Organizational Research Methods

http://orm.sagepub.com

Meta-Analysis: A Constantly Evolving Research Integration Tool

Frank Schmidt

Organizational Research Methods 2008; 11; 96 originally published online Jul 23, 2007;
DOI: 10.1177/1094428107303161

The online version of this article can be found at:
http://orm.sagepub.com/cgi/content/abstract/11/1/96

Published by:
SAGE Publications
http://www.sagepublications.com

On behalf of:

The Research Methods Division of The Academy of Management

Additional services and information for *Organizational Research Methods* can be found at:

Email Alerts: http://orm.sagepub.com/cgi/alerts

Subscriptions: http://orm.sagepub.com/subscriptions

Reprints: http://www.sagepub.com/journalsReprints.nav

Permissions: http://www.sagepub.com/journalsPermissions.nav

Citations (this article cites 25 articles hosted on the
SAGE Journals Online and HighWire Press platforms):
http://orm.sagepub.com/cgi/content/refs/11/1/96

Downloaded from http://orm.sagepub.com at The University of Iowa Libraries on December 19, 2007
© 2008 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.

Organizational Research Methods
Volume 11 Number 1
January 2008 96-113
10.1177/1094428107303161
© 2008 Sage Publications
http://orm.sagepub.com
hosted at
http://online.sagepub.com

Meta-Analysis

A Constantly Evolving Research Integration Tool

Frank Schmidt
University of Iowa

During the past 30 years, meta-analysis has been an indispensable tool for revealing the hidden meaning of our research literatures. The four articles in this special section on meta-analysis illustrate some of the complexities entailed in meta-analysis methods. Although meta-analysis is a powerful tool for advancing cumulative knowledge, researchers can be confused by the complicated issues involved in the methodology. Each of these four articles contributes both to advancing this methodology and to the increasing complexities that can befuddle researchers. In these comments, the author attempts to clarify both of these aspects and provide a perspective on the methodological issues examined in these articles.

Keywords: *meta-analysis; statistical artifacts; parameter estimation; fixed and random models*

Ever since the publication of the first article on meta-analysis in the literature 30 years ago (Schmidt & Hunter, 1977), my colleagues and I, along with many other researchers, have been advancing refinements and improvements in an attempt to increase accuracy and usefulness of this methodology (e.g., see Hunter & Schmidt, 2004). The articles in this feature topic are a continuation of this tradition.

The Wood (in press) Article

This article presents a step-by-step procedure for detecting duplicate studies when compiling studies for a meta-analysis. Duplicate studies are defined as studies based on the same data set and hence not statistically independent. Meta-analysts have always attempted to identify and eliminate such studies using informal procedures and subjective judgment. (Using such procedures, my colleagues and I have detected some such duplicate studies in the personnel selection and related literatures in our work.) The contribution of this article is that it presents a carefully delineated set of steps for doing this (summarized in Wood's Figure 1). Although this procedure does not eliminate the need for subjective judgment, it does systemize the decision process, making it more objective and probably more accurate. It may also increase agreement among meta-analyses conducted by

Author's Note: The author would like to thank Huy Le and In-Sue Oh for valuable comments on an earlier version of this article. Correspondence concerning this article should be addressed to Frank L. Schmidt, Department of Management, Henry B. Tippie College of Business, University of Iowa, Iowa City, IA 52242; e-mail: frank-schmidt@uiowa.edu.

Downloaded from http://orm.sagepub.com at The University of Iowa Libraries on December 19, 2007
© 2008 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.

different researchers. To my knowledge, this is the first article to propose a systematic procedure for the detection of duplicate studies.

It is important to be clear that the problem of nonindependent studies is not a problem specific to meta-analysis. There is nothing about meta-analysis per se that creates this problem. If the traditional alternative to meta-analysis—the narrative review—is used, duplicate studies are just as much a problem for that method of reviewing literatures. The problem of duplicate studies—such as the problem of publication bias or availability bias (Rothstein, Sutton, & Borenstein, 2005)—is a problem related to the broader issue of attainment of accurate cumulative knowledge in science (Hunter & Schmidt, 2004, chap. 13). Meta-analysis is one tool that helps to attain this objective, but not the only one. True experiments (with random assignment), quasiexperimental studies, confirmatory factor analysis, path analysis, and structural equation modeling are examples of other valuable tools. The focus of meta-analysis is on the accurate and precise calibration of relationships among variables and constructs (including moderated relationships). Once these relationships have been calibrated, the focus can move to theory construction and theory testing. At that point, a wider variety of methods become relevant.

If the Wood procedure is used and if the result is the detection of duplicate studies, what is to be done with these studies? Wood discusses a number of alternatives and in the end recommends that the duplicate studies be combined into a single effect size estimate and entered into the meta-analysis as a single study. I concur in this recommendation. This solution salvages the information contained in the duplicate studies, while at the same time preserving statistical independence among the entries into the meta-analysis. This is the same recommendation made by Hunter and Schmidt (2004, chap. 10) for the case of ‘‘conceptual replication.’’ In conceptual replication, there are multiple measures of the same construct taken on the same sample in the same study. Hunter and Schmidt recommend that such measures be combined to produce a single effect size estimate, so that such a sample enters the meta-analysis only once, thus avoiding a violation of the assumption of statistical independence. This is the same principle advocated in the Wood article for dealing with duplicate studies.

Why do research literatures contain duplicate studies? Probably because of the greatly increased pressure to publish today, especially among untenured faculty seeking tenure. This is the same pressure that tempts some researchers to go further and fake data, a problem that has appeared in the biological and physical sciences and has been widely publicized in the media. The result is that there are potentially serious data quality problems in some scientific literatures. The sciences in general have yet to produce really good solutions to these problems. We are fortunate that there has to date been no indication that we have this problem in our literatures.

It is perhaps useful to address the general statistical principles in connection with duplicate studies. If the number of studies in the meta-analysis is large, the statistical expectation is that inclusion of duplicate studies will have little impact on the estimate of the mean effect size. However, the statistical expectation is that duplicate studies will inflate the estimate of the variance of population parameters (Hunter & Schmidt, 2004), because the lack of independence results in an overestimation of total sample size, leading to underestimation of sampling error variance. The result is an undercorrection for sampling error and an overestimation of the standard deviation of population effect sizes (SD_ρ or

Downloaded from http://orm.sagepub.com at The University of Iowa Libraries on December 19, 2007
© 2008 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.

SD_δ).[1] In meta-analyses with a small number of studies, the story can be very different. The duplicate studies may by chance skew the estimated mean effect size, and by chance the sampling error may be equal to or even less than that expected under conditions of independence. Such a departure from statistical expectation is called second-order sampling error (Hunter & Schmidt, 2004, chap. 9). This is important because many published meta-analyses are based on relatively small numbers of studies. Thus, the detection and proper handling of duplicate studies can be important.

