Comparison of Three Meta-Analytic Procedures for Estimating Moderating Effects of Categorical Variables

Herman Aguinis, University of Colorado at Denver
Michael C. Sturman, Cornell University
Charles A. Pierce, University of Memphis

The authors conducted Monte Carlo simulations to compare the Hedges and Olkin, the Hunter and Schmidt, and a refinement of the Aguinis and Pierce meta-analytic approaches for estimating moderating effects of categorical variables. The simulation examined binary moderator variables (e.g., gender—male, female; ethnicity—majority, minority). The authors compared the three meta-analytic methods in terms of their point estimation accuracy and Type I and Type II error rates. Results provide guidelines to help researchers choose among the three meta-analytic techniques based on theory (i.e., exploratory vs. confirmatory research) and research design considerations (i.e., degree of range restriction and measurement error).

The past few decades have seen meta-analysis emerge as a pervasive data-analytic strategy to review bodies of literature quantitatively. Historically, meta-analysis has most often been used to assess direct relationships (e.g., Is X related to Y? What is the effect size associated with an intervention?). At present, however, as numerous areas in management and related fields continue to make theoretical advancements, meta-analytic applications are routinely used to assess moderating effects. A moderating effect occurs when the direct relationship between two variables is contingent on the value of a third, or moderator, variable (Aguinis & Stone-Romero, 1997).

Numerous theories in management and related fields posit the operation of moderating effects of categorical variables (i.e., variables with discrete values; Aguinis, 2004; Aguinis, Petersen, & Pierce, 1999). A perusal of journals and books in management and related fields shows that a large number of primary-level studies have included tests of hypotheses involving categorical moderator variables such as gender, ethnicity, job title, position in an organizational hierarchy, organizational configuration, turnover, and mentoring (e.g., Aguinis, 2007; Aguinis, Boik, & Pierce, 2001; Richard, 2000). In fact, a recent literature review concluded there are few, if any, theories in management and related fields that do not include categorical moderator variables (Aguinis, Beaty, Boik, & Pierce, 2005).

Given the increasing number of primary-level studies testing hypotheses regarding categorical moderator variables, researchers have become interested in testing such hypotheses at the meta-analytic level. In fact, several such meta-analyses have been published recently. For instance, Dalton, Daily, Johnson, and Ellstrand (1999) tested the moderating effect of firm size (small or large) on the relationship between size of board of directors and financial performance. Elsewhere, Phillips (1998) examined the moderating effect of the medium of a realistic job preview (written, verbal, or videotape) on the relationship between realistic job previews and several organizational outcomes. Sturman and his colleagues (Sturman, 2003; Sturman, Cheramie, & Cashen, 2005) tested the moderating effects of type of job performance measure (objective or subjective) on various predictor-job performance relationships. These are only a few examples; numerous additional illustrations of meta-analytic tests of hypotheses regarding moderating effects of categorical variables can be found regularly in journals in management, applied psychology, and related fields.

Similar to primary-level researchers who are faced with a number of options regarding data analysis strategies, meta-analytic-level researchers interested in testing hypotheses regarding categorical moderator variables also face several options. However, unlike primary-level researchers, meta-analytic-level researchers seldom explain or justify their choice for a specific data-analytic approach, and both the execution and reporting of results are highly variable (Cortina, 2003). In fact, Johnson, Mullen, and Salas (1995) noted that the "selection of one approach over another is merely a matter of choice" (p. 94). It seems that meta-analysts often choose a specific meta-analytic technique based on habit, the availability of and their familiarity with a specific software package, or usage trends in specific topic areas rather than the relative merits of the available approaches.

The fact that meta-analysts do not often justify or explain their choice for a specific meta-analytic approach may in part be because of the lack of guidelines in the literature regarding which strategy to use under which conditions. Specifically, which technique is most likely to lead to a Type I error (i.e., erroneously asserting the presence of a moderator variable) and under which conditions? Which technique produces the most accurate estimates of moderating effect magnitude? Which one is most likely to lead to a Type II error (i.e., erroneously dismissing the presence of a categorical moderator variable) and under which conditions? Which technique is most appropriate for testing a priori hypotheses regarding moderating effects of categorical variables? Which is most appropriate for testing post hoc hypotheses?

Present Study

The purpose of the present study is to provide guidelines regarding which of three meta-analytic approaches (i.e., Aguinis & Pierce, 1998; Hedges & Olkin, 1985; Hunter & Schmidt, 1990, 2004) is most appropriate under specific theory development and research design conditions. Although the past decade has produced several comparisons of meta-analytic methods (e.g., Field, 2005; Fuller & Hester, 1999; S. M. Hall & Brannick, 2002; Hough & Hall, 1994; Johnson et al., 1995), there are only a few studies that have compared meta-analytic tests for moderator variables. Six notable exceptions are Cornwell and Ladd (1993), Sagie and Koslowsky (1993), Johnson et al. (1995), Marin-Martinez and Sanchez- Meca (1998), Overton (1998), and Steel and Kammeyer-Mueller (2002). Although these six studies have several merits, they do not provide a complete picture for evaluating which meta-analytic method should be used to test hypotheses about moderator variables. For example, Sagie and Koslowsky examined the accuracy of seven techniques for detecting the presence of moderators meta-analytically, but these seven techniques (including the Q statistic) were all derived from the Hunter and Schmidt (1990) approach. And although Sagie and Koslowsky examined overall homogeneity tests, they did not examine specific tests for moderators. Likewise, although Cornwell and Ladd examined overall homogeneity tests based on the Hunter and Schmidt approach, they too did not examine specific tests for moderators. Unlike Sagie and Koslowsky, and Cornwell and Ladd, Johnson et al. compared specific tests for moderators across three meta-analytic approaches including the Hedges and Olkin (1985) and the Hunter and Schmidt (1990) methods. However, Aguinis and Pierce's (1998) proposed meta-analytic method, which incorporates aspects of the Hedges and Olkin and the Hunter and Schmidt approaches, appeared after publication of the Johnson et al. study. More recently, Overton examined fixed-effects and random-effects meta-analytic approaches, but his study did not examine the Hunter and Schmidt or the Aguinis and Pierce meta-analytic approaches. Hence, Overton's study did not assess the effects of range restriction and measurement error, which are pervasive methodological artifacts examined herein. Unlike Overton, Marin-Martinez and Sanchez-Meca compared the Hunter and Schmidt and the Hedges and Olkin meta-analytic approaches. However, like Overton, Marin-Martinez and Sanchez-Meca did not examine the effects of range restriction and measurement error on tests for moderators. Lastly, although Steel and Kammeyer- Mueller examined meta-analytic tests for moderators, they did not examine the impact of measurement characteristics on the accuracy of moderator tests. In sum, although there has been preliminary research on moderator tests in meta-analysis, there remains unexplored methodological issues in need of direct examination to cover the variety of meta-analytic techniques that exist and the different conditions that meta-analysts face (Viswesvaran & Sanchez,

1998). Furthermore, the accuracy of the Aguinis and Pierce meta-analytic approach has yet to be examined.

In this study, we contribute to the literature exploring moderator tests in meta-analysis by empirically comparing the following three meta-analytic approaches for testing hypotheses regarding moderating effects of categorical variables: (a) Hedges and Olkin (1985), (b) Hunter and Schmidt (1990, 2004), and (c) a refinement of Aguinis and Pierce (1998). The reason for studying the Hedges and Olkin and the Hunter and Schmidt approaches is that they are the two techniques most frequently used in management, applied psychology, and related fields (Johnson et al., 1995). The reason for investigating a refined version of the Aguinis and Pierce technique is that it combines aspects of both the Hedges and Olkin and the Hunter and Schmidt approaches. To date, no empirical research has examined the performance of the Aguinis and Pierce approach. Thus, no empirical study has compared the performance of this more recently proposed approach in relation to the Hunter and Schmidt and Hedges and Olkin approaches regarding the estimation of categorical moderators or, moreover, pointed to the advantages of one approach over the other under the varied conditions usually encountered by researchers. Next, we describe briefly each of these three meta-analytic approaches for testing hypotheses regarding moderating effects of categorical variables.

Hedges and Olkin Approach for Estimating Moderating Effects of Categorical Variables

Hedges and Olkin (1985) proposed a meta-analytic approach to ascertain the magnitude of the relationship between two variables X and Y, the variability of this relationship across primary-level studies, and the moderator variables that account for such variability. The degree of variability of effect-size estimates across studies is assessed with the homogeneity statistic Q. A statistically significant Q suggests that the primary-level effect-size estimates do not estimate a common population effect size, and, therefore, the subsequent search for moderating effects is warranted.

