta-analytic method, or the recent refinements of the original procedures (e.g., Hunter & Schmidt, 1990, 1995), whereas only 6 (20%) used the Hedges and Olkin (1985) method. On the other hand, the review of the 30 most recent meta-analyses published in the journals more closely related to research in experimental, social, and educational psychology (*Psychological Bulletin, Review of Educational Research*) revealed that 23 (77%) studies used Hedges and Olkin's (1985) meta-analytic method and only 2 (7%) applied the Hunter et al. (1982) or Hunter and Schmidt (1990, 1995) procedures. This trend gets even more pronounced for the period of the last 10 years.[1]

Thus, despite the recognized methodological strength (see Cooper & Hedges, 1994; Hedges, 1986; Hedges & Olkin, 1985; Rosenthal, 1984, 1991a, 1991b), and its wide applicability in related fields, Hedges and Olkin's (1985) meta-analytic method seems to be generally overlooked as a meta-analytic alternative for the research synthesis in the field of organizational behavior and management. To date, no methodological or conceptual arguments have been offered as to why the Hedges and Olkin meta-analytic method might not be suitable for application to the field of organizational behavior and management. In fact, recent empirical comparison of the major meta-analytic approaches (Johnson et al., 1995) suggested that the Hedges and Olkin method tends to yield (a) greater statistical power to detect significant effect sizes and (b) more accurate and predictable statistical estimates in terms of central tendency, variability, and moderator analysis, when compared to the other widely used method provided by Hunter et al. (1982), or Hunter and Schmidt (1990, 1995).

The purpose of this article is not to deny or criticize the use of the Hunter et al. (1982) or Hunter and Schmidt (1990, 1995) meta-analytic method in organizational behavior research, but rather to propose that the more frequent use of the Hedges and Olkin (1985) meta-analytic procedures can provide a relevant and powerful methodological alternative to research synthesis in the organizational sciences. In light of this purpose, I outline step by step Hedges and Olkin's procedures for (a) primary meta-analysis, (b) moderator meta-analysis for categorical models, and (c) orthogonal comparisons among classes. Each of these procedures is accompanied by methodological and conceptual rationale underlying the analyses involved. I also present the modified notational system, largely provided by Hedges (1986), which represents a simpler version (without changing the content of analyses) of the notational system used by Hedges and Olkin.

Primary Meta-Analysis

Calculating Single Effect Sizes

In that meta-analysis represents the quantitative summary of individual study findings across an entire body of research (Cooper & Hedges, 1994), the first step is to compute the index of effect magnitude, or effect size, for each study. A given effect size indicates at what specific percentile in the control group an average person in the treatment group would be. For example, "an effect size of +1 indicates that a person at the mean of the control group would be expected to rise to the 84th percentile of the control group after treatment" (Glass, 1976b, p. 13). According to the Hedges and Olkin approach, effect size for every study is computed in the form of index g, which represents the mean difference between the experimental and control groups divided by the pooled standard deviation (Hedges, 1986; Hedges & Olkin, 1985).

$$g = \frac{(\bar{x}_e - \bar{x}_c)}{s} \tag{1}$$

where the pooled standard deviation s is defined as

$$s = \sqrt{\frac{(n^e - 1)(s^e)^2 + (n^c - 1)(s^c)^2}{n^e + n^c - 2}} \tag{2}$$

and where, n^e and n^c represent the experimental and control group sample sizes, and s^e and s^c represent the experimental and control group standard deviations, respectively.

Three issues need to be further clarified when index g is used to compute the effect size magnitude. First, many times, unfortunately, research reports do not provide descriptive statistics necessary to compute effect size g as outlined in Equation 1. In the case that research studies do not report statistical estimates (e.g., $\bar{x}_s$, sd_s) needed to directly calculate effect size g, but they report some other statistic (e.g., t, F, χ^2), Hedges (1981, 1982a) and Rosenthal (1991a, 1991b, 1994) provided computational adjustments necessary to transform different statistical indexes to effect size g. A list of commonly used transformational formulas for these computational adjustments is provided in the appendix.

Second, a possible alternative to pooled standard deviation, in the denominator of Equation 1, might be the control group standard deviation as suggested by Glass et al. (1981). However, it has been largely accepted that pooled standard deviation is a preferable solution because of its smaller sampling error (Hedges & Olkin, 1985; Nouri & Greenberg, 1995).