The Kisamore and Brannick (in press) Article

The focus of this article is on the distinction between fixed effects (FE) and random effects (RE) models in meta-analysis. This is a critically important topic because there is a strange anomaly in the behavioral and social sciences today. As explained in more detail later, the anomaly is the fact that although the FE models are almost never appropriate, the majority of published meta-analyses in education and psychology in general and some related areas have been based on FE models. This is not the case in industrial and organizational psychology literature and certain other literatures such as organizational behavior and business strategy.

The first study in this article shows that when there is real variation in population parameters across studies, the FE model not only cannot detect this variation, but also produces confidence intervals (CIs) that are erroneously narrow. That is, the FE model greatly underestimates the amount of uncertainty in our estimated mean value. Of course, the FE model produces accurate results when there is zero variance across studies in the population values (here the δ values corresponding to the d statistic). The problem is that there are few if any study sets that meet this condition (Hunter & Schmidt, 2000). On the other hand, RE models produce accurate results in both cases: when there is and is not variation in study population values. Study 1 of this article illustrates these points quite well.

There is, however, something in the Table 3 results for study 1 that does not appear quite right. Simulation studies have shown that both the Hedges–Vevea (HV) and Schmidt–Hunter (SH) RE models for the d statistic produce accurate estimates (e.g., Sanchez-Meca & Marin-Martinez, 1998). The values produced by these two RE models are nearly identical in Table 3—and are considerably different from the indicated correct values (the values shown in parenthesis). The authors state that the reason for this is "the small number of studies included in the simulated meta-analysis," and this explanation is correct. Most simulation studies are based on many simulated samples from the underlying population, and the results indicate how accurate the procedure is on average (i.e., in the long run). However, each of the Kisamore-Brannick simulations (one for FE and one for RE data) were based on only one sample of 10 studies from the underlying populations. As a result, there is a huge amount of second-order sampling error in the results and so they are not informative as to the accuracy of the variance estimates produced by the two RE models or the estimates of the mean produced by both RE and FE models. In fact, because it is based on a single simulated sample, many researchers would probably maintain that this procedure does not meet the definition of a simulation study.

Downloaded from http://orm.sagepub.com at The University of Iowa Libraries on December 19, 2007
© 2008 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.

In the second study, the authors reanalyze two published meta-analyses on the Pygmalion effect, one of which (Kierein & Gold, 2000) had used a FE model and one of which (McNatt, 2000) had used a RE model. The authors' reanalysis applied the RE model to both data sets. The results clearly show that it was an error for Kierein and Gold to have used the FE model. Their use of the FE model resulted in an estimate of SD_{δ} of zero, when the actual value was approximately .77! In addition, the FE model produced CIs around the estimated mean values that were erroneously narrow. The authors' bottom line recommendation is that meta-analysts should always use RE models. I strongly concur with this recommendation. It was the recognition that FE models are virtually never appropriate that led John Hunter and me to present only RE models, starting in our earliest publications on meta-analysis (e.g., Hunter, Schmidt, & Jackson, 1982; Schmidt & Hunter, 1977). We have never presented or applied a FE meta-analysis model.

As noted above, FE models lead to erroneously narrow CIs around the mean effect size estimate. How serious is this underestimation of the CI width? My colleagues and I (Schmidt, Oh, & Hayes, 2006) used two different RE methods to reanalyze the data from five FE-based meta-analysis publications in *Psychological Bulletin* that included a total of 54 separate meta-analyses. We found that the average level of underestimation of the width of the CIs by the FE models was 55%. That is, the FE CIs are on average less than half as wide as the actual CIs. This amounts to a very serious overestimation of the degree of certainty of the mean effect sizes. Hence, the use of FE models instead of RE models is not a mere technicality. It leads to major errors.

How common is the use of the FE model in research? Kisamore and Brannick (2008) tabulated the meta-analyses in *Psychological Bulletin* during the 2002 calendar year and found that the majority used the FE model. My colleagues and I (Schmidt et al., 2006) conducted a similar survey for this same journal, covering the much longer period between 1978 and 2006. Of the 169 meta-analyses that could be classified, 129 (or 76%) used only FE methods. Most of these—71%—employed the Hedges and Olkin (1985) FE model. In light of the fact that *Psychological Bulletin* is the premier review journal for the field of psychology in general, it is easy to see that this is a serious problem that is retarding the creation of cumulative knowledge in psychology.

However, as noted above, the picture in industrial/organizational (I/O) psychology is quite different: In the top I/O journals, only about 6% to 7% of meta-analyses use the FE model (Hunter & Schmidt, 2004, pp. 24-26). Although explicit tabulations have not been conducted to my knowledge, my examination of the literature in organizational behavior and business strategy suggests that practice in these areas is similar to that in I/O psychology.

Despite the critical importance of a clear understanding of the implications of the FE–RE distinction for accurate cumulative knowledge, there is much confusion and misunderstanding on this issue. This is reflected in the widespread use of FE models. Kisamore and Brannick (2008) have done a good job of presenting many essential aspects of this distinction, but because of the importance of this issue, additional information is likely to be useful. Hedges and Vevea (1998) and Overton (1998) pointed out that the choice of a FE or RE model depends on type of inference that is the goal of the meta-analyst. If the goal is to draw conclusions that are limited to the set of studies at hand and the meta-analyst does not desire to generalize beyond his or her particular

Downloaded from http://orm.sagepub.com at The University of Iowa Libraries on December 19, 2007
© 2008 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.

set of studies, the FE model can be used when population parameters vary and when they do not. Hedges and Vevea refer to this as conditional inference.

The usual goal of research, however, is generalizable knowledge (Toulmin, 1961), which requires generalization beyond the current set of studies to other similar studies that have been or might be conducted. Hedges and Vevea refer to this as unconditional inference. Within this broader objective, the FE model is appropriate only when population parameters do not vary. When population parameters vary, a RE model is required for unconditional inference (i.e., the inference of cumulative knowledge; Field, 2005; Hedges & Vevea, 1998; Raudenbush, 1994). It is important to note that it is typically not possible to know whether the population parameters do or do not vary prior to conducting the meta-analysis. Thus, it would appear to be prudent to always employ the RE model. The FE model is a special case of the RE model, and if population parameters actually do not vary in a particular meta-analysis, the results produced by the RE model are in expectation the same as those that would be produced by the FE model.

As noted above, the objective in meta-analysis is ordinarily to make inferences about a wider population of studies; that is, to draw conclusions that can be generalized beyond the specific set of studies included in the meta-analysis. If this is not the case and the researcher's purpose is only to reach conclusions limited to the specific set of studies in the meta-analysis, the FE model does not underestimate the standard error and the resulting CIs are not too narrow. This follows from the fact that in this case there is no sampling error in the sampling of study population parameters, because the set of studies at hand is not viewed as a sample of a larger number of studies that might exist or could be conducted (Hedges & Vevea, 1998; Hunter & Schmidt, 2000; Overton, 1998; Raudenbush, 1994). In this case, generalization of conclusions is only to a hypothetical set of studies that is identical to the study set at hand except for simple sampling error; that is, to a set of studies with exactly the same study population parameter values, study for study, and differing only in the sampling of participants (usually people) within studies.