In tests for moderating effects, each study is assigned a numerical value based on the moderator (e.g., gender, 1 = female, 0 = male) and grouped according to this coding scheme. The difference between mean group effect sizes is assessed by computing a between-group homogeneity statistic QB. The presence of the moderator is indicated by a statistically significant QB, which suggests a difference between the mean effect-size estimates across groups.

Hunter and Schmidt Approach for Estimating Moderating Effects of Categorical Variables

Hunter and Schmidt (1990, 2004) proposed a meta-analytic approach based on psychometric principles and contended that a substantial portion of the variability observed in an X-Y relationship

across primary-level studies is the result of artifactual sources of variance (e.g., sampling error, measurement error in the dependent or criterion variable, range restriction). Stated differently, across-study variability in effect-size estimates may be because of (a) methodological and statistical artifacts and/or (b) moderating effects. Consequently, to estimate better moderators of the X-Y relationship in the population, meta-analysts should correct for artifactual across-study variability by controlling it via research design and/or subtracting it from the total observed variance. Hunter and Schmidt argued that unless the artifactual variability is removed, across-study variability may be attributed to "false" moderating effects, and, thus, researchers may commit a Type I statistical error.

Although various rules can be used (Sagie & Koslowsky, 1993), the existence of substantive variability (i.e., not because of artifactual sources) in primary-level effect-size estimates is typically assessed with the 75% rule. The 75% rule indicates that if less than 75% of variance in effect-size estimates is because of artifacts (i.e., sampling error variance, unreliability, and range restriction), it is likely that there is substantive variance and, thus, the search for moderators is warranted (Hunter & Schmidt, 1990). The argument for the 75% rule is that researchers can never correct for all artifacts that cause variance across studies because researchers 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 because of extraneous factors). The Hunter and Schmidt approach assumes that if correctable factors account for at least 75% of the across-study variance, then it is likely the remaining variance is accounted for by uncorrectable factors.

In tests for moderating effects, similar to Hedges and Olkin's approach, studies are assigned a numerical value based on the moderator and grouped accordingly. Then, although statistical significance testing is not advocated by the Hunter and Schmidt approach, the mean effect sizes can be compared across groups using a chi-square or t statistic.

Aguinis and Pierce Approach for Estimating Moderating Effects of Categorical Variables

The Aguinis and Pierce (1998) procedure is based on the Hedges and Olkin (1985) approach, yet it incorporates study-level corrections for methodological and statistical artifacts advocated by the Hunter and Schmidt approach. The Aguinis and Pierce procedure includes the following three steps. First, study-level effect-size estimates are corrected for methodological and statistical artifacts (e.g., measurement error in the independent and dependent variables, range restriction). Second, the homogeneity of corrected study-level effect-size estimates is tested using a modified version of Hedges

and Olkin's Q statistic (i.e., Q'). Note that, in contrast to Q, Q' tests for the homogeneity of individually corrected (for statistical and methodological artifacts) effect-size estimates. Q' and, more generally, the Aguinis and Pierce approach have never been empirically examined in terms of Type I and Type II error rates. Finally, hypotheses regarding moderator effects are tested using a modified homogeneity statistic Q'B. Note that, in contrast to QB, the modified homogeneity statistic Q'B tests for between-group differences in mean corrected effect-size estimates.

Following Hunter and Schmidt (1990), Aguinis and Pierce (1998) noted that the adjustment for the variance of individually corrected correlations should be computed using the square of a correction factor (i.e., corrected correlation/observed correlation), typically labeled a. However, using this compound factor is inappropriate because the Aguinis and Pierce approach uses ds (i.e., standardized difference between mean scores) as the focal effect-size estimate and not correlation coefficients. Thus, the appropriate correction factor is corrected d/observed d and not corrected r/observed r. The present study incorporates this refinement into the Aguinis and Pierce procedure (more detail on this issue is provided in the Procedure and Dependent Variables section).

Method

Overview

Monte Carlo simulations were conducted to compare the relative performance of the Hedges and Olkin (1985), Hunter and Schmidt (1990, 2004), and Aguinis and Pierce (1998) procedures when meta-analytically testing for the presence of a categorical moderator variable. To facilitate the comparison, we simulated situations involving a moderator variable with two levels only (e.g., gender). First, we specified the simulation parameters. Second, we generated the primary-level data to be subsequently combined into sets to simulate 656,100 separate meta-analyses. Third, each set of effect sizes (i.e., each meta-analysis) was analyzed using the three meta-analytic approaches. Fourth, we examined the accuracy of the point estimates (i.e., mean effect-size estimate vs. population effect size specified in the first step) and the accuracy of hypothesis tests regarding Type I and Type II error rates.

Manipulated Parameters

The following parameters were manipulated in the simulation:

Number of studies and sample sizes. The number of studies in each meta-analysis was set to 10, 50, or 100. These values cover ranges typically found in management and related fields (e.g., Russell

et al., 1994) and are consistent with values used in evaluations of meta-analytic procedures (e.g., Aguinis, 2001; Aguinis & Whitehead, 1997).

In primary-level studies involving categorical moderator variables in general and binary moderators in particular, it is often the case that the number of data points (e.g., individuals) differs across levels of the moderator (e.g., women vs. men, ethnic majority vs. ethnic minority group members). Thus, the simulation included the following pairings of sample size for each of the two moderator-based subgroups: (a) n1 = 50, n2 = 20; (b) n1 = 50, n2 = 50; (c) n1 = 100, n2 = 20; (d) n1 = 100, n2 = 50; (e) n1 = 100, n2 = 100; (f) n1 = 250, n2 = 20; (g) n1 = 250, n2 = 50; (h) n1 = 250, n2 = 100; and (i) n1 = 250, n2 = 250. These values were chosen so as to cover a range of possible values including similarly small (e.g., 50 and 50), similarly large (e.g., 250 and 250), and dissimilar (e.g., 100 and 50, 250 and 100) sample sizes across moderator-based subgroups. In addition, these specific values were chosen so as to be representative of situations involving hypothesis tests of binary moderator variables in management, applied psychology, and related fields (e.g., Gerstner & Day, 1997; Huffcutt, Roth, & McDaniel, 1996).

Population moderating effect size. To represent various levels of effect size for the moderating effect, we used X-Y population correlation coefficient values of .1, .3, or .5 for each moderator-based subgroup (i.e., $\rho 1$ = .1, .3, or .5; $\rho 2$ = .1, .3, or .5). Thus, each cell included one value for $\rho 1$ and one value for $\rho 2$, and the magnitude of the population moderating effect assessed as the absolute difference between the correlations in the moderator-based subgroups ($|\rho 1 - \rho 2|$) took on values of 0, .2, or .4. These values cover the typical range found in management and related fields (Aguinis & Stone-Romero, 1997).

Artifactual sources of variance. In addition to sampling error, the two most influential and pervasive sources of artifactual variance are measurement error and range restriction (Hunter & Schmidt, 1990, 2004). Thus, to simulate realistic meta-analytic situations, these variables were incorporated into the simulation design. Regarding measurement error, we varied the level of unreliability for both the predictor and criterion variables. We set reliabilities to .60, .80, and 1.00 to cover values ranging from what can be considered substandard (i.e., .60) to perfect (i.e., 1.00). Regarding range restriction, we varied the extent of restriction in each primary-level study by using the following selection ratios: 10%, 50%, or 100%. These values represent situations ranging from a very severe degree of restriction (i.e., selection ratio = 10%, where only the top 10% of population scores are available in the sample) to a situation where the entire range of population scores is available in the sample (i.e., selection ratio = 100%, where there is an absence of range restriction).

Summarizing the Manipulated Parameters section, the manipulation of the independent variables led to a full factorial design having a total of 6,561 cells or meta-analyses, whereby each meta-analysis represents one unique combination of independent variable values. For each cell, 100 iterations were simulated, thus resulting in a total of 656,100 total simulated meta-analyses. This set of meta-analyses ultimately involved the generation of a total of more than 10 billion simulated individual scores.

Procedure and Dependent Variables

Computer program. The programs were written in Visual Basic. The compiled program was simultaneously executed on three IBM-compatible 1.86 GHz computers. Total computing time to complete the simulations was roughly 514 hours.

