Journal of Educational and Behavioral Statistics
June 2007, Vol. 32, No. 2, pp. 180–205
DOI: 10.3102/1076998606298041

Combining Correlation Matrices: Simulation Analysis of Improved Fixed-Effects Methods

Adam R. Hafdahl
University of Missouri–Columbia

*The originally proposed multivariate meta-analysis approach for correlation matrices—analyze Pearson correlations, with each study's observed correlations replacing their population counterparts in its conditional-covariance matrix—performs poorly. Two refinements are considered: Analyze Fisher Z-transformed correlations, and substitute better estimates of correlations in the conditional covariances. Fixed-effects methods with and without each refinement were examined in a Monte Carlo study; number of studies and the distribution of within-study sample sizes were varied. Both refinements improved element-wise point and interval estimates, as well as Type I error control for homogeneity tests, especially with many small studies. Practical recommendations and suggestions for future methodological work are offered. An appendix describes how to transform Fisher-Z (co)variances to the Pearson-*r *metric.*

Keywords: *meta-analysis; correlation; generalized least squares; Monte Carlo study*

Meta-analysis is a set of quantitative methods for comparing and combining several estimates of some parameter—typically an effect-size index such as a correlation, standardized mean difference, or odds ratio—to improve estimation and inference and investigate generalizability to relevant populations. Typical meta-analysis tasks include estimating the (mean) effect size and constructing a confidence interval (CI) for or testing hypotheses about it, as well as assessing the magnitude and covariates of effect-size heterogeneity among studies. Meta-analysis is employed routinely in integrative research syntheses aimed at building theory, as well as informing policy and practice in the behavioral, social, and medical sciences. Several treatments of research synthesis concepts and techniques are available (e.g., Cooper & Hedges, 1994; Hedges & Olkin, 1985; Hunter & Schmidt, 2004; Lipsey & Wilson, 2001).

Portions of this article are based on a paper presented at the meeting of the Psychometric Society in Chapel Hill, North Carolina, June 21, 2002. I thank Jack Vevea for helpful comments on earlier drafts. Correspondence concerning this article should be addressed to Adam R. Hafdahl, Department of Mathematics, Campus Box 1146, Washington University in St. Louis, St. Louis, MO 63130; e-mail: arhafdah@wustl.edu.

Pearson's product-moment correlation coefficient is one of the most prevalent effect-size indices, in part because of its role as a validity coefficient (Hunter & Schmidt, 2004). Methodologists have amassed an extensive literature proposing, evaluating, and refining meta-analytic methods for the univariate case, wherein each study contributes one correlation (e.g., Burke, 1984; Cornwell & Ladd, 1993; Hedges, 1988, 1989; Law, Schmidt, & Hunter, 1994; Osburn & Callender, 1992). These contributions have led to some consensus about best methods for particular circumstances, although certain issues remain unresolved.

Growing interest in meta-analysis of complex relationships has focused attention on techniques related to synthesizing correlation matrices (Becker, 1992b, 1995), such as meta-analytic explanatory models (e.g., Becker & Schram, 1994; Shadish, 1996) and meta-analytic analogues of exploratory factor analysis (Hafdahl, 2001) and structural equation modeling (M. W. L. Cheung & Chan, 2005; Viswesvaran & Ones, 1995). Scant published work, however, concerns evaluation of multivariate meta-analysis for correlation matrices. One should be wary of extrapolating performance from univariate methods to their multivariate counterparts, which typically include additional model components, require more complex estimation and inference procedures, and are more prone to complications such as missing data.

Given this dearth of evaluative work and in light of evidence that an early strategy for synthesizing correlation matrices performs poorly (see below), my objective in this article is to describe and evaluate improved meta-analytic techniques for a basic situation: a fixed-effects model with no study-level covariates (i.e., moderators). To this end, I first review common meta-analytic models and fixed-effects methods for correlations, then summarize the original multivariate approach's inferior behavior, and describe two refinements that have been proposed. In the balance of the article, I report a Monte Carlo study of these refinements. Sources of attenuation and artifactual variation in correlations, such as differential measurement unreliability or range restriction (Hunter & Schmidt, 2004), will not be considered here. Also, more complex and flexible random- and mixed-effects models that incorporate between-studies heterogeneity will be mentioned in passing only.

Meta-Analytic Models and Methods for Correlation Matrices

To explicate specifics, the following sections review the meta-analytic models and methods that serve as a basis for the refinements and Monte Carlo study described later. For a more detailed presentation, see Becker (1992b, 1995) or Becker and Schram (1994).

Multivariate Meta-Analysis

Two multivariate models for a correlation matrix are presented first, followed by generalized least squares (GLS) estimation and inference techniques for the simpler model.

Models. Suppose we are interested in the $p^* = p(p-1)/2$ distinct correlations between p multivariate-normal variables, and we obtain from each of k independent studies its sample size n_i, $i = 1, \ldots, k$, and observed correlations $\mathbf{r}_i$ as estimates of the population correlations $\boldsymbol{\rho}_i$ ($\mathbf{r}_i$ and $\boldsymbol{\rho}_i$ are column-vectors of the p^* distinct correlations from their respective matrices). Here ρ_{ij}, $j = 1, \ldots, p^*$, in $\boldsymbol{\rho}_i$ is not the disattenuated true correlation central to validity generalization studies. One standard meta-analytic model for $\mathbf{r}_i$ can be expressed in two parts. The within-study model

$$\mathbf{r}_i = \boldsymbol{\rho}_i + \mathbf{e}_i \tag{1}$$

expresses observed correlations' (co)variation around their population counterparts for Study i. Although the distribution of r is skewed for $\rho \neq 0$ and e_{ij} is not strictly independent of ρ_{ij} (Hedges, 1989), we customarily assume $\mathbf{e}_i \sim N_{p^*}(0, \mathbf{V}_i)$, where the $p^* \times p^*$ conditional covariance matrix $\mathbf{V}_i$ is a large-sample approximation whose typical element—the covariance between r_{iab} and r_{icd}—is a function of all correlations among variables a, b, c, and d (Olkin & Siotani, 1976):

$$\mathrm{Cov}(r_{iab}, r_{icd}) = \frac{\left[\begin{array}{c}\frac{1}{2}\rho_{iab}\rho_{icd}\left(\rho_{iac}^2 + \rho_{iad}^2 + \rho_{ibc}^2 + \rho_{ibd}^2\right) + \rho_{iac}\rho_{ibd} + \rho_{iad}\rho_{ibc} - \\ \left(\rho_{iab}\rho_{iac}\rho_{iad} + \rho_{iba}\rho_{ibc}\rho_{ibd} + \rho_{ica}\rho_{icb}\rho_{icd} + \rho_{ida}\rho_{idb}\rho_{idc}\right)\end{array}\right]}{n_i}. \tag{2}$$

When the two correlations share one variable (e.g., variables a and c are the same) or both, this expression involves only three correlations or only one, respectively. The standard approach has been to substitute $\mathbf{r}_i$ elements for $\boldsymbol{\rho}_i$ elements in Equation 2 but treat $\mathbf{V}_i$ as known.

The between-studies model expresses population correlations' (co)variation among studies. For the simplest nontrivial fixed-effects (i.e., homogeneous) case, this model is $\boldsymbol{\rho}_i = \boldsymbol{\rho}$, where $\boldsymbol{\rho}$ is the population correlation matrix common to all studies. In contrast, for the simplest random-effects (i.e., heterogeneous) case, the between-studies model is $\boldsymbol{\rho}_i = \boldsymbol{\mu}_\rho + \mathbf{u}_i$, where the random effects $\mathbf{u}_i$ have mean vector $\mathbf{0}$ and between-studies covariance-component matrix $\mathbf{T}$.

Substituting each between-studies model in the within-study model yields $\mathbf{r}_i = \boldsymbol{\rho} + \mathbf{e}_i$ (fixed effects) and $\mathbf{r}_i = \boldsymbol{\mu}_\rho + \mathbf{u}_i + \mathbf{e}_i$ (random effects). Whereas in the former model observed correlations (co)vary due solely to random subject sampling, the latter includes (co)variation due to study characteristics (e.g., methodological features, subject attributes; Hedges & Vevea, 1998). Either model may also include study-level covariates (Kalaian & Raudenbush, 1996), which will not be considered further here.