Schulze (2004, pp. 38, 195) stated that it is difficult for a meta-analyst to decide whether his/her purpose is this limited generalization and also difficult for a reader of the meta-analysis to evaluate such a decision and that this creates difficulties in interpreting FE results when $S_\delta^2 > 0$ or $S_\rho^2 > 0$. More importantly, it has been pointed out that such conclusions are of limited scientific value (Hedges & Vevea, 1998; Hunter & Schmidt, 2000; National Research Council, 1992; Overton, 1998; Schulze, 2004). The goal of science is cumulative knowledge and cumulative knowledge is generalizable knowledge (Bechtel, 1988; Phillips, 1987; Toulmin, 1961). Researchers are interested in general principles, not in describing a particular set of studies. Hence, it would appear that the FE model would rarely be appropriate for most research purposes.

The National Research Council (1992, p. 147) stated that fixed effects models "tend to understate actual uncertainty" in research findings and recommended "an increase in the use of random effects models in preference to the current default of fixed effects models" (p. 2; see also pp. 185-187). Others have also cautioned that when the goal is generalizable knowledge, use of FE models can lead to inflated Type I error rates and erroneously narrow confidence intervals (e.g., Field, 2003; Hedges, 1994; Hedges & Vevea, 1998; Hunter & Schmidt, 2000; Overton, 1998; Raudenbush, 1994; Rosenthal, 1991).

Downloaded from http://orm.sagepub.com at The University of Iowa Libraries on December 19, 2007
© 2008 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.

Hedges and Vevea (1998) stated that although there is no statistical (sampling) foundation or justification for generalizing FE findings beyond the specific studies in the meta-analysis, there can be, by analogy with the practices of some primary researchers using analysis of variance (ANOVA) in experiments, an extra statistical or judgment-based basis for such wider generalization (pp. 487-488). They proposed that just as primary researchers using fixed effects ANOVA designs in experiments sometimes generalize their conclusions beyond the specific fixed levels of treatment included in their experiments, so also could meta-analysts using FE models, based on the subjective judgment that new studies will be "sufficiently similar" to those in the meta-analysis. In ANOVA, a FE design is one in which all levels of the treatment that are of interest are included in the design, whereas a RE model in ANOVA is one in which only a sample of treatment levels of interest is included in the study. It was by analogy with this distinction in ANOVA that Hedges and Olkin (1985, p. 149) originally labeled the two different models in meta-analysis as FE and RE models (Hedges & Vevea, 1998). Hence, in FE meta-analysis models, the studies included in the meta-analysis are assumed to constitute the entire universe of relevant studies, whereas in RE models the studies are taken to be a sample of all possible studies that might be conducted or might exist on the subject. However, the National Research Council (1992, pp. 46, 139) report indicates that there are problems with this analogy:

> The manner in which the terms "fixed effects" and "random effects" are used in the meta-analysis literature is somewhat different from the classical definitions used in other techniques of statistics such as analysis of variance, where "fixed effects" is the term required to deny the concept of a distribution of the true effects, $\delta_1 \ldots \delta_K$, and "random effects" supposes that the δ_i are sampled from a population and therefore have a distribution. (National Research Council, 1992, p. 46)

An example might help to clarify the meaning of this National Research Council statement. A study of the effects of training time on job performance might include zero training time, 10 hr of training time, and 20 hr of training time. In FE ANOVA, treatments are fixed at these levels and these levels are considered the only ones of interest. In the FE ANOVA, the idea that there is a naturally occurring distribution of training times or potential training times is explicitly denied. This is different from the FE model in meta-analysis in two ways. First, in meta-analysis the researcher does not specify (or fix) the parameter values (ρ_i or δ_i) in the individual studies included in the meta-analysis; instead, these values are accepted as they happen to be sampled in the set of studies at hand. That is, they are observed and not manipulated. The second difference results from the first: Because the researcher does not fix the parameter values included in the studies but accepts them as they happen to have occurred, there appears to be no basis or rationale for postulating or assuming that these parameter values do not have a distribution across studies, which is the key assumption of the FE model in ANOVA. This is the reason why the National Research Council (1992) report rejected the analogy between FE models in ANOVA and FE models in meta-analysis.

However, had the National Research Council accepted this analogy at the conceptual level, this would still have left open the question of whether the broader generalizations

Downloaded from http://orm.sagepub.com at The University of Iowa Libraries on December 19, 2007
© 2008 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.

sometimes made by researchers from fixed effects ANOVA-based experiments are justified. The fact that experimenters sometimes make such generalizations cannot be viewed as a justification (Schulze, 2004). As Hedges and Vevea (1998) pointed out, this practice has no statistical foundation and is based only on subjective judgment. The National Research Council (1992) report concluded that unless population parameters actually do not vary, FE models will yield CIs that are too narrow (and inflated Type I error rates) when there is any generalization to studies beyond the specific ones included in the meta-analysis. This is also the conclusion of Field (2001, 2003, 2005), Hunter and Schmidt (2000), Overton (1998), Schulze (2004), and others.

In our research (Schmidt et al., 2006), we found that authors employ the FE model even when their own data indicates it is inappropriate. Hedges and Olkin (1985) specify that the FE model should be applied only if the homogeneity test (*Q* statistic) is nonsignificant, indicating that this test can detect no evidence of heterogeneity of population values. However, in 85% of the 52 FE published meta-analyses that we examined the *Q* statistic was statistically significant. Yet the authors nevertheless applied the FE model.

As with the FE model, there are potential conceptual problems associated with use of the RE model. In that model, the studies in the meta-analysis are viewed as a sample from a larger universe of studies that exist or could be conducted. Hedges and Vevea (1998) pointed out that this larger universe is often poorly defined and ambiguous in nature. However, Schulze (2004, pp. 40-41) noted that this is not a problem specific to meta-analysis or RE models in meta-analysis but one that characterizes virtually all samples used in primary and other research. Rarely in research is the target population of participants fully enumerated and delimited; in fact, data sets used frequently consist of something close to convenience samples (i.e., a set of participants for whom it was possible to obtain data). Viewed in this light this problem appears less serious. Another potential problem with RE models is the fact that in the estimation of the between-study parameter variance (S^2_δ or S^2_ρ) the number of data points is the number of studies. Hence if the number of studies is small, estimation of this quantity can have less than sterling accuracy (Hedges & Vevea, 1998; Hunter & Schmidt, 1990, 2004; Raudenbush, 1994). Kisamore and Brannick (2008) also noted this problem. One implication of this is that estimates of between-study parameter variance from RE models should be considered only approximate when the number of studies is small.