Simulation procedure. The simulation involved the following four steps. First, bivariate (X, Y) arrays of size n were generated from multivariate normal populations with a mean of zero (i.e., $\mu_X = \mu_Y = 0$), unit variance (i.e., $\sigma_X = \sigma_Y = 1.0$), and correlation ρ. The value of ρ depended on whether an individual score belonged in moderator-based Subgroup 1 or Subgroup 2 (i.e., ρ_1 or ρ_2). Thus, primary-level studies had sample size n_1 or n_2 and effect size estimates ρ_1 or ρ_2, depending on whether the study included individuals in one or the other moderator-based category (e.g., men or women). Second, after the values for ρ_1 and ρ_2 were specified and the initial values generated, random error was added to the X and Y variables to manipulate reliability. Third, the n_1 and n_2 scores were sorted in descending order on X and truncated at the nth value to manipulate range restriction. Identical to Millsap (1989) and Aguinis and Whitehead (1997), the number of individual scores generated (n_1 and n_2) was manipulated so that n_1/n and n_2/n would equal the specified selection ratio. Fourth, the 656,100 meta-analyses were analyzed using each of the three meta-analytic approaches. Specifically, we implemented the following procedures.

Hedges and Olkin (1985) approach. We implemented the following steps:

1. Converted study-level rs to ds using Wolf's (1986, p. 35) formula. The Hedges and Olkin approach was originally designed to analyze ds. This is the reason why the most popular software package that implements the Hedges and Olkin approach is called DSTAT (Johnson, 1993). However, the Hedges and Olkin approach also allows for the analysis of rs, but such analysis requires the controversial r to Fisher's z transformation (S. M. Hall & Brannick, 2002). When a meta-analyst has a data set including rs instead of ds, he or she has the choice to (a) convert rs to ds and analyze ds or (b) convert rs to Fisher's zs and

analyze Fisher's zs (which is a controversial procedure). Accordingly, we implemented an r to d transformation instead of the more controversial r to Fisher's z transformation. Nevertheless, we conducted preliminary analyses based on Fisher's z in addition to ds. The correlation between the point estimates generated using these two procedures was .998. Considering the negligible difference in results, we conducted all analyses based on the less controversial ds. Finally, we used ds and not gs (Glass, 1976) because g has a small sample bias, whereas d is an unbiased estimator of effect size (Hedges & Olkin, 1985, p. 81).

2. Computed Var(d). This is the variance of the effect-size estimates (Hedges & Olkin, 1985, p. 151, Equation 8).
3. Computed di+ (Hedges & Olkin, 1985, p. 152, Equation 9). This is the weighted mean effect-size estimator in each of the two moderator-based subgroups.
4. Computed d++ (Hedges & Olkin, 1985, p. 152, Equation 10). This is the grand mean of all effect-size estimates.
5. Computed Var(di+) (Hedges & Olkin, 1985, p. 152, Equation 14).
6. Computed Q (Hedges & Olkin, 1985, p. 153, Equation 16). This is the homogeneity statistic used to ascertain whether study-level effect sizes estimate a common population effect size.
7. Computed QB (Hedges & Olkin, 1985, p. 154, Equation 17). This is the homogeneity statistic used to ascertain whether the effect-size estimates differ across moderator-based subgroups. Note that QB approximates a chi-square distribution with 1 degree of freedom because the moderator examined has two levels.

Hunter and Schmidt (1990, 2004) approach. We implemented the following steps:

1. Computed the weighted average mean r.
2. Corrected each study-level r for the following artifacts:
   2.1. Corrected for range restriction (Hunter & Schmidt, 1990, p. 128; 2004, p. 107).
   2.2. Corrected for unreliability in X and Y variables (Hunter & Schmidt, 1990, p. 121; 2004, p. 96). Thus, the correction for measurement error was performed on the unrestricted X and Y variables.
3. Computed Var(ei) (Hunter & Schmidt, 1990, p. 148; 2004, p. 123). This is the sampling error variance of corrected correlation coefficients.

4. Computed a and multiplied Var(ei) by a2 to obtain ve (Hunter & Schmidt, 1990, p. 146; 2004, p. 122). a is a correction factor, and ve is a superior estimate of the sampling error variance of corrected correlation coefficients.
5. Computed w_i for each sample (Hunter & Schmidt, 1990, p. 148; 2004, p. 123). This is a weight for each study and is the product of sample size and the artifact correction factor.
6. Computed mean r, Var(r), and Ave(ve) (Hunter & Schmidt, 1990, p. 150; 2004, pp. 125-126).
7. Computed whether more than 75% of variance was accounted for by sampling error variance (Hunter & Schmidt, 1990, p. 165; 2004, p. 145).

Although the Hunter and Schmidt (1990, 2004) procedure does not advocate the use of null hypothesis significance testing, once the above information was calculated, a statistical significance test was performed to determine whether there was a difference between the correlations across the two moderator-based subgroups. Specifically, we used the following equation (Neter, Wasserman, & Whitmore, 1988, p. 402):

$$t = \frac{|\bar{\bar{r}}_1 - \bar{r}_2|}{\sqrt{\frac{Var(r_1)}{k} + \frac{Var(r_2)}{k}}}$$

where $\bar{\bar{r}}_1$ and $\bar{r}_2$ are the average corrected correlations for each of the moderator-based subgroups and Var(r_1) and Var(r_2) are the variance in these correlations. These values are all calculated using the same method as the overall estimates (from step 6 above), except that there is a separate estimate for each of the groups under examination in the study. Note that either r--based or ρ-based notations can be used because $\bar{r}$ is the best estimate of the population correlation ρ. Finally, k represents the number of studies in the meta-analysis, which in our simulation is the same across the two subgroups, and the t distribution has ($2k - 2$) degrees of freedom.

Aguinis and Pierce (1998) approach. We implemented the following steps:

1. Corrected each primary-level effect-size estimate (i.e., correlation coefficient) for the following artifacts:
   1.1. Corrected for range restriction (Hunter & Schmidt, 1990, p. 128; 2004, p. 107).
   1.2. Corrected for unreliability in X and Y variables (Hunter & Schmidt, 1990, p. 121; 2004, p. 96).
2. Converted rs to ds, using the formula reported by Wolf (1986, p. 35).
3. Computed Var(d) (Hedges & Olkin, 1985, p. 151, Equation 8).

4. Computed d_{i+} (Hedges & Olkin, 1985, p. 152, Equation 9).
5. Computed d_{++} (Hedges & Olkin, 1985, p. 152, Equation 10).
6. Computed Var'(d) by adjusting Var(d). Aguinis and Pierce (1998, p. 585, Equation 8)noted that this adjustment should be made by multiplying Var(d) by a correction factor for range restriction and measurement error (cf. Hunter & Schmidt, 1990, pp. 121, 146). However, the correction factor was referenced to the correlation coefficient despite the fact that Var'(d) is the variance of corrected ds and not the variance of corrected rs. Referencing the correction factor to r is appropriate for the Hunter and Schmidt (1990, 2004) method but inappropriate if the focal effect-size estimate is d. The compound correction factor should therefore be based on the ds. Thus, $\mathrm{a} = \text{corrected d} \mid \text{observed d}$, and we refined the Aguinis and Pierce approach by calculating Var'(d) as: $\text{Var'(d)} = (\text{corrected d} \div \text{observed d})^2 \cdot \text{Var(d)}$.
7. Computed Q' (Aguinis & Pierce, 1998, p. 584, Equation 5). This is the homogeneity statistic used to ascertain whether corrected study-level effect sizes estimate a common population effect size.
8. Computed Q'_B (Aguinis & Pierce, 1998, p. 585, Equation 9). This is the homogeneity statistic based on corrected effect-size estimates. Note that Q'_B approximates a chi-square distribution with 1 degree of freedom.

Dependent Variables

For each of the three meta-analytic approaches, we computed (a) point estimates of the population correlation coefficients, (b) Type I error rate (α was set at .05), and (c) Type II error rate (for situations in which the moderating effect is > 0).

Key Accuracy Checks

To assess the key accuracy of the computer programs, two doctoral students who were not otherwise involved in this project and were taking an independent study class on meta-analysis checked all the algorithms. No procedural errors were found, and a few minor typographical errors were discovered and corrected. Then, to assess further the accuracy of the algorithms used to implement each of the three meta-analytic procedures, the second author replicated illustrative meta-analyses reported in Hedges and Olkin (1985), Hunter and Schmidt (1990, 2004), and Aguinis and Pierce (1998). Results obtained using our computer programs were identical to those reported by Hedges and Olkin,

Hunter and Schmidt, and Aguinis and Pierce. Altogether, these key accuracy checks demonstrate the validity of the computer programs used in the simulation.