Third, for small samples ($n < 10$), g has a slight tendency to overestimate population effect size δ (Hedges & Olkin, 1985). To correct for this bias, effect size estimate g should be multiplied by the correction factor, provided by Hedges (1981), which depends on the sample size of the study. This correction, analogous to the sample estimate of variance, gives an unbiased effect size estimator (d). Because d provides

a more stable and accurate estimate of the population effect size δ, and is easy to calculate, its use in subsequent analyses is suggested (Hedges, 1982a, 1982b, 1986; Hedges & Olkin, 1985).

$$d = g\left(1 - \left(\frac{3}{(4n^e + 4n^c - 9)}\right)\right) \tag{3}$$

Hedges (1981) showed that unbiased estimator d for every g has an approximately normal sampling distribution when all studies share common effect size with mean population effect size δ and variance (v).[2] Variance (v) of the unbiased effect size (d) is determined by the sample sizes of both experimental and control groups and the value of unbiased effect size (d).

$$v = \left(\frac{n^e + n^c}{n^e n^c}\right) + \left(\frac{d^2}{2(n^e + n^c)}\right) \tag{4}$$

The possibility to calculate the nonsystematic variance v for every d observation is the major advantage of the Hedges and Olkin (1985) and Hedges (1986) approach because it allows for estimating the systematic effects by using degrees of freedom among different d values while at the same time allowing for estimating nonsystematic variance necessary for statistical tests.

Outlier Analysis

SAMPLE SIZE OUTLIERS

Before meta-analysis proceeds to combining estimates of individual effect sizes, sample size outlier analysis should be conducted first. This analysis is recommended regardless of the overall meta-analytic method used, but especially when Hedges and Olkin (1985) procedures are employed. The scenario that leads to caution is when by analyzing sample sizes of each research report that satisfied the criteria for inclusion in the meta-analysis, we discover that some studies conducted multiple case studies each involving only one subject ($n = 1$). Although the idiosyncratic value of these experiments should be recognized (see Davis & Luthans, 1984), when Hedges and Olkin's meta-analytic procedures are used, effect sizes based on single subjects should be examined in a separate analysis. There are several methodological and conceptual reasons for this suggestion.

First, because effect size estimate g tends to slightly overestimate population effect size δ for small sample sizes ($n < 10$) (Hedges & Olkin, 1985), using an unbiased estimator d, which provides a more reliable estimate of effect magnitude, is suggested (Hedges, 1986; Hedges & Olkin, 1985). By applying the correction factor provided in Equation 3, which depends on the sample size of the study, the magnitude of g is reduced in every instance of the small sample size, except when $n = 1$. In the case of $n = 1$, d actually inflates the value of the already positively biased estimator g. Second, the nomographs for exact confidence intervals for population effect size δ when the lesser of n^e or n^c is less than 10, provided by Hedges and Olkin (1985), include the