Estimation and inference. Becker (2000) provided an overview of multivariate meta-analysis, and many meta-analytic methods for dependent standardized mean differences (Gleser & Olkin, 1994; Kalaian & Raudenbush, 1996) are applicable to correlation matrices. I focus here on the fixed-effects model, which is most defensible for (subsets of) studies with (subsets of) correlations that exhibit homogeneity or when generalizing to studies such as those in the meta-analytic sample (Hedges & Vevea, 1998). One approach is to use the GLS estimator

$$\hat{\boldsymbol{\rho}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{r} = \left(\sum_{i=1}^{k}\mathbf{V}_i^{-1}\right)^{-1}\sum_{i=1}^{k}\mathbf{V}_i^{-1}\mathbf{r}_i = \left(\sum_{i=1}^{k}\mathbf{W}_i\right)^{-1}\sum_{i=1}^{k}\mathbf{W}_i\mathbf{r}_i, \qquad (3)$$

where the design matrix $\mathbf{X}$ consists of k stacked $p^* \times p^*$ identity matrices; $\mathbf{V}$ is block-diagonal with blocks $\mathbf{V}_1, \ldots, \mathbf{V}_k$, each $p^* \times p^*$; $\mathbf{r} = (\mathbf{r}_1', \ldots, \mathbf{r}_k')'$ is a kp^*-element column vector; and study i's weight matrix $\mathbf{W}_i$ is just $\mathbf{V}_i^{-1}$. The (estimated) sampling covariance matrix of $\hat{\boldsymbol{\rho}}$,

$$\hat{\mathbf{V}}(\hat{\boldsymbol{\rho}}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} = \left(\sum_{i=1}^{k}\mathbf{V}_i^{-1}\right)^{-1} = \left(\sum_{i=1}^{k}\mathbf{W}_i\right)^{-1}, \qquad (4)$$

can be used to construct large-sample CIs or confidence regions for or test hypotheses about one or more elements of $\boldsymbol{\rho}$. One may test omnibus (i.e., for all p^* correlation-matrix elements) homogeneity or model specification using the multivariate heterogeneity statistic

$$\begin{aligned} Q_M = \mathbf{r}'\mathbf{V}^{-1}\mathbf{r} - \hat{\boldsymbol{\rho}}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})\hat{\boldsymbol{\rho}} &= \sum_{i=1}^{k}(\mathbf{r}_i - \hat{\boldsymbol{\rho}})'\mathbf{V}_i^{-1}(\mathbf{r}_i - \hat{\boldsymbol{\rho}}) \\ &= \sum_{i=1}^{k}(\mathbf{r}_i - \hat{\boldsymbol{\rho}})'\mathbf{W}_i(\mathbf{r}_i - \hat{\boldsymbol{\rho}}), \end{aligned} \qquad (5)$$

which is distributed approximately as chi-square on $(k-1)p^*$ degrees of freedom under $H_0 : \boldsymbol{\rho}_i = \boldsymbol{\rho}\,\forall\, i$ (or, equivalently, $H_0 : \mathbf{T} = \mathbf{0}$); Q_M exceeding its critical value indicates heterogeneity in observed correlation matrices because of sources other than subject-level sampling error.

In typical practice, not all studies contribute all correlation-matrix elements (Becker, 1992a). One strategy is to substitute simple estimates for missing $\boldsymbol{\rho}_i$ elements in Equation 2, then modify Equations 3–5 by essentially deleting rows and columns of $\mathbf{r}$, $\mathbf{X}$, and $\mathbf{V}$ that correspond to missing correlations (Becker & Schram, 1994). Other approaches to this pervasive problem have been studied (S. F. Cheung, 2000; M. W. L. Cheung & Chan, 2005; Furlow & Beretvas, 2005).

Univariate Meta-Analysis

Univariate meta-analysis employs special cases of the above models and methods for the $p^* = 1$ correlation between $p = 2$ bivariate-normal variables.

Models. With each of k independent studies contributing n_i and r_{ij} as an estimate of ρ_{ij}—the jth element of $\boldsymbol{\rho}_i$—the within-study model becomes

$$r_{ij} = \rho_{ij} + e_{ij}. \tag{6}$$

We assume $e_{ij} \sim N(0, v_{ij})$ and hence $r_{ij} \sim N(\rho_{ij}, v_{ij})$, where

$$v_{ij} = (1 - \rho_{ij}^2)^2 / n_i \tag{7}$$

is a large-sample approximation for the conditional variance.[1] The fixed- and random-effects between-studies models are $\rho_{ij} = \rho_j$ and $\rho_{ij} = \mu_{\rho j} + u_{ij}$, respectively, where the random effects u_{ij} have mean 0 and between-studies variance component τ_j^2.[2] By substitution, the combined models are $r_{ij} = \rho_j + e_{ij}$ (fixed effects) and $r_{ij} = \mu_{\rho j} + u_{ij} + e_{ij}$ (random effects).

Estimation and inference. One approach for estimating the fixed-effects parameter uses weighted least squares (WLS) with weights $w_{ij} = 1/v_{ij}$ (Shadish & Haddock, 1994):

$$\hat{\rho}_j = \sum_{i=1}^{k} w_{ij} r_{ij} \Bigg/ \sum_{i=1}^{k} w_{ij}. \tag{8}$$

The (estimated) sampling variance of $\hat{\rho}_j$ is just

$$\hat{\mathrm{V}}(\hat{\rho}_j) = 1 \Bigg/ \sum_{i=1}^{k} w_{ij}. \tag{9}$$

Under the null hypothesis $H_0 : \rho_{ij} = \rho_j \forall i$ (or $H_0 : \tau_j^2 = 0$), the univariate heterogeneity statistic

$$Q_U = \sum_{i=1}^{k} w_{ij}(r_{ij} - \hat{\rho}_j)^2 \tag{10}$$

is distributed approximately as chi-square on $k - 1$ degrees of freedom. Again, standard practice is to substitute r_{ij} for ρ_{ij} in Equation 7 but treat v_{ij} and w_{ij} as known (cf. Hunter & Schmidt, 1994).

As others have noted, WLS amounts to GLS with each study's observed correlations treated as pairwise independent by setting off-diagonals of $\mathbf{V}_i$ to zero. Specifically, with diagonal $\mathbf{V}_i$, $\mathbf{W}_i$ is diagonal with elements that are just w_{ij}, and with complete data, Equations 3 and 4 yield estimates and sampling variances identical to those from applying Equations 8 and 9 separately to each element. (Q_M in Equation 5 reduces to the sum of $p^* Q_U$ statistics from Equation 10.)

Fisher's Z-transformation. Some authors recommend using Fisher's variance-stabilizing Z transformation[3] before meta-analyzing correlations (Hedges, 1988; Hedges & Olkin, 1985). The univariate models and methods in Equations 6–10 extend to Z-transformed correlations: Z and ζ replace r and ρ, respectively, and Equation 7 becomes $v_{ij} = 1/(n_i - 3)$; point and interval estimates are typically transformed back to the Pearson-r metric. Although several authors have debated theoretically or examined empirically this approach's merits and demerits (e.g., Alexander, Scozzaro, & Borodkin, 1989; James, Demaree, & Mulaik, 1986; Law, 1995; Schmidt, Hunter, & Raju, 1988; Silver & Dunlap, 1987), there does not appear to be consensus that analyzing Z-transformed correlations yields uniformly superior or inferior results.

Multivariate Meta-Analysis: Shortcomings, Advantages, and Refinements

Of the few published explanations of GLS for correlations, most describe analyzing Pearson correlations by substituting observed $\mathbf{r}_i$ for $\boldsymbol{\rho}_i$ in $\mathbf{V}_i$ (Becker, 1992b; Becker & Schram, 1994; cf. Becker, 2000). This pioneering approach—hereafter abbreviated multi-r_O—has undergone little scrutiny until very recently. In the following sections, I summarize Monte Carlo evidence of multi-r_O's (mis)behavior, highlight multivariate meta-analysis benefits, and consider potential improvements upon multi-r_O.

Troublesome Performance

Hafdahl (2001) reported a Monte Carlo study of methods for synthesizing correlation matrices prior to exploratory factor analysis. Most relevant to the present article are three fixed-effects methods (random-effects methods were also examined): univariate meta-analysis of Pearson correlations (uni-r_O) and Z-transformed correlations (uni-Z), and multi-r_O, all as described above. For both uni-r_O and multi-r_O, conditional (co)variances are estimated from $\mathbf{r}_i$, so these differ only in that uni-r_O ignores off-diagonal $\mathbf{V}_i$ elements. In the case of fixed-effects data, for each of between 5 and 200 simulated studies (median and mean n_i were 232 and 415), $\mathbf{r}_i$ was generated from a factor model for $p = 12$ variables (via the implied $\boldsymbol{\rho}$), and the above three methods were used to estimate and make inferences about the $p^* = 66$ population correlations.

Both univariate methods performed better than multi-r_O, often substantially so. Empirical bias and standard error were always larger for multi-r_O than for either univariate method; the latter contradicts multi-r_O's usually smaller estimated standard error (i.e., $\sqrt{\hat{V}(\hat{\rho}_j)}$). Homogeneity tests were not examined. Especially troubling was multi-r_O's declining relative performance with more studies—especially for CI coverage, which typically dropped near 0% for $k \geq 100$ while remaining near nominal for both univariate methods.