The Aguinis, Sturman, and Pierce (in press) Article

The focus of this article is on the use of meta-analysis to detect and calibrate categorical moderators. The effectiveness of three different approaches to doing this is examined via the most heroic computer simulation effort that I have ever seen; truly a momentous amount of work went into these extensive simulations! The overall conclusion is that, in general, the Hunter–Schmidt (HS; Hunter & Schmidt, 2004; Hunter, Schmidt, & Jackson, 1982) and the Aguinis–Pierce (AP; Aguinis & Pierce, 1998) approaches are about equally accurate. The Hedges–Olkin (HO; Hedges & Olkin, 1985) approach is much less accurate, but this is because it fails to include any correction for measurement error or range restriction. That this procedure would be inaccurate could be known a priori.

Downloaded from http://orm.sagepub.com at The University of Iowa Libraries on December 19, 2007
© 2008 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.

The AP procedure is essentially the HO approach with corrections for measurement error and range restriction added. The HO approach of which the AP procedure is a modification is that for the *d* statistic. The assumption in this study is that the original data from the individual studies is in the form of correlations. After correction for measurement error and range restriction these correlations are transformed into *d* values, the meta-analysis is run on the *d* values, and the final results are then back-transformed into the correlation (*r*) metric. It is these final results that are then interpreted.

In real-world meta-analytic data, there can be more than one moderator. For example, if sex is a moderator and we split out the studies into those conducted on males and those conducted on females, we might find that the studies in each group are not homogenous. That is, within each group there can be additional moderators—for example, if age were a second moderator it would operate within both the male and female studies. This is why the ideal search for moderators using meta-analysis is a hierarchical meta-analysis, in which all possible combinations of potential moderators are examined (Hunter & Schmidt, 2004, pp. 424-426). In the simulated data used in the Aguinis et al. study here, there was always only a single (dichotomous) moderator. This means that once we split out on that moderator, all within-group variance is due solely to sampling error and other artifacts. This explains why the AP model could be as accurate as the HS model once the single existing moderator had been identified. The AP model, like the particular HO model from which it is adapted, is an FE model, whereas the HS model is an RE model. (See the Kisamore and Brannick article [2008] and my discussion of that article above; technically, the AP model is a "mixed effects" model [Overton, 1998]—that is, one that assumes that once the moderator is identified, the within-group data are homogeneous.) Hence, if there had been more than one moderator, as in the example above, the HS RE model would still have been accurate within each of the study subgroups, whereas the AP FE model would not have been. This follows from the fact that any FE model assumes that all variance across studies is because of sampling error, which is the case in these simulated data once the single moderator is factored out. But this need not be the case in real data. So this is a limitation of the AP procedure. What is needed is a RE version of the AP procedure, which would allow users to be confident of the accuracy of the results in real data in which there may be more than one moderator.

In addition to measurement error, the Aguinis et al. simulation studies took into account range restriction. However, the simulation assumed direct range restriction, which rarely occurs in real data (Thorndike, 1949). Most range restriction is indirect. Hunter, Schmidt, and Le (2006) presented a new procedure for correcting for indirect range restriction in meta-analysis (see also Schmidt, Oh, and Le, 2006) that can be applied when the information required for the traditional indirect range restriction correction is not available. This procedure has been shown via simulation studies to be more accurate than use of the correction for direct range restriction (Le & Schmidt, 2006). It would be interesting to see an examination of the relative accuracy of the AP and HS procedures under the more realistic conditions of indirect range restriction. The programs contained in the Schmidt and Le (2004) program package incorporate this correction for indirect range restriction, but only for the HS meta-analysis methods.

The Aguinis et al. study clearly shows that, within the data limitations of the simulation discussed above, the AP procedure is quite accurate. This is perhaps somewhat surprising,

Downloaded from http://orm.sagepub.com at The University of Iowa Libraries on December 19, 2007
© 2008 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.

given that one step in that procedure as applied in this case is really unfounded logically. This is the step in which the correlations are transformed to *d* values. Consider a hypothetical example. Suppose the moderator question is: Does the sex of salespeople moderate the correlation between extraversion and sales success (measured by amount sold)? Pearson correlations are available computed solely on men and solely on women salespeople. The first step in the AP procedure is to correct these correlations for measurement error and range restriction. The second step is to convert these corrected *r* values to *d* values, and herein lies the problem. The *d* statistic is by definition the standardized difference between two groups (i.e., the difference between the means of two groups divided by the pooled within group standard deviation). The "*d* values" in our example do not reflect the difference between two groups, because the original correlation is that between two continuous measures within a single group (either the male group or the female group). Hence, these values are logically not values of the *d* statistic.

Now, it is well known that correlations can be transformed to *d* values and vice versa—but this is true only when the correlation is a point biserial correlation (a special form of the Pearson *r*) between group membership and some continuous variable (Hunter & Schmidt, 2004, chap. 7). An example would be the correlation between the dichotomous variable of sex and continuous scores on a job satisfaction measure. Another example would be the correlation between membership in the experimental vs. the control group and the continuous dependent variable. In these cases, the correlation is a reflection of group differences and can be transformed to a meaningful *d* statistic (i.e., a *d* statistic that reflects the size of the difference between the two groups). This is not the situation in many cases—including those examined in the Aguinis et al. article. So in the AP procedure we have a step that does not make sense logically, and yet in the end, when the resulting meta-analysis results are transformed back to the *r* metric, they are still accurate.

It should be noted that the *d* to *r* conversion is not needed or applied in the AP procedure if study outcomes are reported in the *d* metric. This conversion is made only when study results are in the *r* metric. But it should be noted that when study results are presented in the *d* metric, it will generally not be possible for the AP procedure to correct for range restriction or for measurement error in the independent variable (as explained in Hunter & Schmidt, 2004, chap. 7). However, it is still possible to correct for measurement error in the dependent variable measure, and so the AP procedure remains somewhat different from the H&O procedure in this respect. The H&O procedure generally does not include correction for any artifacts.

The reason given by the authors for this *r* to *d* transformation is to avoid the use of the *r* to Fisher's *z* transformation, which has been shown to produce fairly serious biases in the meta-analysis of correlations (Field, 2005; Schulze, 2004). The HO procedure for meta-analysis of correlations employs the Fisher's *z* transformation, and so Aguinis et al. did not want the AP method to be an adaptation of this procedure. However, I believe it would have been preferable to use as the starting point the HO procedure for correlations, leaving out the Fisher's *z* transformation. This would avoid the need to make an *r* to *d* transformation that has no logical foundation. In addition, such a procedure might be at least slightly more accurate. However, doing this would not solve the FE vs. RE problem described earlier. That is, the new procedure employed within each moderator group would still be a FE model, with all the potential problems entailed in any real-world case in which there is

Downloaded from http://orm.sagepub.com at The University of Iowa Libraries on December 19, 2007
© 2008 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.

more than one moderator operating. However, Hedges and Olkin (1985; see also Hedges & Vevea, 1998) do present an RE model for correlations that could be adapted by adding corrections for measurement error and range restriction. In fact, this was done by Hall and Brannick (2002). So the problem is not insuperable. This would be a desirable modification of the AP procedure.