Results

First, we present overall results regarding point estimation, homogeneity tests, and moderating effect tests. Then, we describe more detailed results pertaining to the performance of homogeneity tests (Type I and Type II error rates) and moderating effect tests (Type I and Type II error rates).

Overall Performance of the Three Meta-Analytic Approaches: Point Estimation

Table 1 summarizes the overall performance of the three meta-analytic approaches regarding point estimation collapsing across all 6,561 design conditions. Table 1 shows that the Hunter and Schmidt approach yielded the most accurate point estimates (i.e., smallest errors) of the overall correlation coefficient collapsing across moderator-based subgroups. The Aguinis and Pierce approach was second in performance. A t test revealed that the Aguinis and Pierce approach had a statistically significant greater level of error than the Hunter and Schmidt approach (at $p < .0001$); however, the difference in mean errors between these two approaches was only .006.

The Hedges and Olkin approach had the greatest level of error, exceeding the other two methods by more than .08. This is an expected finding because the Hunter and Schmidt and Aguinis and Pierce techniques include explicit corrections for artifacts (e.g., unreliability of measurement, range restriction) that produce a bias in the point estimates (i.e., ρ_1 and ρ_2 and, consequently, $|\rho_1 - \rho_2|$). Specifically, the presence of these artifacts causes the estimates of ρ_1 and ρ_2 to be downwardly biased. Thus, correcting for these artifacts produced more accurate estimates of overall population correlation coefficients collapsing across subgroup membership. Because of the greater accuracy of the Hunter and Schmidt and Aguinis and Pierce procedures regarding the overall correlation, the estimation of the moderator-based subgroup correlations (i.e.,ρ_1 and ρ_2) and the moderating effect (i.e., $|\rho_1 - \rho_2|$) is also more accurate.

Table 1 also shows a correlation matrix of point estimates derived using each of the three meta-analytic approaches. Results show that there is an almost perfect correlation between absolute errors in point estimates produced by the Hunter and Schmidt and Aguinis and Pierce approaches (i.e., $r = .98$, $p < .0001$), meaning that the point estimates are similarly affected by the various simulated conditions. The relationship between the Hedges and Olkin and the Aguinis and Pierce approaches was second in magnitude (i.e., $r = .82$, $p < .0001$), followed closely by the correlation in absolute errors between the Hunter and Schmidt and Hedges and Olkin approaches (i.e., $r = .79$, $p < .0001$).

Table 1
Summary of Point Estimates for the Three Meta-analytic Approaches

Method	|Error|	*SD*	H-S	H-O	A-P
Hunter and Schmidt (1990, 2004; H-S)	.107	.076	1.00		
Hedges and Olkin (1985; H-O)	.193	.116	.79	1.00	
Aguinis and Pierce (1998; A-P)	.113	.080	.98	.82	1.00

Note: |Error| = average difference between the absolute value of true (i.e., parameter generated via simulation) versus estimated (i.e., empirically derived) correlation for entire data set (i.e., collapsing across subgroup membership). The correlation matrix shows relationships between point estimates obtained using each of the three approaches. Analyses use the two correlations from each simulation case; thus, the sample size for the above analyses is 1,312,200 (6,561 conditions × 100 cases per condition × 2 correlations per case). All correlations are significant at $p < .0001$.

Table 2
Summary of Type I and Type II Error Rates for Overall Homogeneity Tests for the Three Meta-Analytic Approaches

	Hunter and Schmidt (1990, 2004) 75% rule	Hedges and Olkin (1985) *Q*	Aguinis and Pierce (1998) *Q*'
Type I error			
$\|\rho_1 - \rho_2\| = 0$	.06	.11	.20
Type II error			
$\|\rho_1 - \rho_2\| = .2$	.82	.68	.60
$\|\rho_1 - \rho_2\| = .4$	.47	.37	.32

Note: Type I error rate was set at .05. The number of simulation conditions for which $|\rho_1 - \rho_2| = 0$ is 218,700; the number of simulation conditions for which $|\rho_1 - \rho_2| = .2$ is 291,600; and the number of simulation conditions for which $|\rho_1 - \rho_2| = .4$ is 145,800.

Overall Performance of Homogeneity Tests: Type I and Type II Error Rates

Table 2 summarizes results regarding the performance of the three meta-analytic approaches pertaining to homogeneity tests. As described above, homogeneity tests indicate whether the primary-level studies estimate a common population effect size. To test for homogeneity, the Hunter and Schmidt approach uses the 75% rule, the Hedges and Olkin approach uses the Q statistic, and the Aguinis and Pierce approach uses the Q' statistic.

Table 2 shows results regarding Type I and Type II error rates collapsing across all cells in the design for situations including true population homogeneity (i.e., $|\rho_1 - \rho_2| = 0$) and situations including true population heterogeneity (i.e., $|\rho_1 - \rho_2| = .2$ and $|\rho_1 - \rho_2| = .4$). Results show that the Hunter and Schmidt procedure was superior to the Hedges and Olkin and Aguinis and Pierce techniques with regard to the control of Type I error rates at the preset level (i.e., $\rangle = .05$). In the absence of a moderating effect in the population (α, $|\rho_1 - \rho_2| = 0$), the Hunter and Schmidt 75% rule yielded a Type I error rate of .06 as

compared to error rates of .11 for the Hedges and Olkin Q statistic and .20 for the Aguinis and Pierce Q' statistic.

Table 2 also shows results regarding Type II error rates for a moderating effect size of $|p_1 - p_2| = .2$ and a moderating effect size of $|p_1 - p_2| = .4$. Overall, as expected, Type II error rates (i.e., incorrectly failing to reject a null hypothesis of $|p_1 - p_2| = .0$) were greater for the smaller moderating effect (i.e., .2) for all three meta-analytic approaches. In addition, Table 2 shows that Type II error rates for $|p_1 - p_2| = .2$ were .60 for the Aguinis and Pierce approach, .68 for the Hedges and Olkin method, and .82 for the Hunter and Schmidt procedure. This overall high Type II error situation improves for $|p_1 - p_2| = .4$. Specifically, Type II error rates ranged from a low of .32 (Aguinis and Pierce) to a high of .47 (Hunter and Schmidt). The Hedges and Olkin approach yielded an intermediate error rate of .37.

Overall Performance of Moderating Effect Tests: Type I and Type II Error Rates

Table 3 summarizes results regarding the performance of the three meta-analytic approaches regarding moderating effect tests (i.e., tests of whether the mean effect-size estimates differ across moderator-based subgroups). This table shows results regarding Type I and Type II error rates collapsing across all cells in the design. The results shown in Table 3 indicate that the three approaches yield virtually identical Type I and Type II error rates. All three approaches had comparable Type I error rates (i.e., .06 for Hunter and Schmidt and Hedges and Olkin and .05 for Aguinis and Pierce). Type II error rates at both $|p_1 - p_2| = .2$ and $|p_1 - p_2| = .4$ did not differ by more than .02 across the three approaches.

Performance of Homogeneity Tests

Homogeneity tests allow meta-analysts to decide whether the primary-level effect-size estimates come from different populations, and, therefore, the search for a moderating effect is warranted. Next, we report results regarding Type I and Type II error rates for each of the three meta-analytic approaches.

Table 4
Type I Error Rates for Homogeneity Tests for the Three Meta-Analytic Approaches

#	k	n_1	n_2	$\rho_1 = \rho_2$	SR (%)	Rel X	Rel Y	H-S (75%)	H-O (Q)	A-P (Q')
1	10							.14	.06	.09
2	50							.03	.11	.21
3	100							.01	.17	.31
4		50	20					.08	.18	.29
5		50	50					.05	.11	.21
6		100	20					.08	.16	.26
7		100	50					.05	.10	.18
8		100	100					.04	.08	.16
9		250	20					.08	.15	.24
10		250	50					.05	.09	.18
11		250	100					.05	.08	.15
12		250	250					.04	.07	.14
13				.1				.06	.09	.10
14				.3				.06	.11	.17
15				.5				.05	.14	.33
16					10			.08	.12	.22
17					50			.05	.11	.21
18					100			.04	.11	.18
19						.6		.06	.11	.22
20						.8		.06	.12	.21
21						1.0		.05	.11	.16
22							.6	.06	.11	.23
23							.8	.06	.11	.21
24							1.0	.06	.12	.17

Note: # = case number; k = number of primary-level studies included in the meta-analysis; n_1 = sample size in moderator-based Subgroup 1; n_2 = sample size in moderator-based Subgroup 2; ρ_1 = X-Y correlation for moderator-based Subgroup 1; ρ_2 = X-Y correlation for moderator-based Subgroup 2; SR = selection ratio; Rel X = reliability in variable X; Rel Y = reliability in variable Y; H-S = Hunter and Schmidt (1990, 2004) approach; H-O = Hedges and Olkin (1985) approach; A-P = Aguinis and Pierce (1998) approach.