Other unpublished or very recently published work corroborates the poor performance of GLS via multi-r_O. Becker and Fahrbach (1994) found that it performed worse than several competing methods on a variety of criteria, including the omnibus homogeneity test (Equation 5). M. W. L. Cheung and Chan (2005) and S. F. Cheung (2000) compared three methods' omnibus homogeneity test and found multi-r_O generally inferior. Collectively, these studies provide compelling evidence that multi-r_O cannot be recommended for synthesizing correlation matrices.

Advantages of Multivariate Meta-Analysis

The poor performance of multi-r_O might persuade some to rely on univariate meta-analysis for correlation matrices. Although this may be valid for a single element (Becker & Schram, 1994; Gleser & Olkin, 1994), GLS in principle affords advantages and opportunities. First, consider testing homogeneity for two or more correlation elements: The multivariate test (Equation 5) accomplishes this for all elements simultaneously, whereas applying the univariate test (Equation 10) to every element would either inflate experiment-wise Type I error rate or—if a multiplicity adjustment were used—diminish power (S. F. Cheung, 2000).

Second, one may wish to make inferences about (functions of) two or more elements of $\boldsymbol{\rho}$. For example, Olkin and Finn (1990, 1995) described tests comparing certain dependent correlation elements to each other or to specified values, as well as a method to derive CIs for differences between simple, partial, or multiple correlations. These techniques use estimated correlations and their sampling covariance matrix; whereas multivariate meta-analysis provides the latter as $\hat{\mathbf{V}}(\hat{\boldsymbol{\rho}})$ (Equation 4), $\hat{\mathbf{V}}(\hat{\boldsymbol{\rho}})$ from univariate meta-analysis is diagonal.

Multivariate meta-analysis results can also be used to estimate and make inferences about other models. For instance, Becker (1992b) showed how to estimate standardized multiple regression models from a pooled or mean correlation matrix; inference for coefficients relies on $\hat{\mathbf{V}}(\hat{\boldsymbol{\rho}})$. Similarly, meta-analytic correlations have been used to analyze path models (e.g., Brown & Hedges, 1994; Emmers-Sommer & Allen, 1999; Hamilton, 1998; Roesch & Weiner, 2001), factor models (e.g., Grube, Bilder, & Goldman, 1998; Kavale, 1982; Klein, Wesson, Hollenbeck, Wright, & DeShon, 2001), and other structural equation models (SEMs). Whereas these meta-analytic SEM endeavors often involve ad hoc inferences based on univariate meta-analysis (S. F. Cheung, 2000; Furlow & Beretvas, 2005), more principled methods include using $\hat{\boldsymbol{\rho}}$ and $\hat{\mathbf{V}}(\hat{\boldsymbol{\rho}})$ from multivariate meta-analysis (M. W. L. Cheung & Chan, 2005).

Of course, these analysis opportunities are hardly advantageous if multivariate meta-analysis exhibits poor statistical properties. Improving on multi-r_O would allow synthesists to exploit the above benefits without sacrificing the validity of their statistical conclusions.

Proposed Refinements

Understanding better the standard multivariate strategy's poor performance would facilitate improving it. Two plausible explanations and corresponding refinements involve the conditional covariance matrix **V**, whose off-diagonal elements distinguish multi-r_O from uni-r_O. Others have proposed and studied these or similar refinements under some conditions.

Fisher-Z correlations. First, the univariate (WLS) and multivariate (GLS) within-study models both posit normality of observed correlations. Suppose this multivariate distribution's univariate marginals are approximately normal, as WLS assumes, but multivariate normality approximates poorly the joint distributions for pairs and larger sets of correlations. Then $\mathbf{V}_i$ might capture poorly the dependence between correlations and yield nonoptimal weights for Equations 3–5, and GLS may produce inferior estimates, standard errors, and homogeneity tests.

This leads to one multi-r_O refinement: Employ GLS with matrices of Z-transformed correlations, whose sampling distributions may better conform to multivariate normality. After Z-transforming every correlation from every study, the models and methods under *Multivariate Meta-Analysis* above still apply, except that **Z** and $\boldsymbol{\zeta}$ replace **r** and $\boldsymbol{\rho}$, respectively, and the typical element of $\mathbf{V}_i$ (Equation 2) is replaced by the conditional covariance between observed Z-transformed correlations $Z_{i\mathrm{ab}}$ and $Z_{i\mathrm{cd}}$. This covariance is still a function of the six relevant Pearson population correlations but with a slightly different denominator (Steiger, 1980):

$$\mathrm{Cov}(Z_{i\mathrm{ab}}, Z_{i\mathrm{cd}}) = \frac{n_i \mathrm{Cov}(r_{i\mathrm{ab}}, r_{i\mathrm{cd}})}{(n_i - 3)(1 - \rho_{i\mathrm{ab}}^2)(1 - \rho_{i\mathrm{cd}}^2)}. \tag{11}$$

For diagonal elements, this expression simplifies to the uni-Z conditional variance, which requires no estimation. The off-diagonals, however, are functions of $\boldsymbol{\rho}_i$. Convention dictates substituting $\mathbf{r}_i$ for $\boldsymbol{\rho}_i$, a method I refer to as multi-Z_O. Whether this will outperform multi-r_O is difficult to predict from previous studies of these methods' univariate counterparts, which do not require estimating covariances in $\mathbf{V}_i$. Becker and Fahrbach (1994) found that multi-Z_O outperformed multi-r_O but exhibited unacceptable Type I error control for multivariate tests (e.g., Equation 5).

One ostensible drawback of analyzing Fisher-Z correlations is that $\hat{\mathbf{V}}(\hat{\boldsymbol{\zeta}})$, the Fisher-Z analogue of Equation 4, is not in the Pearson-r metric (M. W. L. Cheung & Chan, 2005, footnote 5). Because many of the more sophisticated procedures mentioned above rely on $\hat{\mathbf{V}}(\hat{\boldsymbol{\rho}})$, a method for transforming $\hat{\mathbf{V}}(\hat{\boldsymbol{\zeta}})$ to $\hat{\mathbf{V}}(\hat{\boldsymbol{\rho}})$ is presented in the appendix to mitigate this hindrance.

Conditional-covariance estimation. Second, consider sampling error introduced by substituting $\mathbf{r}_i$ for $\boldsymbol{\rho}_i$ in $\mathbf{V}_i$. Although the diagonal of $\mathbf{V}_i$ includes exactly the same sampling error as the univariate conditional variances (Equation 7), the impact of sampling error off the diagonal of $\mathbf{V}_i$ may be sizeable: Operations on $\mathbf{r}_i$ elements to compute each conditional covariance (Equation 2) and invert $\mathbf{V}$ (Equations 3–5) may compound and propagate their sampling error and, in turn, distort GLS weights and impair multi-r_O performance.

A profitable strategy under the fixed-effects model, where $\boldsymbol{\rho}_i = \boldsymbol{\rho}$, may be to replace $\boldsymbol{\rho}_i$ elements in every study's $\mathbf{V}_i$ with those from an estimate of $\boldsymbol{\rho}$ more stable than $\mathbf{r}_i$. Such substitution is not novel (e.g., Hedges, 1983; Law et al., 1994; Osburn & Callender, 1992). This method, whereby estimated $\boldsymbol{\rho}$ replaces $\boldsymbol{\rho}_i$, will be referred to as multi-r_E or multi-Z_E for Pearson-r or Fisher-Z correlations, respectively. Becker and Fahrbach (1994) and S. F. Cheung (2000) reported that this estimated-ρ refinement improved performance over its observed-r_i GLS counterparts; Furlow and Beretvas (2005) found that multi-r_E outperformed uni-r_O and uni-Z.