Although the AP and HS procedures generally provided very similar results in these studies, this was not the case in one area: omnibus homogeneity tests. (See Table 2 and Tables 4 through 6.) In situations in which the researcher postulates a priori moderator hypotheses, these tests are usually not used (or at least should not be used). Instead, the moderator hypotheses are (or should be) tested by breaking the data out into subgroups or by regressing study outcomes onto hypothesized moderator variables (Hunter & Schmidt, 2004). In situations in which there are no priori moderator hypotheses, omnibus homogeneity tests are often used in an attempt to determine whether study results are heterogeneous (Hunter & Schmidt, 2004, pp. 401-406). If the test suggests they are, this suggests one or more moderators may be operating to produce variability in outcomes across studies. (This is not a certain conclusion, however, because the apparent heterogeneity could be caused by artifacts not corrected for rather than by moderators.) In the AP procedure (and the HO procedure) the chi-square Q statistic is used, whereas in the HS method one can use the 75% rule. The rationale for the 75% rule is that if 75% of the observed variance of r values or d values is explained by sampling error and other artifacts, then it is likely that the remaining 25% is due to the several artifacts that it is not possible to correct for (Hunter & Schmidt, 2004, pp. 54 and 146; Schmidt et al., 1993). These additional artifacts occur in real data but generally not in simulated data.

In comparing the results produced by these two approaches in this study we see the classic trade off between Type I and Type II errors. In the Aguinis et al. study results, the 75% rule in general does a much better job of controlling Type I errors. Using the 75% rule, one typically has only about a 5% or 6% chance of concluding there is a moderator when there is none. The Type I error rate is considerably higher for both the original HO Q test and the modified Q test used in the AP method. On the other hand, the 75% rule has a higher Type II error rate, meaning that if there is a moderator, you are less likely to detect it. It is simply a fact that it is not possible for any procedure to have both a low Type I error rate and a low Type II error rate—unless the moderator is quite large (rare) and/or the number of studies is very large (also quite rare; Hunter & Schmidt, 2004, pp. 68-71). So we have to live with this trade off.

As in the present study, Sackett, Harris, and Orr (1986) compared the Type I and Type II error rates for the Q statistic and the SH procedure using simulation methods. As in the present study, they varied the number of studies, the sample size of studies, the level of variation in population correlations, mean values of population correlations, and level of measurement error (although they did not include the artifact of range restriction). As with the present Aguinis et al. study, the Sackett et al. (1986) study was quite thorough. Yet its findings were very different. Sackett et al. found that under all conditions the SH procedure had a Type II error rate lower than or equal to that of the Q statistic. On the other hand, they found that the SH procedure had a higher Type I error rate than the Q statistic. That is, the pattern of findings in the Sackett et al. study was the exact opposite of that reported in the present study. The Aguinis et al. study does cite the Sackett et al. study but

Downloaded from http://orm.sagepub.com at The University of Iowa Libraries on December 19, 2007
© 2008 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.

does not discuss the large difference in the final conclusions. The major cause of the contradictory results is that the Sackett et al. study used a cutoff of 90% of variance accounted for, not 75% as used by Aguinis et al. The 90% cutoff is arguably more appropriate for simulated data than the 75% value. As noted above, the rationale for the 75% values is that in real data one is typically unable to correct for some artifacts, such as typographical and other data errors and contamination or deficiency in the criterion measure; and such artifacts could easily account for 25% of the observed between-study variance. Therefore, if in real data corrected artifacts account for a high proportion of the observed variance (e.g., 75%) then virtually all observed variance is probably artifactual. With simulated data, these uncorrectable artifacts do not exist, and so a higher percentage variance cutoff is more appropriate. The reader should note that in simulated data a 90% cutoff causes the Type I error rate to be higher, because the 90% criterion makes it more difficult to conclude that there is no moderator present. On the other hand, when a moderator does in fact exist, the 90% cutoff produces higher power than the 75% cutoff—because (again) the standard for concluding there is no moderator is more difficult to meet. Hence, both the Type I and the Type II error rates found for the SH procedure in simulated data depend on the cutoff criterion used. In light of the rationale presented above, it is probably the case that results obtained in simulated data with the 90% cutoff provide a more accurate picture of what happens in real data than results obtained with the 75% cutoff.

One final observation is relevant here. Sackett et al., found that the Type I error rate for the Q test was close to the nominal 5% rate, whereas Aguinis et al. found much higher Type I error rates for their Q test. It is not clear what the explanation is for this difference, but it may have to do with the adaptation of the Q test in the AP procedure used to make the test appropriate for corrected correlations.

Given the trade-off between Type I and Type II errors discussed above, should a researcher prefer a procedure with a low Type I error rate but a higher Type II error rate? Or should a procedure with a higher Type I error rate but a lower Type II error rate be preferred? The Aguinis et al. article argues that it is more important to have a low Type II error rate than a low Type I error rate. The choice between these two types of procedure depends on one's estimation of how frequently moderators occur. If moderators are rare, then the important type of error is Type I error. In this situation, the real danger lies in concluding you have found a moderator when in fact there is no moderator. There are many opportunities to make this (Type I) error and few opportunities to make a Type II error. This may in fact be the case in most literatures. Many researchers seem to believe that moderators are very frequent (even ubiquitous), but the fact is that there is very little empirical evidence to support this assumption (Schmidt, Hunter, Pearlman, & Hirsh, 1985). That is, there are very few well-established, replicated moderators; and claims of moderator findings are notorious for their failure to replicate, even in large *N* studies. The belief that moderators are pervasive can be argued to be based on observation of variations in data that are mostly due to of sampling error and other artifacts; such variations often appear to be "real." Therefore, it may be that real moderators (i.e., population-based moderators large enough to be of practical or theoretical significance) are rare, and so the critical task may be to ensure control of Type I errors. The Aguinis et al. study finds that the 75% rule does this better than the *Q* statistic, but the Sackett et al. (1986) study found the opposite.

Downloaded from http://orm.sagepub.com at The University of Iowa Libraries on December 19, 2007
© 2008 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.