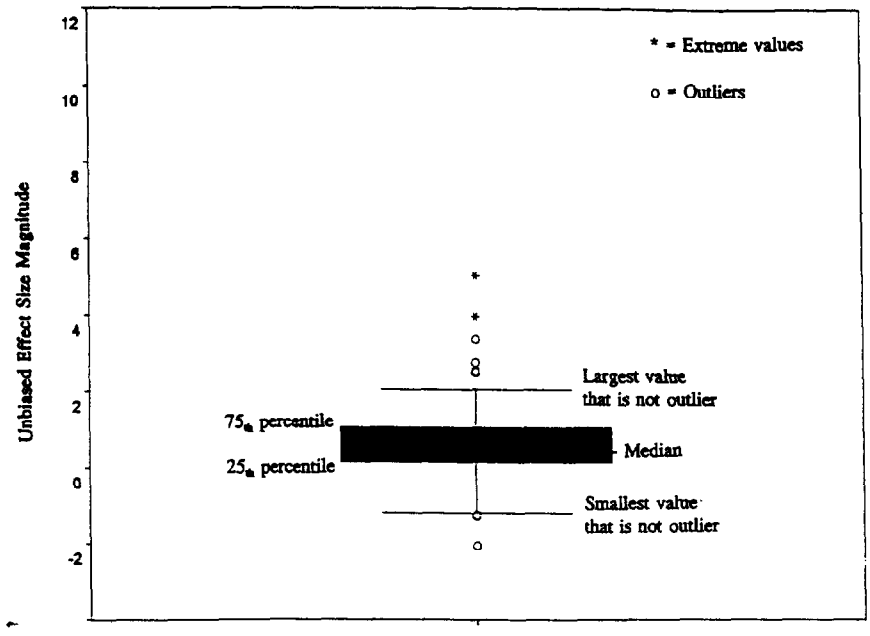

Figure 1: Schematic Analysis of Outliers and Extreme Values

values for effect size magnitude when $2 \leq n \leq 10$, thus again excluding the possibility of $n = 1$. Third, when there are studies with sample size outliers (especially as extreme as $n = 1$), the procedure of weighting the average effect size estimate $d.$ with the inverse of its variance will be unduly biased considering the relatively greater sample estimate of variance based on single subjects (see also Osburn & Callender, 1992).[3] Finally, the strong possibility for capitalization on chance precludes a reliable generalization of the findings based on single subjects in terms of external validity (see Cook & Campbell, 1977).

EFFECT SIZE OUTLIERS AND EXTREME VALUES

After the sample size outlier analysis, but still before combining individual estimates of effect sizes, we also need to estimate the relative stability of unbiased effect size magnitudes by determining their outliers and extreme values for the entire sample of k studies. A commonly used procedure is to conduct a schematic plot analysis (Light, Singer, & Willett, 1994) because of its accuracy (and easy use) in determining outlier and extreme values (see also Huffcutt & Arthur, 1995). An example of a schematic plot analysis is shown in Figure 1.

In a schematic plot analysis, effect magnitudes that are positioned 1.5 to 3 lengths from the upper or lower edge of the 50% interquartile range (e.g., Tukey's hinges) are considered outliers, and values that are more than 3 lengths from the interquartile range are considered extreme values (e.g., Tukey, 1977). Upper and lower "viskers" represent the highest and lowest values of unbiased effect size d that are not considered outliers. In this example, the upper boundary of effect size magnitudes corresponds to the value of 2, meaning that the average person in the control group will rise 2 standard

deviations from the mean (98th percentile) after the treatment. By having this information, we can easily see that any increase beyond this limit would not have much of a meaningful significance. Even if only a modest proportion of all effect sizes is deleted, the customary procedure for dealing with sample size outliers is to conduct two analyses (e.g., Stajkovic & Luthans, 1997, 1998; Williams & Livingstone, 1994): one with outliers and extreme values included and one with them excluded. Also, conceptual analysis of contexts and/or processes of the studies that generated effect magnitude and/or sample size outliers and extreme values should be conducted to determine whether there were any across-studies, systematically occurring common characteristics that may provide additional insight into the "deviant" data (see Behrens, 1997; Stajkovic & Luthans, 1998).

Combining Estimates of Individual Effect Sizes

The next analytical question in every meta-analysis is how to combine estimates of effect sizes. Although one way is to simply take an average of (d), the more precise procedure to determine whether all studies share the common effect size is to combine estimates of single unbiased effect sizes by calculating a weighted average effect size, labeled ($d_.$), that incorporates variances v_i to v_k for each d_i to d_k (Hedges, 1986). Thus, to compute the weighted average effect size ($d_.$) across k studies, we weight each effect size estimate (d) by the inverse of its variance.[4]

$$d_. = \frac{\sum_{i=1}^{k} (w_i d_i)}{\sum_{i=1}^{k} w_i} \tag{5}$$

where,

$$w_i = 1/v_i = \frac{2(n_i^e + n_i^c)\, n_i^e n_i^c}{2(n_i^e + n_i^c)^2 + n_i^e n_i^c d_i^2} \tag{6}$$

If all k studies share a common population effect size, the weighted average effect size $d_.$ has approximate normal distribution with the mean of δ and variance of $v_.$, where,

$$v_. = \frac{1}{\sum_{i=1}^{k} w_i} \tag{7}$$

After calculating the weighted average effect size ($d_.$), and the variance of weighted average effect size ($v_.$), the hypothesis that the common population effect size $\delta = 0$ is tested to ensure that the average effect size ($d_.$) across all k studies occurred beyond

chance. To do that, we can either (a) compare the ratio of $d.^2/v.^2$ to χ^2 distribution for $df = 1$ at a certain significance level (α) (Hedges, 1986) or (b) obtain, say, 95% confidence intervals (Hedges & Olkin, 1985). Either way, we intend to determine if there is a significant main effect for the average treatment across k studies.