Complete-data consequences. When every study contributes a complete set of observed correlations $\mathbf{r}_i$, replacing $\boldsymbol{\rho}_i$ in $\mathbf{V}_i$ with either $\boldsymbol{\rho}$ or $\hat{\boldsymbol{\rho}}$ has noteworthy effects on GLS and its relationship with WLS. To illustrate for multi-r_E, let $\boldsymbol{\Psi}_i = n_i\mathbf{V}_i$ and $N = n_1 + \cdots + n_k$. If $\boldsymbol{\rho}_i = \boldsymbol{\rho}$,[4] then $\boldsymbol{\Psi}_i = \boldsymbol{\Psi}$, $\mathbf{V}_i^{-1} = n_i\boldsymbol{\Psi}^{-1}$, and Equation 3 reduces to

$$\hat{\boldsymbol{\rho}} = \left(\sum_{i=1}^{k} n_i\boldsymbol{\Psi}^{-1}\right)^{-1} \sum_{i=1}^{k} n_i\boldsymbol{\Psi}^{-1}\mathbf{r}_i = \left(\sum_{i=1}^{k} n_i\right)^{-1} \boldsymbol{\Psi}\boldsymbol{\Psi}^{-1} \sum_{i=1}^{k} n_i\mathbf{r}_i = \sum_{i=1}^{k} n_i\mathbf{r}_i \Big/ N. \tag{12}$$

That is, with complete data from every study the GLS estimator is just the sample-size weighted mean of correlation matrices (S. F. Cheung, 2000). By a similar argument, Equation 4 reduces to

$$\mathbf{V}(\hat{\boldsymbol{\rho}}) = \left(\sum_{i=1}^{k} n_i\hat{\boldsymbol{\Psi}}^{-1}\right)^{-1} = \hat{\boldsymbol{\Psi}}\left(\sum_{i=1}^{k} n_i\right)^{-1} = \hat{\boldsymbol{\Psi}}/N, \tag{13}$$

where $\hat{\boldsymbol{\Psi}}$ is just $\boldsymbol{\Psi}$ with $\hat{\boldsymbol{\rho}}$ substituted for $\boldsymbol{\rho}$. Analogous results hold for $\hat{\boldsymbol{\zeta}}$ and $\hat{\mathbf{V}}(\hat{\boldsymbol{\zeta}})$ from multi-Z_E, with Equation 11, $n_i - 3$, and $N - 3k$ replacing Equation 2, n_i, and N, respectively.

Furthermore, consider the univariate special case ($p = 2$, $p^* = 1$): The WLS estimator and its sampling variance (Equations 8 and 9) reduce to scalar versions of Equations 12 and 13 and are equivalent to their counterparts in $\hat{\boldsymbol{\rho}}$ and $\hat{\mathbf{V}}(\hat{\boldsymbol{\rho}})$. Thus, with complete data and $\boldsymbol{\rho}_i = \boldsymbol{\rho}$ or $\boldsymbol{\rho}_i = \hat{\boldsymbol{\rho}}$, WLS yields identical estimates and standard errors as GLS for both Pearson-r and Fisher-Z correlations, which implies that the off-diagonals in $\mathbf{V}_i$ cannot improve $\boldsymbol{\rho}$ estimation.

These simplifications also ease computational burden considerably: $\hat{\boldsymbol{\rho}}$ need not involve $\boldsymbol{\Psi}$, $\hat{\mathbf{V}}(\hat{\boldsymbol{\rho}})$ requires computing the $p^*(p^* + 1)/2$ elements of $\boldsymbol{\Psi}$ only once (from $\hat{\rho}$), and for $\hat{\rho}$ and $\hat{\mathbf{V}}(\hat{\rho})$, the potentially large $\mathbf{X}$ and $\mathbf{V}$ need not be stored or manipulated. Moreover, to obtain element-wise standard errors, one can avoid entirely the tedious computation of $\boldsymbol{\Psi}$ off-diagonals.

Distinct meta-analytic methods. The above considerations lead to meta-analytic methods that differ according to whether (a) dependence between correlations is ignored via univariate WLS versus incorporated via multivariate GLS, (b) Pearson-r versus Fisher-Z correlations are analyzed, and (c) ρ_{ij} in the conditional (co)variances (Equation 2, 7, or 11) is replaced by r_{ij} versus $\hat{\rho}_j$ (or ρ_j). Crossing these factors yields seven distinct meta-analytic methods, six of which were described explicitly above: Hafdahl (2001) examined multi-r_O, uni-r_O, and uni-Z (for which the r_{ij} vs. $\hat{\rho}_j$ distinction is irrelevant); each of multi-Z_O and multi-r_E implements one refinement of multi-r_O; and multi-Z_E implements both refinements. The remaining method, uni-r_E, employs WLS with Pearson correlations by replacing ρ_{ij} in v_{ij} (Equation 7) with $\hat{\rho}_j$.

Monte Carlo Study of Refined and Existing Methods

In the balance of this article, I report a Monte Carlo study of several meta-analytic methods for correlation matrices, with particular attention to the above multivariate GLS refinements. It appears no published study has compared two or more of these multivariate methods. The rare unpublished work comparing two or more multivariate methods is limited by choice of methods, conditions, or evaluation criteria. Becker and Fahrbach (1994) compared all four multivariate methods described above (and others) on several criteria but only under the special case of $p = p^* = 3$ with constant n_i (i.e., $n_i = n \, \forall \, i$) and relatively few studies ($5 \leq k \leq 30$). S. F. Cheung (2000) used random n_i and included more studies ($15 \leq k \leq 150$ before deleting cases) but compared multi-r_O and multi-r_E only in terms of the omnibus homogeneity test, used relatively large n_i ($50 \leq n_i \leq 250$, $\mathrm{E}[n_i] = 230$), and examined only missing-data conditions. Addressing these limitations would solidify the evidence base for practical recommendations.

Method

The present Monte Carlo study is similar to fixed-effects segments of Hafdahl (2001) in that different methods were used to meta-analyze correlation matrices from varying numbers of studies, and accuracy, precision, and CI coverage were assessed; homogeneity tests were also examined. To cover more realistic conditions in which meta-analysts might work with correlation matrices, I varied the n_i distribution and used smaller correlation matrices. Simulations were conducted using SAS/IML (version 8.02). Computing code and data are available upon request.

Target population. Generating fixed-effects data requires specifying p and each correlation value for $\boldsymbol{\rho}$. Larger correlation matrices include more correlation pairs that share no variables, whose sampling covariances depend on six correlations instead of three (Equation 2); they also permit more opportunity for propagation of sampling error. A correlation matrix for $p = 4$ variables, with $p^* = 6$ correlations, was used. This avoids properties of the $p = p^* = 3$ case that might limit generalizability (e.g., three pairs of elements share no variables).

Particular correlation values were chosen to cover a range of positive values and reflect realistic structure. Specifically, a hypothetical recursive path model was established on the basis of theorized relationships among the four key variables in a particular social-cognitive research domain (self-discrepancy theory; Higgins, 1987): actual-ideal (I) and actual-ought (O) self-discrepancy, dejection (d), and agitation (a). Consistent with the theory, the standardized I$\rightarrow$d ($\beta_{dI} = .7$) and O$\rightarrow$a ($\beta_{aO} = .6$) paths are stronger than the I$\rightarrow$a ($\beta_{aI} = .1$) and O$\rightarrow$d ($\beta_{dO} = .2$) paths, and the exogenous I and O are correlated ($\rho_{OI} = .3$). This path model implies a correlation matrix with elements $\boldsymbol{\rho} = (\rho_{OI}, \rho_{dI}, \rho_{dO}, \rho_{aI}, \rho_{aO}, \rho_{ad})' = (.300, .760, .410, .280, .630, .322)'$.

Sample sizes. Both the number of studies and within-study sample sizes were varied. A given replication's meta-analytic sample included $k = 5$, 10, 20, 50, 100, or 200 studies, and n_i for each study was drawn from a positively skewed distribution typical of research syntheses (Osburn & Callender, 1992). Specifically, $n_i = \langle(\bar{n}/2)[(X_i - 3)/\sqrt{6}] + \bar{n}\rangle$, where $\bar{n} = 30, 100$, or 300; $X_i \sim \chi^2(3)$; and $\langle a\rangle$ denotes the integer nearest a. Hence, the (expected) n_i distribution is the same shape regardless of $\bar{n}$, with $\mathrm{E}(n_i) \approx \bar{n}$ and $\mathrm{V}(n_i) \approx (\bar{n}/2)^2$.

Replications and simulated data. The sample-size factors constitute a 6×3 ($k \times \bar{n}$) between-subjects factorial design, with the same $\boldsymbol{\rho}$ in all 18 cells. In each cell, 5,000 replications were run to yield relatively precise rejection rates ($SE[\hat{\alpha}] \approx .003$ for $\alpha = .05$) and other results. For a given replication, the observed correlation matrix from each of k simulated primary studies was generated using the SAS/IML functions RANNOR and ROOT (Cholesky decomposition): After n_i observations were generated from a 4-variate normal distribution with $\boldsymbol{\rho}$ as specified, Pearson correlations in $\mathbf{r}_i$ were computed from these n_i observations to yield fixed-effects data.

Meta-analysis. The k observed correlation matrices from each replication in a given cell were meta-analyzed using all seven methods delineated above. In this complete-data situation, multi-r_E and uni-r_E yielded equivalent estimates and standard errors, as did multi-Z_E and uni-Z; for these four methods, their $\hat{\boldsymbol{\rho}}$ was substituted for $\boldsymbol{\rho}$ to compute $\mathbf{V}_i$ or v_{ij}. Omnibus (Equation 5) and element-wise (Equation 10) homogeneity tests were used for multivariate and univariate methods, respectively. Estimation performance was evaluated by accuracy, precision, and efficiency of correlation estimates (transformed to the Pearson-r metric for

Fisher-Z methods); inference, by coverage and rejection rates for correlation CIs and homogeneity tests, respectively.