Type I error rates. Table 4 shows results regarding Type I error rates for homogeneity tests for all values of each variable manipulated in the design. The level of Type I errors ranged from .01 to .14 for the Hunter and Schmidt procedure (M = .06) and from .06 to .18 for Hedges and Olkin (M = .11). The Aguinis and Pierce approach had higher levels of error for the homogeneity test, with errors ranging from .09 to .33 (M = .20).

Type II error rates. Table 5 shows results regarding Type II error rates for homogeneity tests for $|p_1 - p_2|$ = .2. Results indicate that Type II error rates were lowest for the Aguinis and Pierce approach (i.e., M = .60, range = .38-.83). The Hunter and Schmidt and Hedges and Olkin approaches yielded higher rates (i.e., M = .82, range = .55-.90; M = .68, range = .44-.86, respectively).

Table 5
Type II Error Rates for Homogeneity Tests for $|\rho_1 - \rho_2| = .2$ for the Three Meta-Analytic Approaches

#	k	n_1	n_2	ρ_1, ρ_2	SR (%)	Rel X	Rel Y	H-S (75%)	H-O (Q)	A-P (Q')
1	10							.75	.86	.83
2	50							.84	.66	.57
3	100							.88	.51	.39
4		50	20					.90	.74	.64
5		50	50					.90	.76	.67
6		100	20					.89	.75	.67
7		100	50					.88	.73	.65
8		100	100					.82	.67	.59
9		250	20					.88	.75	.67
10		250	50					.84	.70	.62
11		250	100					.73	.58	.51
12		250	250					.55	.44	.38
13				.1, .3				.83	.69	.66
14				.3, .1				.81	.70	.66
15				.3, .5				.84	.67	.54
16				.5, .3				.82	.66	.53
17					10			.88	.79	.70
18					50			.85	.70	.62
19					100			.74	.54	.48
20						.6		.88	.76	.66
21						.8		.72	.67	.58
22						1.0		.77	.61	.56
23							.6	.88	.76	.64
24							.8	.82	.67	.59
25							1.0	.77	.61	.57

Note: # = case number; k = number of primary-level studies included in the meta-analysis; n_1 = sample size in moderator-based Subgroup 1; n_2 = sample size in moderator-based Subgroup 2; ρ_1 = X-Y correlation for moderator-based Subgroup 1; ρ_2 = X-Y correlation for moderator-based Subgroup 2; SR = selection ratio; Rel X = reliability in variable X; Rel Y = reliability in variable Y; H-S = Hunter and Schmidt (1990, 2004) approach; H-O = Hedges and Olkin (1985) approach; A-P = Aguinis and Pierce (1998) approach.

Table 6 displays results regarding Type II error rates for homogeneity tests for a moderating effect size of $|\rho_1 - \rho_2| = .4$. As expected, because of the increase in the magnitude of the moderating effect from .2 to .4, the pattern of error rates is similar to that displayed in Table 5, but they are smaller in magnitude. Nevertheless, despite a fairly large moderating effect, Type II error rates are near .50 and higher for some conditions. The Aguinis and Pierce approach yielded the lowest mean error rate (i.e., .32), the Hedges and Olkin method was somewhat larger ($M = .38$), and the Hunter and Schmidt approach was the highest ($M = .48$).

Performance of Moderating Effect Tests

Tests of moderating effects involve comparing mean effect-size estimates across moderator-based subgroups. Next, we report results regarding Type I and Type II error rates for each of the three meta-analytic approaches.

Table 6
Type II Error Rates for Homogeneity Tests for $|\rho_1 - \rho_2| = .4$ for the Three Meta-Analytic Approaches

#	k	n_1	n_2	ρ_1, ρ_2	SR (%)	Rel X	Rel Y	H-S (75%)	H-O (Q)	A-P (Q')
1	10							.63	.59	.63
2	50							.46	.30	.24
3	100							.46	.17	.11
4		50	20					.74	.53	.46
5		50	50					.60	.45	.38
6		100	20					.69	.51	.44
7		100	50					.49	.39	.33
8		100	100					.35	.29	.25
9		250	20					.63	.50	.44
10		250	50					.40	.33	.28
11		250	100					.23	.21	.18
12		250	250					.10	.10	.08
13				.1, .5				.49	.37	.33
14				.5, .1				.45	.36	.31
15					10			.63	.53	.46
16					50			.49	.48	.32
17					100			.28	.19	.17
18						.6		.63	.49	.42
19						.8		.45	.34	.29
20						1.0		.34	.26	.24
21							.6	.60	.46	.39
22							.8	.45	.35	.30
23							1.0	.36	.29	.26

Note: # = case number; k = number of primary-level studies included in the meta-analysis; n_1 = sample size in moderator-based Subgroup 1; n_2 = sample size in moderator-based Subgroup 2; ρ_1 = X-Y correlation for moderator-based Subgroup 1; ρ_2 = X-Y correlation for moderator-based Subgroup 2; SR = selection ratio; Rel X = reliability in variable X; Rel Y = reliability in variable Y; H-S = Hunter and Schmidt (1990, 2004) approach; H-O = Hedges and Olkin (1985) approach; A-P = Aguinis and Pierce (1998) approach.

Type I error rates. As Table 3 shows in aggregate, and Table 7 shows for various characteristics of the meta-analyses, Type I error rates for the three methods were very similar. Type I error rates indicate the proportion of instances in which the test concludes incorrectly that the two moderator-based subgroups estimate different population correlation coefficients.

Table 7 shows that Type I error rates for the Hedges and Olkin approach were very close to the .05 preset value for all cases. Specifically, error rates for the Hedges and Olkin approach ranged from .05 to .07 with a mean of .055. The Hunter and Schmidt approach produced similar results, with error rates ranging from .05 to .07 and a mean of .059. The Aguinis and Pierce approach yielded similar error rates, ranging from .04 to .06 with a mean of .050.

Results from the Hedges and Olkin and Hunter and Schmidt approaches correlated at .79 ($p < .0001$). The Aguinis and Pierce approach showed a similarly strong degree of relationship with the Hedges and Olkin (i.e., $r = .82$, $p < .0001$) and the Hunter and Schmidt (i.e., $r = .68$, $p < .0001$) approaches.

Type II error rates. Table 8 shows results regarding Type II error rates for moderating effect tests for $|\rho_1 - \rho_2| = .2$. This table indicates that Type II error rates were similar across all three techniques, ranging from a low of .12 to a high of .69. The mean error rates were .36 for the Hunter and

Schmidt approach, .36 for the Hedges and Olkin approach, and .38 for the Aguinis and Pierce approach. Results yielded by the three approaches were inter-correlated above $r = .84$ ($p < .0001$).

Table 7
Type I Error Rates for Tests of Moderating Effects for the Three Meta-Analytic Approaches

#	k	n_1	n_2	$\rho_1 = \rho_2$	SR (%)	Rel X	Rel Y	H-S (t)	H-O (Q_B)	A-P (Q'_B)
1	10							.06	.06	.05
2	50							.06	.05	.05
3	100							.06	.05	.05
4		50	20					.07	.06	.05
5		50	50					.05	.05	.04
6		100	20					.07	.06	.05
7		100	50					.05	.05	.05
8		100	100					.05	.05	.05
9		250	20					.07	.06	.05
10		250	50					.06	.05	.05
11		250	100					.06	.06	.06
12		250	250					.05	.05	.05
13				.1				.07	.05	.04
14				.3				.06	.05	.05
15				.5				.05	.06	.06
16					10			.06	.05	.05
17					50			.05	.05	.05
18					100			.06	.07	.06
19						.6		.06	.05	.05
20						.8		.06	.06	.05
21						1.0		.06	.05	.05
22							.6	.06	.06	.05
23							.8	.06	.05	.05
24							1.0	.06	.06	.05

Note: # = case number; k = number of primary-level studies included in the meta-analysis; n_1 = sample size in moderator-based Subgroup 1; n_2 = sample size in moderator-based Subgroup 2; ρ_1 = X-Y correlation for moderator-based Subgroup 1; ρ_2 = X-Y correlation for moderator-based Subgroup 2; SR = selection ratio; Rel X = reliability in variable X; Rel Y = reliability in variable Y; H-S = Hunter and Schmidt (1990, 2004) approach; H-O = Hedges and Olkin (1985) approach; A-P = Aguinis and Pierce (1998) approach.