Testing for Homogeneity of Effect Sizes

Weighted average effect size ($d.$) represents an unbiased estimate of the population effect size only if single effect sizes are consistent among each other across all k studies (Hedges, 1982a, 1982b; Hedges & Olkin, 1985). If single effect sizes do not statistically deviate from each other more than what is expected by chance (nonsignificant value of the Q_t homogeneity statistic shown below), the estimates of effect sizes differ only by unsystematic sampling error, and we conclude that the model of single effect size fits the data adequately (Hedges & Olkin, 1985). However, if heterogeneity of effect sizes across k studies is significant (significant treatment-by-study interaction), that would indicate that the differences in individual estimates of effect sizes may be large enough (as represented by the significant value of the Q_t statistic) to reject the homogeneity hypothesis that single effect size estimates are drawn from the same population (Hedges, 1986; Hedges & Olkin, 1985).

To test for within-group homogeneity of effect sizes (treatment-by-study interaction), we use the Q_t homogeneity statistic (Cooper & Hedges, 1994; Hedges, 1982a; Hedges, 1986; Hedges & Olkin, 1985), which represents the weighted sum of squares of the effect size estimates d_i to d_k about the weighted mean ($d.$).[5]

$$Q_t = \sum_{i=1}^{k} w_i (d_i - d.)^2 \tag{8}$$

The Q_t value is compared to the χ^2 distribution for $df = (k - 1)$ where k = number of studies.[6]

Moderator Analysis

The grounds for any moderator analysis are first found in the theoretical predictions that are then tested by the means of meta-analytic statistical tests. Thus, if the sample of k studies consists of largely or significantly heterogeneous effect sizes, which is in concert with theoretical expectations, the moderator analysis is performed to identify the sources of systematic variations. However, before we relate the theoretically determined coded characteristics of the studies to variations in effect sizes, we need to examine the consistency of coding for both judges and the variables themselves. We do this by calculating interrater reliability and, usually neglected, "effective" reliability, respectively.

Interrater and "Effective" Reliability

Interrater reliability among a number of judges who coded the studies should be determined for each moderator category and for combined moderator categories by

computing the mean agreement among judges across aggregated categories (see Bullock & Svyantek, 1985, for the specifics on this procedure). Although a number of correlation coefficients could be used for this procedure (see Rosenthal, 1991a, 1991b), the Spearman correlation coefficient (Rho) is recommended because the moderator categories usually represent nominal characteristics of different studies.[7]

Although most meta-analyses report interrater reliability, very few determine effective reliability. The effective reliability is defined as an estimate of the reliability of the variables coded with a comparable group of judges (Rosenthal, 1991). Determining the effective reliability of the mean interrater reliability indicates the probability that a similar group of two other raters would reach the same conclusions regarding the variables coded. Effective reliability is determined by using the Spearman-Brown formula (Walker & Lev, 1953) of the modified notation form (Rosenthal, 1991).

$$R = \frac{nr}{1 + (n-1)r} \tag{9}$$

where R = effective reliability, n = number of judges, and r = mean interrater reliability.

Nature of Moderator Analysis

Moderator analysis for categorical models is conducted by employing Hedges and Olkin's (1985) procedure for fitting parametric fixed effect categorical models to the effect sizes.[8] In particular, we use theoretically predetermined moderators to create groupings that share some common characteristics conceptually expected to influence effect sizes (Hedges, 1986). The partitioning of studies is initiated with moderators believed to be conceptually most related to effect sizes. Three sets of tests, necessary to determine whether the moderator(s) adequately explain the treatment-by-study interaction, and the nature of the moderation, are to be performed: (a) test for homogeneity of effect sizes within moderator groups, (b) test for homogeneity of effect sizes between moderator groups, and (c) multiple orthogonal comparisons among classes (second-level partitioning) (Hedges, 1982b; Hedges, 1986; Hedges & Olkin, 1985).

Homogeneity of Effect Sizes Within Moderator Groups

First, we test for the homogeneity of effect sizes within each newly created grouping (first-level partitioning) to determine whether the grouping variable adequately explains the study-by-treatment interaction and to ensure that no other sources of systematic variance are present in that particular group. To test for within-group homogeneity of effect sizes, we use the Q_w homogeneity statistic (Hedges & Olkin, 1985; see also Hedges, 1982a, 1982b, 1986).

$$Q_w = Q_{w1} + Q_{w2} + Q_{w3} + \ldots + Q_{wp}, \tag{10}$$

where p = number of groups, and Q_{w1} to Q_{wp} are homogeneity statistics for each respective group, derived from the H_t homogeneity statistic, calculated as if each group were an entire collection of studies. The total Q_w thus represents an overall test of homogeneity of effect sizes within the partitioned groups across k studies (Hedges, 1986; Hedges & Olkin, 1985). The value of the Q_w homogeneity statistic is compared to the χ^2 distribution for $df = (k - p)$, where k = number of studies and p = number of groups. Although the Q_w statistic is used as an indicator of overall homogeneity of effect sizes within groups, microanalyses of individual Q_{w1} to Q_{wp} homogeneity statistics is recommended to gain additional insights about the nature of the moderation for each group.