Results

Comparisons among multi-r_O and its refinements will be emphasized. Detailed element-wise results will be presented for only ρ_{Ol} (.30) and ρ_{dl} (.76), which behaved similarly to the four smaller and two larger correlations, respectively. Likewise, because patterns of inference results were similar at all three confidence and significance levels examined, only 95% CIs and homogeneity tests at $\alpha = .05$ will be presented.

To facilitate comparisons across factors more under the meta-analyst's control—method and k—figures are arranged with one plot for each combination of $\bar{n}$ and ρ and one connected series for each method across k. The methods' defining features are represented by plot symbol (closed [GLS] vs. open [WLS], circle [r] vs. diamond [Z]) and line style (dotted [observed-r_i] vs. dashed [estimated-ρ] vs. solid [no ρ]). In each figure plots with the same $\bar{n}$ share the same ordinate scale, but in some figures the scale varies across $\bar{n}$ to show important large-$\bar{n}$ patterns.

Accuracy. Figure 1 displays empirical bias for estimates of selected correlations, where bias$(\hat{\rho}_j)$ = E$(\hat{\rho}_j - \rho_j)$. Estimates for multi-r_E were attenuated slightly toward zero (i.e., negative bias), whereas the other methods' estimated correlations were inflated to some degree. Most methods were less biased with larger $\bar{n}$ or, less noticeably, smaller k. Although bias exhibited no clear relationship with ρ for most methods, multi-r_O was more accurate for larger ρ.

Accuracy varied considerably among multivariate methods. As expected, multi-r_O was substantially biased—usually enough to inflate an estimate's second decimal place or, when $\bar{n} = 30$, its first decimal place. Each refinement alone improved accuracy: Multi-Z_O and multi-r_E were always less biased than multi-r_O, especially for smaller ρ; this bias decrease was typically 3 to 7 times for multi-Z_O and 15 to 20 times for multi-r_E. Multi-Z_E, which incorporates both refinements, was often negligibly more accurate than multi-r_E. Only multi-Z_E and multi-r_E maintained absolute bias below .01 in all conditions and below .005 when $\bar{n} \geq 100$.

Uni-r_O was more accurate than multi-r_O, typically exhibiting two to three times less bias and nearly always maintaining bias below .05. Uni-r_O was, however, less accurate than all refined methods, especially multi-Z_E and multi-r_E (equivalent to uni-Z and uni-r_E, respectively).

Precision. In line with sampling variance expressions (e.g., Equations 7 and 9), empirical standard error for every method's correlation estimates—where $SE(\hat{\rho}) = \sqrt{\mathrm{E}\{[\hat{\rho}_j - \mathrm{E}(\hat{\rho}_j)]^2\}}$—was smaller with larger $\bar{n}$, k, or ρ. Among multivariate methods, multi-r_O was always least precise; multi-Z_O, markedly more

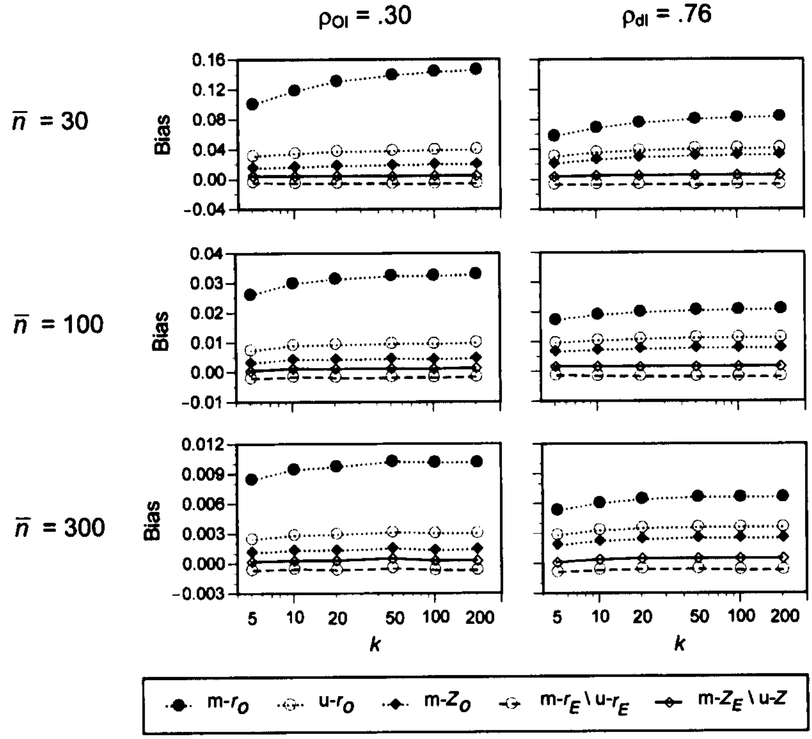

FIGURE 1. *Empirical bias for estimated common Pearson correlation ($\hat{\rho}_j$) by meta-analytic method, number of studies (k), mean within-study sample size ($\bar{n}$), and population value (ρ_j).*

Note: Meta-analytic methods: m-r_O = multi-r_O, u-r_O = uni-r_O, m-Z_O = multi-Z_O, m-r_E = multi-r_E, u-r_E = uni-r_E, m-Z_E = multi-Z_E, u-Z = uni-Z.

precise; and multi-r_E and multi-Z_E, the most precise with very similar standard errors (multi-r_E was slightly more precise than multi-Z_E for smaller ρ and vice versa for larger ρ). Disparities in precision were most pronounced with smaller $\bar{n}$ and larger k. Finally, uni-r_O was always more precise than multi-r_O, usually similar in precision to multi-Z_O (although slightly more precise when $\bar{n} = 30$) and always less precise than multi-r_E and multi-Z_E.

Relative efficiency. Empirical mean square error (*MSE*), where $MSE(\hat{\rho}_j) = E[(\hat{\rho}_j - \rho_j)^2] = [\text{bias}\ (\hat{\rho}_j)]^2 + [SE(\hat{\rho}_j]^2$, is reported in terms of relative efficiency vis-à-vis $\hat{\rho}^s$, a "subject-level" estimate of the correlation matrix obtained by treating a replication's N observations as one large data set.[5] Relative efficiency for a given method's $\hat{\rho}_j$ is just $MSE(\hat{\rho}^s_j)/MSE(\hat{\rho}_j)$; values below 1.0 indicate the meta-analytic estimator is less efficient than its subject-level counterpart. Figure

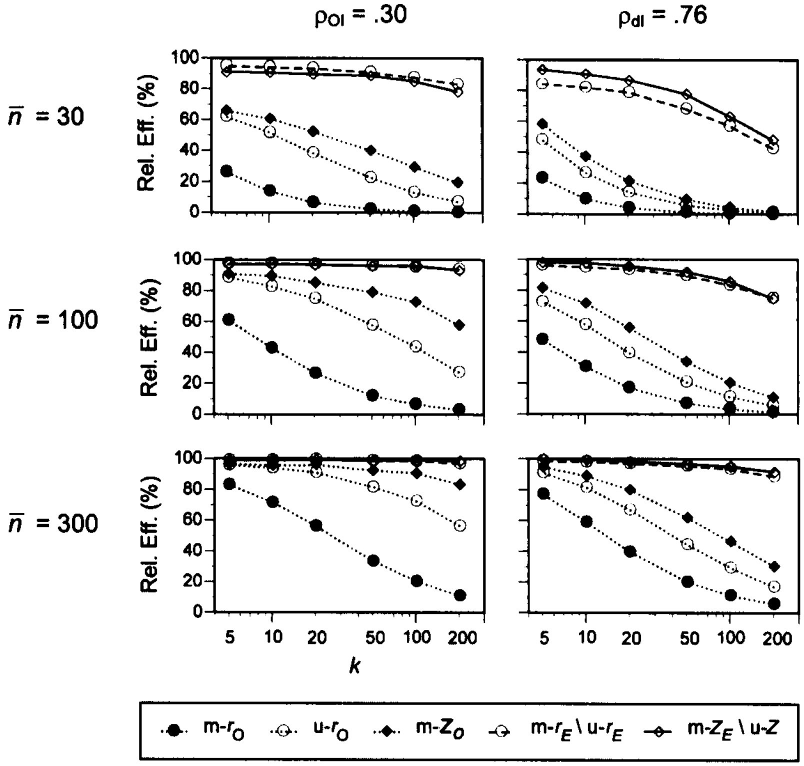

FIGURE 2. *Empirical relative efficiency (vs. subject-level estimator $\hat{\rho}_j$s) for estimated common Pearson correlation ($\hat{\rho}_j$) by meta-analytic method, number of studies (k), mean within-study sample size ($\bar{n}$), and population value (ρ_j).*

Note: See Figure 1 for methods. Rel. Eff. = relative efficiency.