Table 9 displays similar results regarding Type II error rates for moderating effect tests for an effect size of $|\rho_1 - \rho_2| = .4$. As expected, because of the increase in magnitude of the moderating effect from .2 to .4, these error rates are similar across conditions to those displayed in Table 8, but they are smaller in magnitude. The Hedges and Olkin approach had a mean error rate of .13, followed closely by the Hunter and Schmidt and Aguinis and Pierce approaches with error rates of .14 and .16, respectively. Error rates were highly correlated for all three approaches, each comparison being above $r = .83$ ($p < .0001$).

Discussion

The impetus for the present study included three factors. First, meta-analytic tests of hypotheses regarding moderating effects of categorical variables are increasingly pervasive in management and related fields. Also, because meta-analytic tests for categorical moderators differ from those for continuous moderators, there is a specific need to investigate the procedures most widely

used in the organizational sciences. Second, in contrast to primary level researchers, meta-analysts do not seem to have guidelines regarding which technique to use, and under which conditions, to test hypotheses regarding moderating effects of categorical variables. Third, although recent Monte Carlo investigations have been published regarding the performance of meta-analytic techniques, these studies have not been comprehensive in the inclusion of meta-analytic approaches. Consequently, the purpose of the present study was to compare the Hedges and Olkin, the Hunter and Schmidt, and a refinement of the Aguinis and Pierce approaches for estimating moderating effects of categorical variables meta-analytically. This study examined these techniques under a wide range of situations typically encountered by applied psychology and management researchers and assessed the techniques' accuracy in terms of Type I and Type II error rates and their ability to estimate the population effect size. Next, we discuss implications of the present results for theory and conduct of meta-analysis.

Table 9
Type II Error Rates for Tests of Moderating Effects for $|\rho_1 - \rho_2| = .4$ for the Three Meta-Analytic Approaches

#	k	n_1	n_2	ρ_1, ρ_2	SR (%)	Rel X	Rel Y	H-S (t)	H-O (Q_B)	A-P (Q'_B)
1	10							.35	.32	.35
2	50							.05	.05	.08
3	100							.01	.01	.03
4		50	20					.25	.25	.31
5		50	50					.17	.15	.18
6		100	20					.22	.22	.27
7		100	50					.12	.11	.12
8		100	100					.08	.07	.07
9		250	20					.20	.20	.26
10		250	50					.10	.09	.10
11		250	100					.05	.04	.05
12		250	250					.02	.02	.02
13				.1, .5				.13	.13	.17
14				.5, .1				.14	.13	.13
15					10			.22	.21	.26
16					50			.13	.12	.14
17					100			.06	.05	.05
18						.6		.21	.20	.25
19						.8		.12	.11	.13
20						1.0		.08	.07	.08
21							.6	.18	.18	.22
22							.8	.13	.12	.14
23							1.0	.10	.09	.10

Note: # = case number; k = number of primary-level studies included in the meta-analysis; n_1 = sample size in moderator-based Subgroup 1; n_2 = sample size in moderator-based Subgroup 2; ρ_1 = X-Y correlation for moderator-based Subgroup 1; ρ_2 = X-Y correlation for moderator-based Subgroup 2; SR = selection ratio; Rel X = reliability in variable X; Rel Y = reliability in variable Y; H-S = Hunter and Schmidt (1990, 2004) approach; H-O = Hedges and Olkin (1985) approach; A-P = Aguinis and Pierce (1998) approach.

Implications for Theory

The present results lead to several meaningful conclusions and implications for the accumulation of knowledge, theory advancement, and the conduct of meta-analysis in management, applied psychology, and other social sciences. More than a decade ago, J. A. Hall and Rosenthal (1991), writing about meta-analysis, asserted: "If we want to know how well we are doing in the biological,

psychological, and social sciences, an index that will serve us well is how far we have advanced in our understanding of the moderator variables of our field" (p. 447).

Results shown in Tables 1 to 9 suggest that all three meta-analytic methods lead to errors and to a lack of understanding regarding the operation of moderator variables. A significant problem with each of the three techniques we investigated is that they often lead to the erroneous conclusion that there are no moderating effects. Stated differently, although Type I error rates are fairly close to the preset level, Type II error rates are in many conditions quite large (i.e., .50 and larger). Specifically, the overall mean Type II error for homogeneity tests was .50 for the Aguinis and Pierce, .58 for the Hedges and Olkin, and .71 for the Hunter and Schmidt approaches. In addition, overall, Type II error rates increase as range restriction becomes more severe and measurement error increases. Thus, in most cases, and unless there is a very strong a priori theory-based rationale, most meta-analysts may not conduct a moderating effect test in the presence of homogeneity of effect sizes. That is, the first step in a meta-analysis usually includes a test of whether effect-size estimates are homogeneous. Given the uniformly low power for homogeneity tests across approaches, a meta-analyst is likely to conclude incorrectly that the data do not warrant specific subgroup-based comparisons given that effect-size estimates are believed to estimate the same population effect. Given that Statistical Power = 1 – Type II error, the present results imply that the statistical power to detect categorical moderators meta-analytically is well below the recommended .80 level (Cohen, 1988). In short, these results show that in approximately 60% of cases (depending on which approach one implements), a moderating effect test may not even be conducted despite the fact that there is a moderating effect in the population. Although previous work has investigated this statistical power issue (Cornwell & Ladd, 1993; Hedges & Pigott, 2001; Overton, 1998; Sackett, Harris, & Orr, 1986; Spector & Levine, 1987), the present study included a more complete set of situations (e.g., measurement error and range restriction) and comparisons across meta-analytic procedures (i.e., inclusion of the Aguinis & Pierce, 1998, procedure). Our results show that the presence of range restriction and measurement error worsens the statistical power problem. And, unfortunately, the implementation of the more recently proposed Aguinis and Pierce approach does not mitigate the power problem observed in the other two more established approaches.

An implication of these results for theory advancement and the accumulation of knowledge in management and related fields is that meta-analysts may reject correct hypotheses positing the operation of moderator variables. The failure to detect population moderating effects is detrimental to the advancement of management and related fields for several reasons. First, theoretical models including moderating effects may be incorrectly discarded. This type of model misspecification can lead

to serious errors in prediction. Second, a meta-analysis reporting null results is likely to influence researchers to abandon a line of research involving the moderated relationship that was investigated. If the meta-analysis incorrectly reported the null result, this could have serious negative effects, particularly because meta-analytic reviews are often more influential than primary-level studies. It may take several years for researchers to reconsider testing a hypothesis regarding a moderating effect after a meta-analysis has demonstrated (perhaps incorrectly) that the moderating effect does not exist. As a consequence of an incorrect result because of low statistical power, the process of knowledge accumulation in management, applied psychology, and other social sciences can be seriously delayed.

Implications for the Conduct of Meta-Analysis

The present results also have implications for the conduct of meta-analysis. First, the Aguinis and Pierce approach was analytically described, and it has already been cited (e.g., Collins & Holton, 2004; Sharma & Yetton, 2003) and used in published research (e.g., Webber & Donahue, 2001). However, the present study represents the first attempt to investigate empirically the performance of this meta-analytic procedure.