Homogeneity of Effect Sizes Between Moderator Groups

Second, we test for the homogeneity of effect sizes between groups to examine whether average effect sizes significantly differ across moderator groups. For this test, we use the Q_b homogeneity statistic (Hedges & Olkin, 1985; see also Hedges, 1982a, 1982b, 1986).

$$Q_b = Q_t - Q_w. \tag{11}$$

Q_b value is compared to the χ^2 distribution for $df = (p - 1)$, where p = number of groups. We continue this process of subdividing and testing for within- and between-moderator group fit until the within-group homogeneity, with respect to effect magnitude, is achieved (see Hedges & Olkin, 1985, for the detailed discussion on these procedures).[9] This procedure of fitting models to effect sizes for each class is conceptually analogous to fitting hierarchical log linear models to contingency tables (Hedges & Olkin, 1985).

Graphical Presentation of the Data

Because the upper and lower confidence limits for every average unbiased effect size d_{ij} (where i = particular group from first-level partitioning and j = specific class from second-level partitioning), corresponding to the specific combination of the moderator group and class, are computed in almost every meta-analysis, the data usually lend themselves to useful graphical representations. Although there are many graphical options, a clustered side-by-side schematic plot (e.g., Alliger, 1995; Okun, Olding, & Cohn, 1990) is recommended in that it provides the most comprehensive coverage of the data (see Light et al., 1994). A clustered side-by-side schematic plot would, for example, show magnitudes of average unbiased effect sizes for two groups (e.g., manufacturing and service organizations) categorized by types of different classes (e.g., types of interventions), and their attendant confidence intervals (see Stajkovic & Luthans, 1997). The schematic plot of this type usually makes the next question obvious, which is whether the average effect sizes for each grouping produce significant pairwise differences for each linear combination among different classes (see also Stajkovic & Luthans, 1998).

Orthogonal Comparisons Among Moderator Classes

If it is found that effect sizes for final partitions are homogeneous within groups but heterogeneous between classes, we statistically compare the average effect sizes for different classes within each initial group to determine whether they differ from each other. Average effect sizes for each class are compared by means of linear combinations using orthogonal polynomials (Hedges & Olkin, 1985). Using the modified notations, the value of the contrast estimate has the form of

$$C = c_1 d_{\cdot 1} + \ldots + c_p d_{\cdot p} \tag{12}$$

where $c_1 \ldots c_p$ are known constants, $d_{\cdot 1}$ is the weighted average effect size for the i_{th} class, and p is the number of classes. The estimated variance takes the form of

$$V_c = c_1^2 v_{\cdot 1} + \ldots + c_p^2 v_{\cdot p} \tag{13}$$

where $v_{\cdot 1} \ldots v_{\cdot p}$ are variances for corresponding ($d_{\cdot}$).

Simultaneous Orthogonal Comparisons Among Moderator Classes

In the case that several comparisons were to be performed simultaneously, the probability of at least one Type I error would be greater than the chosen significance level of, say, .05. To control for the probability of inflation of Type I error, the method of Bonferroni inequalities (see Hedges & Olkin, 1985; Keppel, 1991) should be used in determining family-wise α level by making comparisons at significance level of

$$FW_\alpha = \frac{\alpha}{2L} \tag{14}$$

where FW_α = family-wise α level, and l = number of simultaneous comparisons.

Limitations

Whenever meta-analysis includes moderator analysis, regardless of the meta-analytic approach used, several theoretical issues should be addressed. First, it is conceptually possible that the moderator variable might be related to sample size, thus causing nonrandom sample selection error (Russell & Gilliand, 1995). Also called "first order sampling error" (Sackett, Tenopyr, Schmitt, & Kehoe, 1985), another possibility is that nonrandom sample selection error occurs when the moderator is related to the content of a sample. In this case, some screening/moderating process, rather than a moderator itself, operates to select certain types of subjects in or out of a particular sample. Although the sample size cannot "cause" the effect (though it definitely can increase the possibility to detect the effect), the sample content certainly can. Thus, besides the always present possibility that sample differences might be due to a "true" impact of

a certain moderator, differences in effect sizes between different samples might also be due to differences in the way samples were constituted.