2 displays relative efficiency as a percentage. Whereas bias often dominated *MSE* (i.e., |bias| > *SE*) and, hence, relative efficiency for multi-r_O and sometimes did so for the other two observed-r_i methods (especially with smaller $\bar{n}$ and larger k and ρ), standard error dominated relative efficiency for multi-Z_E and multi-r_E.[6] For most methods, especially the observed-r_i methods, relative efficiency increased with larger $\bar{n}$ or smaller k or ρ.

Among multivariate methods, multi-r_O was always least efficient, with relative efficiency typically—across the 108 $k \times \bar{n} \times \rho$ combinations—between 5% (Q_1) and 42% (Q_3) and half of the time below 15% (median). Multi-Z_O was much more efficient, with relative efficiency typically between 41% and 90% and above 72% half of the time. Multi-r_E and multi-Z_E were always most efficient and often nearly indistinguishable, with relative efficiency typically between 90% and 98% and above 96% half of the time; multi-Z_E exhibited a

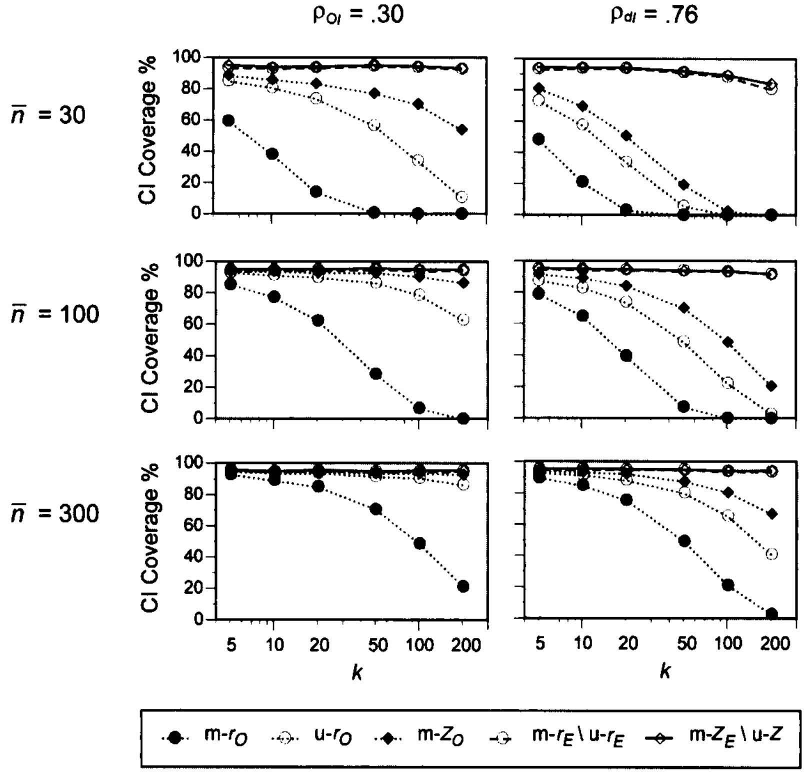

FIGURE 3. *Empirical 95%-CI coverage percentage for common Pearson correlation by meta-analytic method, number of studies (*k*), mean within-study sample size (*$\bar{n}$*), and population value (*ρ_j*).*

Note: See Figure 1 for methods. CI = confidence interval.

notable advantage for $\rho \geq .63$, especially when $\bar{n} = 30$, and a small disadvantage for $\rho \leq .41$.

Uni-r_O was always more efficient than multi-r_O, especially with smaller $\bar{n}$, but less efficient than the refined multivariate methods. Its relative efficiency was typically between 24% and 79% and more often than not below 53%.

Inference about correlation. To assess element-wise inference about correlations, each estimated correlation and its standard error (Equation 4, 9, or 13) was used to construct normal-theory central 90%, 95%, and 99% CIs: $\hat{\theta}_j \pm z_{\alpha/2}\sqrt{\hat{V}(\hat{\theta}_j)}$, where θ is ρ for Pearson-r methods and ζ for Fisher-Z methods. CI coverage rate was estimated as the percentage of replications whose CI included θ. Figure 3 displays 95%-CI coverage rates. In many respects, CI coverage exhibited similar patterns as relative efficiency. For instance, CI coverage for most methods was nearer nominal 95% with larger $\bar{n}$ or smaller k or ρ. All methods exhibited poor CI coverage for large ρ combined with small $\bar{n}$ and large k.

Comparative CI performance among multivariate methods also resembled relative efficiency. Multi-r_O CI coverage was never nominal and was often abysmal, typically between 3% and 76% and half of the time below 38%. Coverage for multi-Z_O was much better—typically between 77% and 94%—but below 90% half of the time and rarely within sampling error of nominal unless $\bar{n} = 300$. Multi-r_E and multi-Z_E were clearly superior and exhibited very similar coverage, although multi-Z_E was slightly better when $\bar{n} = 30$; their coverage was usually nominal or barely below and fell appreciably below only rarely, under the worst conditions—dropping below 90% only for $\rho \geq .63$ when $\bar{n} = 30$ and $k \geq 100$.

CI coverage for uni-r_O was markedly better than for multi-r_O but worse than for all other methods: typically between 57% and 91%, and more often than not below 82%.

Inference about homogeneity. Homogeneity tests were compared among the four multivariate methods (Equation 5) and among the three univariate methods (Equation 10) by estimating homogeneity rejection rate as the percentage of replications for which H_0 was rejected at $\alpha = .10$, .05, or .01. With fixed-effects data, these estimated Type I error rates should be near $100\alpha\%$. Note that these tests merely assess $H_0 : \boldsymbol{\rho}_i = \boldsymbol{\rho}$ or $H_0 : \rho_{ij} = \rho_j$; random-effects methods for estimating $\mathbf{T}$ or τ_j^2 are beyond the present scope (Hafdahl, 2004; Kalaian & Raudenbush, 1996).

Figures 4 and 5 display homogeneity rejection rates for the multivariate and univariate methods, respectively. For most of these seven methods, Type I error was controlled much better with larger $\bar{n}$ or smaller k. A subtler pattern among univariate methods is that with larger ρ, rejection rate tended to improve for uni-r_O but deteriorate for uni-r_E.

As for multivariate tests of omnibus homogeneity, multi-r_O performed worst once again, rarely controlling Type I error rate below 10% and letting it exceed 20% often and even hit 100%. Multi-Z_O performed better, but its rejection rate was above 17% half of the time and sometimes above 50%. Multi-r_E exhibited further improvement, with rejection rates typically between 6% and 16%, but it attained nominal 5% only when $\bar{n} = 300$ with smaller k and sometimes exceeded 20%. Multi-Z_E, with both refinements, controlled Type I errors best: never above 16%, typically between 5% and 7%, and half of the time 6% or lower; even so, its rejection rate was not within sampling error of nominal either when $\bar{n} = 30$ or when $\bar{n} = 100$ and $k \geq 50$. Even at its worst, however, when $\bar{n} = 30$, multi-Z_E controlled Type I error rate about as well as multi-r_E did when $\bar{n} = 100$ and as well as multi-Z_O did when $\bar{n} = 300$.

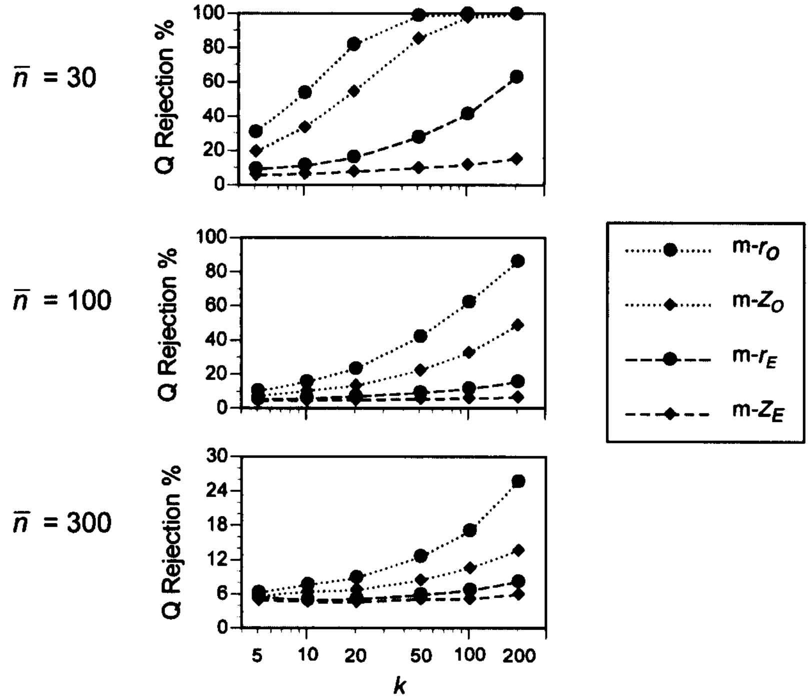

FIGURE 4. *Empirical rejection percentage for multivariate omnibus homogeneity test based on Q_M, by meta-analytic method, number of studies (k), and mean within-study sample size ($\bar{n}$).*
Note: See Figure 1 for methods.