On the other hand, if moderators are frequent, then Type II errors are more serious than Type I errors (because Type II errors have the opportunity to occur more often), and the procedure of preference is the one with the lowest Type II error rate, even if the price paid for its use is an elevated Type I error rate. In fact, this is generally the case when the focus is study *main effects* (correlations or group differences; e.g., Lipsey & Wilson, 1993; Schmidt, 1996; Schmidt, Hunter, & Urry, 1976). The reason for this is that study main effects are almost always present, which means there are many opportunities to commit Type II errors and few opportunities to commit Type I errors. For example, Lipsey and Wilson (1993) reviewed more than 300 meta-analyses of psychological and educational interventions and found that only one had a zero or near zero effect (less than 1% of the interventions). Hence, in such literatures, it is almost impossible to make a Type I error in connection with mean effect sizes, and so only Type II errors are really important. In the case of moderators, however, we have yet to see any comparable empirical evidence that they are real and frequent. Belief in moderators may often be based on seeming plausibility or intuitive hunches more than on any hard evidence. If so, it would not be appropriate to recommend a moderator detection procedure with a high Type I error rate.

The Steel and Kammeyer-Mueller (2008) Article

This article is daunting in its complexity. I have read it carefully twice and I am still not sure whether I fully understand the mechanics of the Bayesian procedure they advocate. The authors of this article are concerned about the practice in meta-analysis—and in many other statistical methods such as ANOVA (Hunter & Schmidt, 2004, pp. 399-401)—of setting negative variance estimates to zero. They correctly state that second order sampling error is a problem in estimating the variance of population parameters in meta-analysis, and they point out that one can obtain an estimate of zero variance when in fact the true value is some positive number. As an alternative to the usual practice of setting negative variance estimates to zero, they present a very complex Bayesian procedure for estimating this variance. They advocate this procedure for general usage, not just for situations in which the initial variance estimate is negative.

Their computer simulation tests of this procedure yielded mixed results, with the traditional procedure being more accurate in some cases and their Bayesian procedure being better in others. I found this pattern of findings somewhat confusing. For example, I was not able to discern any real explanation for why the traditional procedure was superior when $k = 15$ and mean $N = 50$, and when $k = 50$ for all N values. The results seemed to form an odd patchwork. In general, when the number of studies was 30 or more, the two procedures appeared to give similar results; the results differed only when k was small. However, the authors point out that many meta-analyses are based on a small number of studies. On the other hand, the condition of small k is precisely the condition in which no procedure can provide a very accurate estimate of SD_ρ or SD_δ in any single meta-analysis, because of the presence of substantial second order sampling error. When k is small, accuracy can be obtained only by averaging these values across multiple meta-analyses, as described in more detail below.

Downloaded from http://orm.sagepub.com at The University of Iowa Libraries on December 19, 2007
© 2008 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.

It is not clear why the authors presented their simulation results as averages across all values of population variance. Because they present their approach primarily as a way to deal with the issue of negative estimated variances, it would be more informative to report the results separately for situations in which moderator variance is a small percentage of total (observed) effect size variance. It is in situations of this sort that negative variance estimates occur most frequently.

My conclusion is that the authors' Bayesian procedure will probably provide more accurate estimates of SD_ρ or SD_δ for small k meta-analyses for which the conclusion of zero variance is incorrect (i.e., the product of second-order sampling error). However, this does not include all small k meta-analyses that reach a conclusion of zero variance, and there is no way to tell in advance whether any given meta-analysis is one in which the zero variance conclusion is due to second-order sampling error or is in fact a correct conclusion.

In the area of meta-analysis methods, things are often not simple. I believe that the approach taken in this article misses one of the important "big pictures" in meta-analysis. This is the fact that even when the traditional procedure of setting negative variance estimates to zero is used, the variance estimates have a positive (upward) bias. The effect of substituting the Bayesian procedure for the traditional procedure is to further increase this positive bias, by substituting a positive value for the zero value. Why is there a positive bias? There are two reasons. First, in all unbiased subtractive variance estimation procedures, setting negative estimates to zero creates a positive bias. Hedges and Vevea (1998) discuss this bias in some detail as it affects their random effects meta-analysis procedure. This bias also explains the positive bias in the HS methods in estimates of SD_ρ that was found in Law, Schmidt, and Hunter (1994). That is, a procedure that is completely unbiased when the negative variance estimates are not set to zero has a positive bias when they are set to zero. In the Bayesian procedure advocated here, negative values are not set to zero but to some positive value; hence the result in an increase in the upward bias of estimated SD_ρ or SD_δ values.

It is important to note that in their Tables 1 and 2, Steel and Kammeyer-Mueller reported *negative* biases for estimates of population variances by the traditional procedure (i.e., the procedure that sets negative variance estimates to zero), instead of the expected positive biases. The explanation for this apparent anomaly can be found in Hunter and Schmidt (2004, pp. 407-408). The statistic Steel and Kammeyer-Mueller averaged across simulation results reduces to $1 - \text{Var}(e)/\text{Var}(r)$, where Var($e$) is sampling error variance and Var(r) is the variance of the observed study correlations. As explained in Hunter and Schmidt, the second term in this expression has a positive bias because of the fact that sampling error can produce extremely small values of Var(r), even values as small as zero, resulting in extremely large or even infinite values for the ratio Var(e)/Var(r). The solution is to invert this ratio—that is, to use the ratio Var(r)/Var(e)—compute the average of this ratio, and then to take this inverse of this average value as the estimate of average percentage variance accounted for. With the inverse ratio the most extreme possible value is zero and so there is no longer a negative bias in the estimate of the average percentage of variance accounted for. Because of these considerations, the negative biases reported by Steel and Kammeyer-Mueller for the traditional procedure should actually be positive.

Downloaded from http://orm.sagepub.com at The University of Iowa Libraries on December 19, 2007
© 2008 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.

However, the setting of negative variance estimates to zero is not the only source of upward bias. Variance produced by artifacts that cannot be corrected for is the other source. As detailed in Schmidt et al. (1993) there are 10 or more variance-producing artifacts that typically cannot be corrected. The combination of these two different sources of upward bias means that on average meta-analytic estimates of the variance (or *SD*) of population parameters are considerably too large. Steel and Kammeyer-Mueller focus on the fact that in some cases in which the variance estimate is zero, this value may be an underestimate. But they appear to completely miss the bigger picture of upward bias. The effect of this upward bias is to create the false impression that moderators—especially larger moderators—are more frequent than they actually are.