Considering that in meta-analyses examinations of moderator effects are based on residual variances (Hedges & Olkin's, 1985, $Q_t - \chi^2$—homogeneity of effect sizes, or Hunter & Schmidt's, 1990 75% rule), the moderator analysis can detect the presence of a moderator effect, but it cannot determine the true conceptual processes causing the effect (Russell & Gilliand, 1995). For example, Russell and Gilliand (1995) outlined several specific possibilities in which any meta-analyses cannot determine the true process causing the effect: (a) moderator itself could be the true cause of variations in effect sizes, (b) moderator could be related to sample size, (c) moderator could be related to sample content, (d) moderator can covary with both sample size and sample content, and the most complex option, (e) moderator can be related to sample content but not to sample size, or vice versa.

In each of these cases, residuals-based meta-analyses would suggest heterogeneity of effect sizes within a group of studies, or, in other words, significant treatment-by-study interaction, when in fact the moderator effect conceptually might be due to nonrandom sample selection error. This possibility simply cannot be determined by meta-analyzing (regardless of method used) findings from individual studies. Only primary research with random assignment of subjects to experimental and control groups might, so far, adequately address this problem (Campbell & Stanley, 1963; Cook & Campbell, 1977). Addressing the question of nonrandom sampling error appears to be a topic in need of further methodological and conceptual development within the field of research synthesis.

Conclusion

By outlining specific step-by-step procedures for primary meta-analysis, moderator meta-analysis, and orthogonal comparisons among classes, I hope to have facilitated the future application of Hedges and Olkin's (1985) meta-analytic approach to organizational studies. Also, by providing methodological and conceptual rationale underlying each of the analyses and a modified notational system, the future use of this approach should be further enhanced.

In summary, I believe that the Hedges and Olkin (1985) method can provide another powerful meta-analytic alternative to the research synthesis in the field of organizational behavior and management in general. This seems particularly important given the fact that some of the procedures and corrections of the Hunter and Schmidt (1995) approach are necessarily incompatible with the Hedges (1986) and Hedges and Olkin (1985) analytical framework outlined above (Hedges & Olkin, 1985; Hunter & Schmidt, 1995; Johnson et al., 1995).

APPENDIX
Transforming Different Effect Size Estimates to Effect Size *g*

1. Calculating *g* from *t* or *F*

Since $t^2 = F$ and (A1)

$$t = \text{SQR}(F), \text{ when } n^e \neq n^c \tag{A2}$$

$$g = t\ \{\text{SQR}[(1/n^e) + (1/n^c)]\}, \text{ or} \tag{A3}$$

$$g = t\ \{[\text{SQR}(n^e + n^c)] / [\text{SQR}(n^e\, n^c)]\}, \text{ or when } n^e = n^c \tag{A4}$$

$$g = (2\ t) / (\text{SQR}\ N), \text{ where } N = n^e + n^c \tag{A5}$$

2. Calculating g from r

$$g = \{r / [\text{SQR}(1 - r^2)]\}\ \{\text{SQR}[(df(n^e + n^c)) / (n^e\, n^c)]\} \tag{A6}$$

3. Calculating g from significance levels

When ambiguous significance level p is given such as significant at .05, .01, we first obtain r by using a lower limit estimate (conservative estimate) to find Z level from the table of standard normal deviates, where

$$r = \phi = Z / (\text{SQR}\ N), \text{ and then apply Equation (A6) to find } g. \tag{A7}$$

4. Calculating g from χ^2

(a) When $\chi^2\ df = 1$

To find g from χ^2 for $df = 1$, we use the same procedures as described for significance levels; however, this time first finding Z from the table of standard normal deviates for exact $\chi_{(1)}{}^2$ value, then apply Equation (A7) to find r, and then apply Equation (A6) to find g.

(b) When $\chi^2\ df > 1$

To find g from χ^2 when $df > 1$ relatively complex computations of linear trend in the five means are required. These procedures are detailed by Rosenthal and Rosnow (1985, 1991).