Among the three univariate methods, uni-r_O exhibited the worst Type I error control: above 10% half of the time, and occasionally above 50%. Although uni-r_E controlled Type I error better, typically maintaining it between 5% and 8% and rarely above 15%, it exceeded nominal over half of the time. Only uni-Z maintained Type I error control between 4% and 6% in all conditions, nearly always within sampling error of nominal.

Summary

In most respects estimation and inference improved, and differences among methods diminished with larger ($\bar{n}$) or fewer (k) studies; for larger correlations (ρ), most methods' relative efficiency and CI coverage worsened. On the basis of their relative performance under conditions examined here, the multivariate meta-analytic methods can be ordered approximately from better (i.e., lower bias and standard error, higher relative efficiency, CI coverage and homogeneity

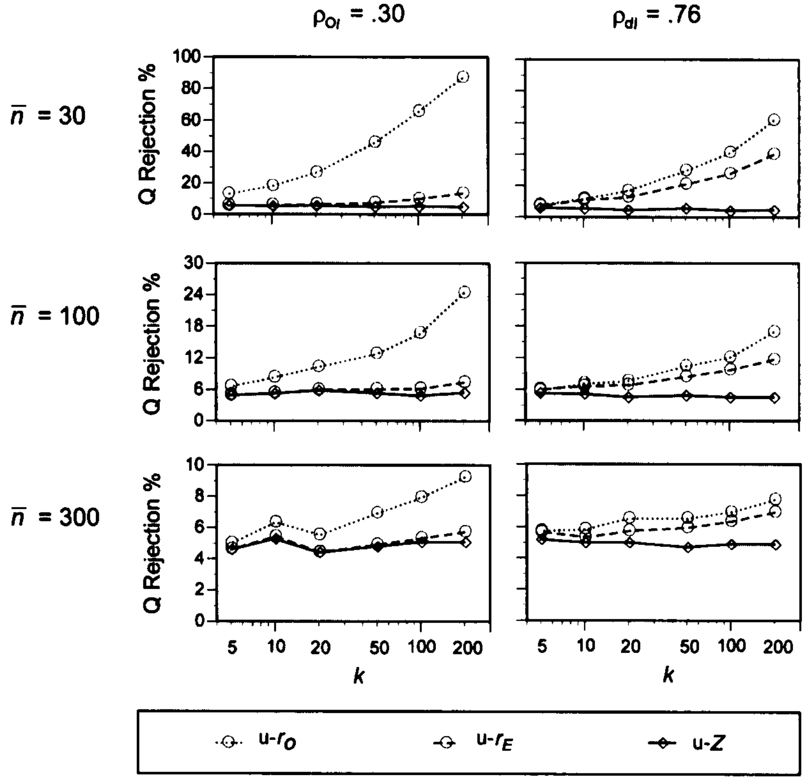

FIGURE 5. *Empirical rejection percentage for univariate element-wise homogeneity test based on Q_U, by meta-analytic method, number of studies (*k*), mean within-study sample size (*n̄*), and population value of common Pearson correlation (*ρ_j*).*
Note: See Figure 1 for methods.

rejection rates nearer nominal) to worse as follows: multi-Z_E, multi-r_E, multi-Z_O, multi-r_O. Element-wise results for uni-r_O were typically better than for multi-r_O but worse than for all other methods. Some aspects of this ordering are sharper than others: The three observed-r_i methods were nearly always inferior, and multi-r_O nearly always performed most poorly and often substantially so, whereas in some conditions multi-r_E outperformed multi-Z_E or uni-r_O was superior to multi-Z_O. This ordering is related to the importance of the three distinctions among methods:

1. Using estimated versus observed correlations in the sampling (co)variances proved to be most influential: The three estimated-ρ methods (multi-Z_E, multi-r_E, uni-r_E) always outperformed their observed-r_i counterparts (multi-Z_O, multi-r_O, uni-r_O), often considerably.

2. Methods based on Fisher-Z correlations typically performed better than those based on Pearson-r correlations: Multi-Z_O performed better than multi-r_O on all criteria in most conditions; although multi-r_E was sometimes slightly less biased and more efficient than multi-Z_E, the multi-Z_E homogeneity test was notably superior; and uni-Z always performed better than uni-r_O and about equally well as or, for homogeneity tests, appreciably better than uni-r_E.
3. The univariate (WLS) versus multivariate (GLS) distinction—pertinent only to element-wise results for observed-r_i methods—favored univariate methods: Estimation and inference for uni-r_O and uni-Z were always better than for multi-r_O and multi-Z_O, respectively.

Discussion

The primary purpose of the Monte Carlo study was to assess two refinements to an early GLS method for multivariate meta-analysis of correlation matrices. Each refinement improved on multi-r_O's estimation and inference with respect to all criteria considered. Although simply analyzing Fisher-Z correlations improved performance, multi-Z_O still behaved poorly in many respects. Using estimated instead of observed correlations in the conditional (co)variances (i.e., multi-r_E) improved performance further, but Type I error rates for homogeneity tests were still unacceptably high. Finally, employing both refinements (i.e., multi-Z_E) brought the homogeneity Type I error rate near nominal, except with small or many primary studies.

These findings suggest that sampling error in observed correlations is the most important reason multi-r_O performs poorly, as Becker and Fahrbach (1994) posited. Fisher's Z-transformation alone did improve performance and was necessary to maintain multivariate homogeneity rejection rates at nominal—due perhaps to homogeneity tests' greater reliance on normality via their quadratic forms—but it was less effective than the estimated-ρ refinement. On a related note, when $\bar{n} = 30$, the observed-r_i correlation estimates exhibited excessive positive kurtosis and (except for multi-Z_O) positive skewness, especially for larger k; these substantial violations of WLS and GLS normality assumptions corroborate the observed-r_i methods' especially poor CI coverage and homogeneity Type I error control with many small studies.

The Monte Carlo findings also replicate and extend Hafdahl's (2001) results from fixed-effects meta-analysis of fixed-effects data. Namely, multi-r_O exhibited inferior element-wise estimation and inference compared to two standard univariate methods (uni-r_O and uni-Z). Moreover, with smaller primary studies and a smaller correlation matrix than in Hafdahl (2001), uni-Z often outperformed uni-r_O substantially, especially in terms of homogeneity tests. The present findings also agree in essence with those from previous simulations (Becker & Fahrbach, 1994; M. W. L. Cheung & Chan, 2005; S. F. Cheung, 2000; Furlow & Beretvas, 2005) regarding comparisons among similar methods under conditions comparable to those examined here.

Recommendations

On the basis of the Monte Carlo results, three recommendations can be offered for synthesists planning fixed-effects meta-analysis of correlation matrices:

1. The most advisable strategy is to substitute estimated correlations for ρ_i elements in the conditional (co)variances (Equation 2, 7, or 11) and analyze either Fisher-Z correlations (i.e., multi-Z_E or uni-Z) or their Pearson-r counterparts (i.e., multi-r_E or uni-r_E). Not only do these estimated-ρ methods exhibit superior estimation and inference performance, but with complete data from every study, they are computationally easier. Choosing between multi-r_E versus multi-Z_E depends somewhat on one's circumstances and focal results: Although each strategy yields slightly more accurate, precise, and efficient point estimates under certain conditions, multi-Z_E more often yields nominal CI coverage and controls homogeneity Type I errors substantially better—albeit not always within nominal (see S. F. Cheung, 2000, for alternative tests).

2. Substituting observed correlations for ρ_i elements in conditional (co)variances and employing either GLS with Z-transformed correlations (i.e., multi-Z_O) or WLS with Pearson correlations (i.e., uni-r_O) may yield acceptable results under some conditions but cannot be endorsed generally. Except with very few very large primary studies, performance for both methods deteriorates, and their Type I error control for homogeneity tests is quite poor.

3. Avoid substituting observed correlations for ρ_i elements in the conditional (co)variances and employing GLS with Pearson correlations (i.e., multi-r_O). This original strategy's estimation and inference results are often quite unacceptable and at best perceptibly inferior to those from other methods. Note that although this recommendation concurs with Becker's (2000) suggestion to avoid using GLS with untransformed correlations, GLS of Pearson correlations with the estimated-ρ refinement (i.e., multi-r_E) does have some merit.