Steel and Kammeyer-Mueller are correct in stating that estimates of SD_ρ or SD_δ are often fairly uncertain in meta-analysis. As they point out, the reason is that, for this estimate, the *N* is the number of studies (*k*), not the cumulative *N* across the studies. We have learned from meta-analysis that a single study can almost never answer a question. For that we need multiple studies—preferably a large number—and application of meta-analysis to these studies. Meta-analysis usually provides fairly precise estimates of mean effect sizes, but unless *k* is fairly large, a single meta-analysis cannot usually provide a really accurate estimate of SD_ρ. What is needed here is an average across multiple meta-analyses—a second-order meta-analysis (Hunter & Schmidt, 1994, pp. 406-408). That is, in meta-analysis domains in which theoretical considerations and hypothesized moderators are identical or similar, we can increase the accuracy of estimates of SD_ρ or SD_δ by averaging across meta-analyses and hence averaging out second-order sampling error. Hunter and Schmidt (2004, pp. 406-407) give an example of this. Of course, such an average estimate would still have the upward bias caused by inability to correct for a number of variance-producing artifacts. However, it would be possible to eliminate the upward bias created by setting negative variance estimates to zero by retaining the negative values in computing the average value. Steel and Kammayer-Mueller would probably object to this on grounds that a variance cannot be negative, and it is true that a negative estimate is "nonsense" for any particular meta-analysis. But when the focus is on producing the least biased averaged estimate, retaining the negative values reduces bias. Again, as I stated above, methodological questions in meta-analysis can be complex.

The problem then is that the Steel and Kammayer-Mueller Bayesian procedure would increase the upward bias in the second-order meta-analysis, beyond even the bias resulting from setting the negative values to zero. In this sense, I would have to say that this procedure would not contribute to cumulative knowledge.

Steel and Kammeyer-Mueller call for the placing of Bayesian confidence intervals (CIs) around estimates of the variance of population values. A technical point is relevant here: Given the Bayesian nature of these intervals, it appears they would be credibility intervals, not confidence intervals, because they are based on the variance of the population parameters, and not on sampling error estimates.[2] But these intervals do nonetheless reveal the substantial uncertainty in these estimates whenever *k* is small, showing that their Bayesian estimation procedure, like all alternative estimation procedures "cannot get blood out of a turnip." That is, as noted above, when *k* is small there is no way to get an accurate estimate of SD_ρ or SD_δ in any single meta-analysis. This raises the question of whether we should

Downloaded from http://orm.sagepub.com at The University of Iowa Libraries on December 19, 2007
© 2008 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.

spend a great deal of time and effort developing and using a procedure aimed precisely at doing this. Recall that it is only when k is small that the Steel-Kammeyer-Mueller procedure yields estimated population variances different from those produced by the traditional (and much simpler) procedure. As noted above, when k is small the only way to get accurate estimates of this value is through use of second-order meta-analysis.

Steel and Kammeyer-Mueller present evidence that zero estimates of population variance are not rare in meta-analyses in the literature. They also quote Stewart and Roth (2001) as saying that in their meta-analysis "analyses of all *non-outliers* [italics added] showed that all variance was a result of research artifacts" (p. 149) In fact, the inappropriate use of outlier analyses to throw out large and small observed values is a major contributor to false conclusions of zero population variance. These procedures typically eliminate observed correlations or d values that are +2.00 or –2.00 *SD*s from the mean value. Hence, they are based on the false assumption that sampling error cannot produce values that extreme. Many if not most of the "outliers" so eliminated from the meta-analysis are nothing more than rare but still expected larger sampling errors. The result is underestimation of the variance of the observed values, resulting in underestimation of the population variance. The problem with outlier analysis in meta-analysis is that it is almost impossible to distinguish between large sampling errors and true outliers (i.e., actual erroneous data). As a result, many valid data points with large sampling errors are thrown out. This is not a problem in the estimation of the mean effect size, but it creates a downward bias in estimation of the population variance (Hunter & Schmidt, 2004, pp. 196-197). This is why we recommend against the use of outlier analysis in meta-analysis in Hunter and Schmidt (2004).

Could the downward bias in estimates of SD_ρ and SD_δ created by applying outlier analysis help to cancel out the upward biases discussed above? It might, but generally there is no way to know whether this downward bias is of the right size to counter the upward biases discussed earlier. My belief is that we should strive to reduce or eliminate all biases. Introducing a bias in the hopes that it will cancel out another bias does not seem like good practice.

Steel and Kammeyer-Mueller maintain that all zero estimates of population variance are "nonsense estimates." They argue that such estimates are implausible and that there is always positive variance across studies. Bear in mind that what is being discussed here is the variance that is left after all artifactual variance has been removed—not observed variance and not the partially corrected estimate of population variance produced in meta-analyses because of the fact that it is never possible to correct for all variance-producing artifacts. That is, we are talking about variation at the level of construct relations. In my judgment, the authors' position on this question is based on subjective judgment and is contrary to much empirical evidence. For example, based on an extensive empirical database, Schmidt et al. (1993) concluded the variability of true (operational) validity for a wide variety of cognitive tests used in employment, controlling for job complexity level, was zero or so near to zero as to be indistinguishable from zero. This issue is also discussed in Schmidt and Raju (2007). Steel and Kammeyer-Mueller do not mention this evidence; possibly they do not find it plausible. But plausibility is a subjective judgment; we are talking here about a very large amount of empirical evidence that contradicts this subjective judgment, at least in this one research area. We suspect that if such extensive

Downloaded from http://orm.sagepub.com at The University of Iowa Libraries on December 19, 2007
© 2008 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.

empirical databases were compiled in other areas, the results would often be similar (Schmidt & Raju, 2007). Here I refer the reader to the comments on the frequency of moderators that I made earlier in connection with the Aguinis et al. (2008) article. The evidence suggests that moderators are often solipsistic: they exist in the minds of researchers but not in real data populations.

This article discusses three different non-Bayesian methods of estimating SD_ρ or SD_δ that the authors say could be compared to their Bayesian method: the Hedges and Vevea (1998) method, the Raju and Drasgow (2003) method, and the Hunter and Schmidt (2004) method. They chose to compare their method only to the HS method but stated they could have also presented comparisons with the other two methods. Actually, this is not correct. The HV and Raju–Drasgow estimates of population variances are in the Fisher's *z* metric, not the *r* metric. These *SD* estimates cannot be converted to the correlation metric (e.g., Hall & Brannick, 2002; Hunter & Schmidt, 2004, pp. 82-83, 203-205; Schulze, 2004), and so cannot be compared to either the Steel-Kammeyer-Mueller estimates or the HS estimates, both of which are in the correlation metric. Although any value of Fisher's *z* or any mean Fisher's *z* can be converted to *r*, it is not possible to convert a variance or *SD* in Fisher's *z* metric to the *r* metric. This is a serious limitation for all methods using the Fisher's *z* metric, because to be usable the final results have to be expressed in the correlation metric.

Summary

These four articles, each in its own way, along with this commentary, make a contribution to improvement in understanding of meta-analysis methods among researchers. The overall theme that I take away from this feature topic is the complexity of the methodological issues involved in meta-analysis. But although these issues are complex, they are not impenetrable. Over time we as a field can come to an ever-increasing understanding of them and their implications for the conclusions we draw about cumulative scientific knowledge from our meta-analyses.

Notes

[1] These symbols refer to the standard deviation of the actual underlying population parameters. They are estimates of what the observed standard deviation of outcomes across studies would be if the number of studies were very large, the sample size were infinite in each study, and study results were not affected by measurement error, range restriction, or any other statistical artifacts.