5. Calculating g from proportions

$$g = \{\sin^{-1}[\text{SQR}(p^e)]\} - \{\sin^{-1}[\text{SQR}(p^c)]\}, \tag{A8}$$

where $\sin^{-1}$ is the inverse hyperbolic sine function and p^e and p^c are proportion values for experimental and control groups, respectively.

Notes

1. The list of meta-analyses, databases, dates, and procedures used to make these comparisons is available from the author.

2. Because the variance of the unbiased effect size estimate d conceptually depends on the unknown population parameter δ, we can further use the variance-stabilizing transformation, provided by Hedges and Olkin (1985), to transfer estimator d to a new variate with the simpler distribution. For the simplicity of the notation, I relabeled this new variate $s(d)$, which takes the form of $s(d) = \text{SQR}\ 2 \sinh^{-1}(d/a) = \text{SQR}\ 2 \log\{[(d/a)] + [\text{SQR}(d^2/a^2 + 1)]\}$, where $\sinh^{-1}$ is the inverse hyperbolic function, and $a = \text{SQR}\ \{4 + [2(n^e/n^c)] + [2(n^c/n^e)]\}$.

3. Another possibility to avoid skewing meta-analytic estimates by extremely small (e.g., $n < 10$; Hedges & Olkin, 1985) or large (see Hunter & Schmidt, 1995, for details) sample sizes is to conduct unweighted meta-analysis (e.g., Williams & Livingstone, 1994), in the case of Hedges and Olkin (1985) method, by using g instead of d. However, considering that sample of k studies that might include cases of $n = 1$ usually represents only a minority of experiments, excluding these reports might be a better option than being forced to use less accurate and biased effect size estimator g instead of more stable, unbiased, and weighted effect size estimator d (see Hedges & Olkin, 1985).

4. If a study contains only one effect size, k = number of studies; however, if a study contains multiple effect sizes, and they are included in the analysis, k = total number of effect sizes across all studies.

5. In the notational system presented, the conventional term Q is used (e.g., Cooper & Hedges, 1994; Hedges & Olkin, 1985), although the symbol H (with the same statistical properties) has been and can be substituted (e.g., Hedges, 1982a, 1986).

6. The position taken in this section is that of statistical significance. However, it should be noted that although tests of the significance of Q_t will ordinarily be sufficient to determine if a group of studies share a common effect size, there may be instances in which the value of the Q_t is (a) large yet still nonsignificant (e.g., when there is a large number of studies) or (b) small but significant (sensitive χ^2 distribution-based test "picking up" even the very small variations). In the first case, it is still possible that the studies do not share a common effect size (e.g., Johnson et al., 1995). For example, if the purpose of the meta-analytic review is to investigate if effects sizes depend on the methodological features of the studies, then moderator analyses can still be used, even though the Q_t statistic is not significant. In these occasions, a meta-analyst may wish to reserve judgment regarding whether the studies share a common effect size until moderator analyses are completed. In the second case, a meta-analyst should determine whether the unexplained systematic variance needs to be addressed by the next level of moderation (e.g., theoretical considerations), or if the remaining variance is simply due to method artifacts (e.g., sensitive test) and is not necessarily, in the words of Hunter and Schmidt (1995), meaningful in organizational reality.

7. Cohen's kappa represents another conventional way to examine the reliability of moderator codes for categorical judgments (see Siegel & Castellan, 1988).

8. Hedges and Olkin (1985) also provide procedures for continuous models (modeled on regression) with the same underlying logic (weighted general linear model).

9. More "flexibility" in searching for correctly specified models is allowed when multiple regression procedures are used.

References

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.

Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. *Psychological Methods, 2*, 131-160.

Bullock, R. J., & Svyantek, D. J. (1985). Analyzing meta-analysis: Potential problems, an unsuccessful replication, and evaluation criteria. *Journal of Applied Psychology, 70*, 108-115.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand McNally.

Cook, T. D., & Campbell, D. T. (1977). *Quasi-experimentation*. Boston, MA: Houghton Mifflin.

Cooper, H., & Hedges, L. V. (1994). *Handbook of research synthesis*. New York: Russell Sage.

Cooper, H. M. (1984). *The integrative research review: A systematic approach*. Beverly Hills, CA: Sage.

Davis, T.R.V., & Luthans, F. (1984). Defining and researching leadership as a behavioral construct: An idiographic approach. *Journal of Applied Behavioral Science, 20*, 237-251.

Friske, R., & Treines, G. (1985). *Einführung in die metaanalyse*. Bern: Hans Huber.