The present results show that the Aguinis and Pierce approach yields superior point estimates as compared to the Hedges and Olkin approach, and the Aguinis and Pierce point estimates are nearly identical to those produced by the Hunter and Schmidt approach (see Table 1). Regarding homogeneity tests, Type I error rates for the Aguinis and Pierce approach were inferior (i.e., further from the preset .05 rate) to those generated by the other two procedures, yet Type II error rates were slightly better than those generated by the other two approaches (see Table 2). Regarding moderating effect tests, the Aguinis and Pierce approach yielded Type I and Type II error rates virtually identical to the other two approaches. In addition, Type I and Type II error rates, both for homogeneity and moderating effect tests, were highly correlated with those generated by the other two approaches. Regarding the effects of increasing the severity of range restriction and measurement error, the Aguinis and Pierce approach yielded higher Type I error rates for homogeneity tests as range restriction (i.e., selection ratios changing from 100% to 10%), and measurement error (i.e., reliabilities changing from 1.0 to .60) became more severe. These Type I error rates were higher than those produced by both the Hunter and Schmidt and Hedges and Olkin approaches (see Table 4). On the other hand, the Aguinis and Pierce approach provided a relative advantage regarding Type II error rates for homogeneity tests as range restriction became more severe (see Tables 5 and 6). Regarding tests of moderating effects, as range restriction and measurement error increase in severity, the performance of the Aguinis and Pierce approach is

nearly identical to the performance of the other two approaches regarding both Type I and Type II error rates (see Tables 7-9). In sum, the Aguinis and Pierce procedure provides an overall relative advantage regarding Type II error rates for homogeneity tests, and this relative advantage is accentuated as range restriction becomes more severe. However, the approach performs at similar or lower levels than the other approaches regarding (a) point estimates, (b) Type I error rates regarding homogeneity tests, (c) Type I and Type II rates regarding moderating tests, and (d) Type I and Type II error rates for both homogeneity and moderating effect tests as measurement error becomes more severe. Thus, the Aguinis and Pierce approach is recommended in general for meta-analyses including strong theory-based hypotheses (for which a Type II error may be more costly than a Type I error) and, in particular, for meta-analytic data sets exhibiting severe levels of range restriction.

Second, point estimates of the population correlation coefficients were most accurate for the Hunter and Schmidt approach. Thus, we recommend that this approach be used to estimate the magnitude of the population moderating effect. We provide this recommendation with the caveat that, collapsing across all cells in the design, the point estimates for the Hunter and Schmidt procedure had an overall mean error of .107 (see Table 1). Thus, the estimate is just that: an estimate. It should not be confused with, or assumed to be identical to, the population or true (i.e., prespecified parameter in the simulation) correlation coefficient.

Third, regarding homogeneity tests, the Hunter and Schmidt approach was slightly superior regarding accuracy in Type I error rates. Alternatively, the Aguinis and Pierce procedure was slightly superior regarding Type II error rates. Thus, we recommend that researchers implement the Hunter and Schmidt approach in the absence of strong theory-based hypotheses regarding moderating effects of categorical variables and the Aguinis and Pierce approach in the presence of strong theory-based hypotheses. Stated differently, the Hunter and Schmidt approach seems to be best regarding post hoc attempts to estimate moderating effects because, in the absence of a moderating effect, this meta-analytic technique is best at holding Type I error rates closer to the preset level. On the other hand, the Aguinis and Pierce approach seems to be best for a priori attempts to test moderating effect hypotheses because it provides relatively better statistical power.

Fourth, results regarding moderating effect tests show that both Type I and Type II error rates are similar across approaches. Thus, there seems to be no advantage to using one approach over the other.

Finally, an examination of Tables 4 to 9 allows for a comparison of the relative performance of the three approaches as influenced by range restriction and measurement error. Regarding Type I error

rates for homogeneity tests, Table 4 shows that as range restriction becomes more severe, the Hunter and Schmidt procedure produces the smallest number of errors. And the Hunter and Schmidt procedure also produces the smallest Type I error rates for homogeneity tests as measurement error becomes more severe (i.e., reliabilities for X and Y decreasing from 1.0 to .6). Regarding Type II error rates for homogeneity tests, Tables 5 to 6 show that the Aguinis and Pierce approach is less negatively influenced as range restriction becomes more severe, and the Aguinis and Pierce procedure is less negatively influenced by increases in measurement error as compared with the Hunter and Schmidt and Hedges and Olkin approaches. Regarding Type I error rates for tests of moderating effects, Table 7 shows that all three approaches are nearly equivalent in how well they are affected by increased levels of severity in range restriction and measurement error, although the Aguinis and Pierce method is either the best or tied for the best in 8 of the 9 comparisons (compared to 4 of 9 for Hedges and Olkin and 1 of 9 for Hunter and Schmidt). And regarding Type II error rates for tests of moderating effects, Tables 8 and 9 show that all three approaches are similarly vulnerable to stringent range restriction and large measurement error conditions, with Hedges and Olkin performing best (or tied for best) in 17 of 18 cases (compared to 8 of 18 for Hunter and Schmidt and 1 of 18 for Aguinis and Pierce). In short, as range restriction and measurement error increase in severity, the Hunter and Schmidt procedure produces the best Type I error rates for homogeneity tests, and the Hedges and Olkin procedure produces the best Type II error rates for homogeneity tests. Nevertheless, all three approaches are similarly vulnerable to Type I and Type II error rates regarding moderating effect tests as range restriction becomes more severe and measurement error increases.

Table 10 provides a rank ordering of the three meta-analytic approaches based on their overall relative performance. Table 11 provides a rank ordering of the approaches' performance based on the effects of range restriction and measurement error. These two tables can help researchers choose a particular meta-analytic approach based on theory development (e.g., the need to decrease Type I error in relation to Type II error) and research design considerations (i.e., the presence of severe range restriction and measurement error).

Table 10
Relative Overall Performance of the Three Meta-Analytic Approaches

Point estimates (see Table 1)	
1. Hunter and Schmidt	
2. Aguinis and Pierce	
3. Hedges and Olkin	
Homogeneity tests (see Table 2)	
Type I error rates	Type II error rates
1. Hunter and Schmidt (.06)	1. Aguinis and Pierce (.46)
2. Hedges and Olkin (.11)	2. Hedges and Olkin (.53)
3. Aguinis and Pierce (.20)	3. Hunter and Schmidt (.65)
Moderating effect tests (see Table 3)	
Type I error rates	Type II error rates
1. Aguinis and Pierce (.05)	1. Hunter and Schmidt (.24)
2. Hedges and Olkin and Hunter and Schmidt (.06)	2. Hedges and Olkin (.25)
	3. Aguinis and Pierce (.27)

Note: Ranks are ordered from best to worst. Mean Type I and Type II error rates are shown in parentheses.

Limitations and Research Needs

Although we have made an effort to simulate a wide range of conditions that have been commonly noted in meta-analysis implications, there are still a number of limitations to our study that highlight the need for further research. First, the present study's data were generated assuming normally distributed scores. Thus, although complying with the normality assumption is common practice in Monte Carlo investigations of meta-analytic methods (e.g., Aguinis, 2001; Aguinis & Whitehead, 1997; Millsap, 1989), we acknowledge that the present study's results may not be generalizable to situations in which this assumption is not tenable (Oswald & Johnson, 1998). Thus, future research could investigate the extent to which the relative performance of the three meta-analytic procedures is affected by the presence of non-normal distributions.

Second, for the sake of simplicity, we chose to simulate a situation involving a binary moderator variable. We made this choice because the vast majority of published meta-analytic tests of categorical moderators include moderators with only two levels (e.g., male vs. female, low vs. high, ethnicity coded as majority vs. nonmajority). In addition, this choice is the simplest and most parsimonious. Moreover, we cannot think of any compelling reason why the present results regarding a binary moderator would not generalize to categorical moderators with more than two levels. Nevertheless, we acknowledge that our simulation emulated a situation involving a categorical moderator with two levels only. Thus, future research could examine the extent to which results of the present study generalize to meta-analytic investigations of moderator variables having more than two levels.

Table 11
Relative Performance of the Three Meta-Analytic Approaches as Affected by Increasing Levels of Range Restriction and Measurement Error

Effects of range restriction		
Homogeneity tests (see Tables 4-6)		
Type I error rates	Type II error rates	
1. Hunter and Schmidt	1. Aguinis and Pierce	
2. Hedges and Olkin	2. Hedges and Olkin	
3. Aguinis and Pierce	3. Hunter and Schmidt	
Moderating effect tests (see Tables 7-9)		
Type I error rates[a]	Type II error rates[a]	
1. Aguinis and Pierce	1. Hedges and Olkin	
2. Hedges and Olkin	2. Hunter and Schmidt	
3. Hunter and Schmidt	3. Aguinis and Pierce	
Effects of measurement error		
Homogeneity tests (see Tables 4-6)		
Type I error rates	Type II error rates[b] $\lvert\rho_1 - \rho_2\rvert = .2$	Type II error rates[b] $\lvert\rho_1 - \rho_2\rvert = .4$
1. Hunter and Schmidt	1. Aguinis and Pierce	1. Hunter and Schmidt
2. Hedges and Olkin	2. Hunter and Schmidt	2. Aguinis and Pierce
3. Aguinis and Pierce	3. Hedges and Olkin	3. Hedges and Olkin
Moderating effect tests (see Tables 7-9)		
Type I error rates[a]	Type II error rates[a]	
1. Aguinis and Pierce	1. Hunter and Schmidt	
2. Hedges and Olkin	2. Hedges and Olkin	
3. Hunter and Schmidt	3. Aguinis and Pierce	

Note: Ranks are ordered from best to worst.
a. Although a rank order is provided, error rates are virtually identical across the three approaches.
b. Results are reported separately for each effect-size condition because of a swapping of rank orders between the Aguinis and Pierce and Hunter and Schmidt approaches.