[2] Confidence intervals around the mean reflect the amount of uncertainty in the estimate of the mean due to sampling error. Credibility intervals reflect the estimated amount of variability in the underlying population values (population parameters). Credibility intervals do not reflect sampling error; sampling error is removed prior to calculation of credibility intervals. In contrast, confidence intervals are completely determined by sampling error.

References

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

Downloaded from http://orm.sagepub.com at The University of Iowa Libraries on December 19, 2007
© 2008 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.

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

Bechtel, W. (1988). *Philosophy of science: An overview for cognitive science*. Hillsdale, NJ: Lawrence Erlbaum.

Field, A. P. (2001). Meta-analysis of correlation coefficients: A Monte Carlo comparison of fixed- and random-effects methods. *Psychological Methods, 6*, 161-180.

Field, A. P. (2003). The problem in using fixed-effects models of meta-analysis on real world data. *Understanding Statistics, 2,* 77-96.

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

Hall, S. M., & Brannick, M. T. (2002). Comparison of two random effects methods of meta-analysis. *Journal of Applied Psychology, 74*, 469-477.

Hedges, L. V. (1994). Statistical considerations. In H. Cooper & L. V. Hedges (Eds.),*The handbook of research synthesis* (pp. 29-38). New York: Russell Sage.

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

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3*, 486-504.

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

Hunter, J. E., & Schmidt, F. L. (1994). The estimation of sampling error variance in meta-analysis of correlations: The homogeneous case. *Journal of Applied Psychology*, *79*, 171-177.

Hunter, J. E., & Schmidt, F. L. (2000). Fixed effects vs. random effects meta-analysis models: Implications for cumulative research knowledge. *International Journal of Selection and Assessment, 8*, 275-292.

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

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

Hunter, J. E., Schmidt, F. L., & Le, H. (2006). Implications of direct and indirect range restriction for meta-analysis methods and findings. *Journal of Applied Psychology, 91*, 594-612.

Kierein, N. M., & Gold, M. A. (2000). Pygmalion in work organizations: A meta-analysis. *Journal of Organizational Behavior, 21*, 913-928.

Kisamore, J. L., & Brannick, M. T. (in press). An illustration of the consequences of meta-analysis model choice. *Organizational Research Methods*.

Law, K. S., Schmidt, F. L., & Hunter, J. E. (1994). A test of two refinements in meta-analysis procedures. *Journal of Applied Psychology, 79*, 978-986.

Le, H., & Schmidt, F. L. (2006). Correcting for indirect range restriction in meta-analysis: Testing a new meta-analytic method. *Psychological Methods, 11*, 416-438.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48,* 1181-1209.

McNatt, D. B. (2000). Ancient Pygmalion joins contemporary management: A meta-analysis of the result. *Journal of Applied Psychology, 85,* 314-322.

National Research Council. (1992). *Combining information: Statistical issues and opportunities for research.* Washington, DC: National Academy of Sciences Press.

Overton, R. C. (1998). A comparison of fixed effects and mixed (random effects) models for meta-analysis tests of moderator variable effects. *Psychological Methods, 3*, 354-379.

Phillips, D. C. (1987). *Philosophy, science, and social inquiry*. Oxford, UK: Pergamon.

Raju, N. S., & Drasgow, F. (2003). Maximum likelihood estimation in validity generalization. In K. R. Murphy (Ed.), *Validity generalization: A critical review* (pp. 263-285). Mahwah, NJ: Lawrence Erlbaum.

Raudenbush, S. W. (1994). Random effects models. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 301-322). New York: Russell Sage.

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

Rothstein, H. F., Sutton, A. J., & Borenstein, M. (2005). *Publication bias in meta-analysis: Prevention, assessment and adjustment.* London: Wiley.

Downloaded from http://orm.sagepub.com at The University of Iowa Libraries on December 19, 2007
© 2008 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.

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

Sanchez-Meca, J., & Marin-Martinez, F. (1998). Weighting by inverse variance or by sample size in meta-analysis: A simulation study. *Educational and Psychological Measurement, 58*, 211-220.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1*, 115-129.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62*, 529-540.

Schmidt, F. L., Hunter, J. E., Pearlman, K., & Hirsh, H. R. (1985). Forty questions about validity generalization and meta-analysis. *Personnel Psychology*, *38*, 697-798.

Schmidt, F. L., Hunter, J. E., & Urry, V. (1976). Statistical power in criterion-related validation studies. *Journal of Applied Psychology*, *61*, 473-485.

Schmidt, F. L., Law, K., Hunter, J. E., Rothstein, H. R., Pearlman, K., & McDaniel, M. (1993). Refinements in validity generalization methods: Implications for the situational specificity hypothesis. *Journal of Applied Psychology, 78*, 3-13.

Schmidt, F. L., & Le, H. (2004). *Software for the Hunter–Schmidt meta-analysis methods*. Iowa City: University of Iowa, Department of Management & Organizations.

Schmidt, F. L., Oh, I.-S., & Hayes, T. (2006). *Fixed vs. random effects models in meta-analysis: An empirical comparison of differences in results in the psychological literature*. Manuscript submitted for publication.

Schmidt, F. L., Oh, I.-S., & Le, H. (2006). Increasing the accuracy of corrections for range restriction: Implications for selection procedure validities and other research results. *Personnel Psychology, 59*, 281-305.

Schmidt, F. L., & Raju, N. S. (2007). Updating meta-analytic research findings: Bayesian approaches versus the medical model. *Journal of Applied Psychology, 92*, 297-308.

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

Steel, P. D. G., & Kammeyer-Mueller, J. (in press). Bayesian variance estimation for meta-analysis: Quantifying our uncertainty. *Organizational Research Methods*.

Stewart, W. H., & Roth, P. L. (2001). Risk propensity differences between entrepreneurs and managers: A meta-analytic review. *Journal of Applied Psychology, 86*, 145-153.

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

Toulmin, S. S. (1961). *Foresight and understanding: An enquiry into the aims of science*. New York: Harper.

Wood, J. (in press). Methodology for dealing with duplicate study effects in a meta-analysis. *Organizational Research Methods*.

Frank Schmidt is the Ralph L. Sheets Professor of Human Resources in the Tippie College of Business at the University of Iowa. He was one of the two coinventors of validity generalization methods and has published more than 150 journal articles and book chapters. He received the Distinguished Scientific Contributions Award (with John Hunter) from the American Psychological Association and also received the Distinguished Career Award from the Human Resources Division of the Academy of Management and the Michael R. Losey Human Resources Research Award from the Society for Human Resource Management. He received his doctorate in industrial/organizational psychology from Purdue University.

Downloaded from http://orm.sagepub.com at The University of Iowa Libraries on December 19, 2007
© 2008 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.