Glass, G. V. (1976a). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*, 3-8.

Glass, G. V. (1976b). *Presidential address*. Annual meeting of the American Educational Research Association, San Francisco.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*, 107-128.

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). Estimating effect sizes from a series of independent experiments. *Psychological Bulletin, 92*, 490-499.

Hedges, L. V. (1986). Issues in meta-analysis. *Review of Research in Education, 13*, 353-398.

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

Huffcutt, A. I., & Arthur, W. (1995). Development of a new outlier statistic for meta-analytic data. *Journal of Applied Psychology, 80*, 327-334.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis*. Newbury Park, CA: Sage.

Hunter, J. E., & Schmidt, F. L. (1995). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis*. Beverly Hills, CA: Sage.

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

Keppel, G. (1991). *Design and analysis*. Englewood Cliffs, NJ: Prentice Hall.

Light, R. J. (Ed.). (1983). *Evaluation studies review annual* (Vol. 8). Beverly Hills, CA: Sage.

Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research*. Cambridge, MA: Harvard University Press.

Light, R. J., Singer, J. D., & Willett. (1994). The visual presentation and interpretation of meta-analysis. In H. Cooper & L. V. Hedges (Eds.), *Handbook of research synthesis* (pp. 439-453). New York: Russell Sage.

Nouri, H., & Greenberg, R. H. (1995). Meta-analytic procedures for estimation of effect sizes in experiments using complex analysis of variance. *Journal of Management, 21*, 801-812.

Okun, M. A., Olding, R. W., & Cohn, C. M. (1990). A meta-analysis of subjective well being interventions among elders. *Psychological Bulletin, 103*, 257-265.

Osburn, H. G., & Callender, J. (1992). A note on the sampling variance of the mean uncorrected correlation in meta-analysis and validity generalization. *Journal of Applied Psychology, 77*, 115-122.

Rosenthal, R. (Ed.). (1980). *New directions for methodology of social and behavioral science: Quantitative assessment of research domains* (No. 5). San Francisco, CA: Jossey-Bass.

Rosenthal, R. (1991a). *Meta-analytic procedures for social research*. Newbury Park, CA: Sage.

Rosenthal, R. (1991b). *Meta-analytic procedures for social research* (2nd ed.). Newbury Park, CA: Sage.

Rosenthal, R. (1994). Parametric measures of effect sizes. In H. Cooper & L. V. Hedges (Eds.), *Handbook of research synthesis* (pp. 231-244). New York: Russell Sage.

Rosenthal, R., & Rosnow, R. L. (1985). *Contrast analysis: Focused comparisons in the analysis of variance*. New York: Cambridge University Press.

Rosenthal, R., & Rosnow, R. L. (1991). *Essential of behavioral research: Methods and data analysis* (2nd ed.). New York: McGraw-Hill.

Rosenthal, R., & Rubin, D. (1978). Interpersonal expectancy effects: The first 345 studies. *Behavioral and Brain Sciences, 3*, 377-415.

Rosenthal, R., & Rubin, D. (1988). Comment: Assumptions and procedures in the file drawer problem. *Statistical Science, 3*, 120-125.

Russell, G. J., & Gilliand, S. W. (1995). Why meta-analysis doesn't tell us what the data really mean: Distinguishing between moderator effects and moderator processes. *Journal of Management, 21*, 813-831.

Sackett, P. R., Tenopyr, M. L., Schmitt, N., & Kehoe, J. (1985). Commentary on forty questions about validity generalizations and meta-analysis. *Personnel Psychology, 38*, 697-798.

Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics*. New York: McGraw Hill.

Stajkovic, A. D., & Luthans, F. (1997). A meta-analysis of the effects of organizational behavior modification on task performance, 1975-95. *Academy of Management Journal, 40*, 1122-1149.

Stajkovic, A. D., & Luthans, F. (1998). Self-efficacy and work-related performance: A meta-analysis. *Psychological Bulletin, 12*(4), 240-261.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Walker, H. M., & Lev, J. (1953). *Statistical inference*. New York: Holt, Rinehart, & Winston.

Williams, C. R., & Livingstone, L. (1994). Another look at the relationship between performance and voluntary turnover. *Academy of Management Journal, 37*, 269-298.

Yeaton, W. H., & Wortman, P. M. (Eds.). (1984). *New directions in program evaluation: Issues in data synthesis* (No. 24). San Francisco, CA: Jossey-Bass.