Conclusion

Some authors have issued warnings regarding the fallibility of meta-analysis in general and validity generalization (i.e., Hunter and Schmidt approach) in particular (Bobko & Stone-Romero, 1998; Russell & Gilliland, 1995) and concluded that meta-analysis is no panacea. We echo the concerns expressed by these authors. Meta-analysis is just another data-analytic technique and is no substitute for good theory. Much like any other data-analytic technique, meta-analysis can lead to incorrect conclusions even when all procedures are correctly implemented from a technical standpoint. The present results show that, under some conditions, the probability of detecting the moderating effect of a categorical variable can be as low as .10 (i.e., Type II error rates as high as .90).

The present results provide the following guidelines regarding the conduct of meta-analysis. First, results show that the Hunter and Schmidt approach yields the most accurate estimate for the moderating effect magnitude, and, therefore, it should be used for point estimation. Second, regarding homogeneity tests, the Hunter and Schmidt approach provides a slight advantage regarding Type I error

rates, and the Aguinis and Pierce approach provides a slight advantage regarding Type II error rates. Thus, the Hunter and Schmidt approach is best for situations when theory development is at the initial stages and there are no strong theory-based hypotheses to be tested (i.e., exploratory or post hoc testing). Alternatively, the Aguinis and Pierce approach is best when theory development is at more advanced stages (i.e., confirmatory and a priori testing). Third, all three approaches yield similar overall Type I and Type II error rates for moderating effect tests, so there are no clear advantages of using one approach over the other. Fourth, the Hunter and Schmidt procedure is the least affected by increasing levels of range restriction and measurement error regarding homogeneity test Type I error rates, and the Aguinis and Pierce homogeneity test Type II error rates are least affected by these research design conditions (in the case of measurement error, this is particularly true for effect sizes around .2). Thus, the choice of one approach over the other needs to consider the extent to which range restriction and measurement error are research design issues present in the meta-analytic database to be analyzed. And finally, all three approaches are effectively equally vulnerable to Type I and Type II error rates for homogeneity and moderating effect tests as range restriction becomes more severe and measurement error increases, with perhaps a slight advantage (around .01) to the Aguinis and Pierce approach. In closing, we hope these recommendations will help researchers choose meta-analytic techniques based on theory and research design considerations as opposed to habit, the availability of a user-friendly computer program, or usage trends in specific topic areas.

References

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

Aguinis, H. (2004). *Regression analysis for categorical moderators*. New York: Guilford.

Aguinis, H. (2007). *Performance management*. Upper Saddle River, NJ: Pearson Prentice Hall.

Aguinis, H., Beaty, J. C., Boik, R. J., & Pierce, C. A. (2005). Effect size and power in assessing moderating effects of categorical variables using multiple regression: A 30-year review. *Journal of Applied Psychology*, *90*, 94-107.

Aguinis, H., Boik, R. J., & Pierce, C. A. (2001). A generalized solution for approximating the power to detect effects of categorical moderator variables using multiple regression. *Organizational Research Methods*, *4*, 291-323.

Aguinis, H., Petersen, S. A., & Pierce, C. A. (1999). Appraisal of the homogeneity of error variance assumption and alternatives to multiple regression for estimating moderating effects of categorical variables. *Organizational Research Methods*, *2*, 315-339.

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

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.

Bobko, P., & Stone-Romero, E. F. (1998). Meta-analysis may be another useful research tool, but it is not a panacea. In *Research in personnel and human resources management* (Vol. 16, pp. 359-397). Stamford, CT: JAI.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Collins, D. B., & Holton, E. F. (2004). The effectiveness of managerial leadership development programs: A meta-analysis of studies from 1982 to 2001. *Human Resource Development Quarterly*, *15*, 217-248.

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

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

Dalton, D. R., Daily, C. M., Johnson, J. L., & Ellstrand, A. E. (1999). Number of directors and financial performance: A meta-analysis. *Academy of Management Journal*, *42*, 674-686.

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

Fuller, J. B., & Hester, K. (1999). Comparing the sample-weighted and unweighted meta-analysis: An applied perspective. *Journal of Management*, *25*, 803-828.

Gerstner, C. R., & Day, D. V. (1997). Meta-analytic review of leader-member exchange theory: Correlates and construct issues. *Journal of Applied Psychology*, *82*, 827-844.

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

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

Hall, S. M., & Brannick, M. T. (2002). Comparison of two random-effects methods of meta-analysis. *Journal of Applied Psychology*, *87*, 377-389.

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.

Hough, S. L., & Hall, B. W. (1994). Comparison of the Glass and Hunter and Schmidt meta-analytic techniques. *Journal of Educational Research*, *87*, 292-296.

Huffcutt, A. I., Roth, P. L., & McDaniel, M. A. (1996). A meta-analytic investigation of cognitive ability in employment interview evaluations: Moderating characteristics and implications for incremental validity. *Journal of Applied Psychology*, *81*, 459-473.

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

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage.

Johnson, B. T. (1993). *DSTAT 1.10: Software for the meta-analytic review of research literatures*. Hillsdale, NJ: Lawrence Erlbaum.

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

Marin-Martinez, F., & Sanchez-Meca, J. (1998). Testing for dichotomous moderators in meta-analysis. *The Journal of Experimental Education*, *67*, 69-81.

Millsap, R. E. (1989). Sampling variance in the correlation coefficient under range restriction: A Monte Carlo study. *Journal of Applied Psychology*, *74*, 456-461.

Neter, J.,Wasserman,W., & Whitmore, G. A. (1988). *Applied statistics* (3rd ed.). Newton, MA: Allyn & Bacon.

Oswald, F. L., & Johnson, J. W. (1998). On the robustness, bias, and stability of statistics from meta-analysis of correlation coefficients: Some initial Monte Carlo findings. *Journal of Applied Psychology*, *83*, 164-178.

Overton, R. C. (1998). A comparison of fixed-effects and mixed (random-effects) models for meta-.

Phillips, J. M. (1998). Effects of realistic job previews on multiple organizational outcomes: A meta-analysis. *Academy of Management Journal*, *41*, 673-690.

Richard, O. C. (2000). Racial diversity, business strategy, and firm performance: A resource-based view. *Academy of Management Journal*, *43*, 164-177.

Russell, C. J., & Gilliland, S. W. (1995). Why meta-analysis doesn't always tell you what the data really mean. *Journal of Management*, *21*, 813-831.

Russell, C. J., Settoon, R. P., McGrath, R. N., Blanton, A. E., Kidwell, R. E., Lohrke, F. T., et al. (1994). Investigator characteristics as moderators of personnel selection research: A meta-analysis. *Journal of Applied Psychology*, *79*, 163-170.

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

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

Sharma, R., & Yetton, P. (2003). The contingent effects of management support and task interdependence on successful information systems implementation. *MIS Quarterly*, *27*, 533-555.

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

Steel, P. D., & Kammeyer-Mueller, J. D. (2002). Comparing meta-analytic moderator estimation techniques under realistic conditions. *Journal of Applied Psychology*, *87*, 96-111.

Sturman, M. C. (2003). Searching for the inverted U-shaped relationship between time and performance: Metaanalyses of the experience/performance, tenure/performance, and age/performance relationships. *Journal of Management*, *29*, 609-640.

Sturman, M. C., Cheramie, R. A., & Cashen, L. H. (2005). The impact of job complexity and performance measurement on the temporal consistency, stability, and test-retest reliability of employee job performance ratings. *Journal of Applied Psychology*, *90*, 269-283.

Viswesvaran, C., & Sanchez, J. I. (1998). Moderator search in meta-analysis: A review and cautionary note on existing approaches. *Educational and Psychological Measurement*, *58*, 77-87.

Webber, S. S., & Donahue, L. M. (2001). Impact of highly and less job-related diversity on work group cohesion and performance: A meta-analysis. *Journal of Management*, *27*, 141-162.

Wolf, F. M. (1986). *Meta-analysis: Quantitative methods for research synthesis*. Beverly Hills, CA: Sage.