Conditional (Co) Variance Denominator

Some authors advocate using $n_i - 1$ instead of n_i in the conditional variances for Pearson correlations (Equation 7; see Note 1). When the simulation was rerun with this modification for covariances (Equation 2) as well as variances, the affected methods—multi-r_E, multi-r_O, uni-r_E, and uni-r_O—tended to improve at least slightly on all evaluation criteria, especially when $\bar{n} = 30$. Only homogeneity rejection rates changed materially, however; those for the univariate case agreed closely with corresponding results from Alexander et al.'s (1989) comparisons of Type I error rate between uni-r_E (with $n_i - 1$) and uni-Z under similar conditions. These changes were not sufficient to alter the above patterns or recommendations.

Future Directions

The multivariate methods advocated above provide synthesists with dependable fixed-effects meta-analysis strategies for synthesizing homogeneous correlation matrices under most conditions, the notable exception being inference—especially homogeneity tests—with many small studies. Some unresolved issues warrant further consideration, however.

As with most Monte Carlo studies, several circumstances remain unexamined. Two departures likely to arise in practice are especially pertinent to the generalizability of these Monte Carlo results. First, certain methods might behave differently with larger correlation matrices or particular patterns of values (e.g., negative correlations, certain factor or simplex structures). Second, the performance of fixed-effects methods with random-effects data warrants study, because some meta-analysts use fixed-effects techniques despite appreciable between-studies heterogeneity; although often ill-advised, this may be defensible for generalizing to a narrower universe (Hedges & Vevea, 1998). Predicting the impact of these changes is difficult, but there is little reason to suspect the above advice would change.

Parallel refinements for more complex and flexible multivariate models would benefit meta-analysts. Three such models are immediately apparent: random-effects models, fixed-effects models with one or more study-level covariates, and mixed-effects models that combine these two (Kalaian & Raudenbush, 1996). Hafdahl (2001) also found that a random-effects analogue to multi-r_O performed worse than univariate random-effects methods. Given the ubiquity of between-studies variation in meta-analytic data and the theoretical and practical value of relating it to study features, models that incorporate this heterogeneity are indispensable. Extending the present estimated-ρ refinement to these more complex models is less clear-cut, however, because ρ_i depends on the model for between-studies heterogeneity. Preliminary results for such extensions to the random-effects case are promising (Hafdahl, 2004).

Beyond the development of multivariate methods that perform well under ideal conditions, additional work is needed to adapt these methods to the characteristic untidiness of real meta-analytic data. Becker and Schram (1994) and Shadish (1996) have described several such challenges, with an emphasis on multivariate explanatory models. Typically, problems encountered in univariate meta-analysis (e.g., missing data, artifactual attenuation, assumption violations) are compounded in multivariate meta-analysis and demand more difficult solutions. The potential yield of sound multivariate methods for meta-analyzing correlation matrices, however, merits investment in their further improvement.

Appendix
Approximations to Estimate Pearson-r Results from Fisher-Z Results

First I describe a technique for estimating a single Pearson correlation and this estimate's sampling variance from the corresponding Fisher-Z estimates, based on a Taylor series approximation. This technique extends readily to several correlations.

Univariate Case

Suppose we have $\hat{\zeta}$ and $\hat{V}(\hat{\zeta})$, which estimate the Fisher-Z correlation ζ and this estimate's sampling variance $V(\hat{\zeta})$, such as from meta-analyzing Fisher-Z correlations. Suppose further we want $\hat{\rho}$ and $\hat{V}(\hat{\rho})$, these estimates' Pearson-r counterparts. First note that $\rho = \tanh(\zeta)$, so the derivative of ρ with respect to ζ is

$$\frac{d\rho}{d\zeta} = \frac{d}{d\zeta}\tanh(\zeta) = \mathrm{sech}^2(\zeta) = \left(\frac{2}{e^{\zeta} + e^{-\zeta}}\right)^2, \tag{A1}$$

where $\mathrm{sech}(a)$ is the hyperbolic secant of a. Now, we typically estimate ρ using $\hat{\rho} = \tanh(\hat{\zeta})$, and a large-sample Taylor series approximation leads to the following estimator of $V(\hat{\rho})$:

$$\hat{V}(\hat{\rho}) = \left(\frac{d\hat{\rho}}{d\hat{\zeta}}\right)^2 \hat{V}(\hat{\zeta}) = \mathrm{sech}^4(\hat{\zeta})\hat{V}(\hat{\zeta}). \tag{A2}$$

Multivariate Case

Suppose we have p^* estimates of p^* Fisher-Z correlations and these estimates' $p^* \times p^*$ covariance matrix, $\hat{\boldsymbol{\zeta}}$ and $\hat{\mathbf{V}}(\hat{\boldsymbol{\zeta}})$, such as from multivariate meta-analysis of Fisher-Z correlation matrices, and we want these estimates' Pearson counterparts, $\hat{\boldsymbol{\rho}}$ and $\hat{\mathbf{V}}(\hat{\boldsymbol{\rho}})$. The typical elements of $\hat{\boldsymbol{\zeta}}$ and $\hat{\boldsymbol{\rho}}$ are $\hat{\zeta}_{ab}$ and $\hat{\rho}_{ab}$, whose corresponding diagonal elements in $\hat{\mathbf{V}}(\hat{\boldsymbol{\zeta}})$ and $\hat{\mathbf{V}}(\hat{\boldsymbol{\rho}})$ are $\hat{V}(\hat{\zeta}_{ab})$ and $\hat{V}(\hat{\rho}_{ab})$, respectively; the typical off-diagonal elements of $\hat{\mathbf{V}}(\hat{\boldsymbol{\zeta}})$ and $\hat{\mathbf{V}}(\hat{\boldsymbol{\rho}})$ are $\hat{\mathrm{Cov}}(\hat{\zeta}_{ab}, \hat{\zeta}_{cd})$ and $\hat{\mathrm{Cov}}(\hat{\rho}_{ab}, \hat{\rho}_{cd})$.

Extending the univariate case is straightforward. Let $\mathbf{D}$ be a $p^* \times p^*$ matrix of partial derivatives with typical element $\partial\hat{\rho}_{ab}/\partial\hat{\zeta}_{ab} = \partial\tanh(\hat{\zeta}_{ab})/\partial\hat{\zeta}_{ab} = \mathrm{sech}^2(hat\zeta_{ab})$. (Because $\partial\hat{\rho}_{ab}/\partial\hat{\zeta}_{cd} = 0$, $\mathbf{D}$ is diagonal.) Then the multivariate delta method leads to the following estimator of $\mathbf{V}(\hat{\boldsymbol{\rho}})$:

$$\hat{\mathbf{V}}(\hat{\boldsymbol{\rho}}) = \mathbf{D}\hat{\mathbf{V}}(\hat{\boldsymbol{\zeta}})\mathbf{D}. \tag{A3}$$

In particular, the typical diagonal and off-diagonal elements of $\hat{\mathbf{V}}(\hat{\boldsymbol{\rho}})$ are, respectively,

$$\hat{V}(\hat{\rho}_{ab}) = \mathrm{sech}^4(\hat{\zeta}_{ab})\hat{V}(\hat{\zeta}_{ab}), \tag{A3a}$$

which is just the univariate variance estimate in Equation A2, and

$$\hat{\mathrm{Cov}}(\hat{\rho}_{ab}, \hat{\rho}_{cd}) = \mathrm{sech}^2(\hat{\zeta}_{ab})\,\mathrm{sech}^2(\hat{\zeta}_{cd})\hat{\mathrm{Cov}}(\hat{\zeta}_{ab}, \hat{\zeta}_{cd}). \tag{A3b}$$

Notes

1. An alternative expression for v_{ij} replaces n_i with $n_i - 1$ (Hedges, 1989; Hunter & Schmidt, 2004) to improve the approximation for small samples. Using n_i maintains consistency with the multivariate case and is addressed in the Discussion.

2. In the meta-analysis literature, τ^2 is also denoted by Var(ρ), V(ρ), V_ρ(SD_ρ in square root form), or σ^2_ρ, and its estimator is often denoted by S^2_ρ in the validity generalization literature.

3. Fisher's Z-transformation of ρ is $\zeta = \log[(1+\rho)/(1-\rho)]/2 = \tanh^{-1}(\rho)$, where $\log(a)$ and $\tanh^{-1}(a)$ denote the natural logarithm and hyperbolic arctangent of a, respectively. Transforming ζ to ρ is accomplished by $\rho = (e^{2\zeta} - 1)/(e^{2\zeta} + 1) = \tanh(\zeta)$. For a sample, replace ζ and ρ with Z and r, respectively.

4. Replacing $\boldsymbol{\rho}_i$ elements with any real-valued constants instead of $\boldsymbol{\rho}$ elements, such as from $\hat{\boldsymbol{\rho}}$, also leads to Equations 12 and 13, provided $\mathbf{V}_i$ is positive definite.

5. This subject-level estimator is rarely available to meta-analysts, but it provides a useful basis for comparison as an ideal scenario where N subjects' information is not degraded by partitioning them into studies (Viana, 1993). This estimator was essentially unbiased ($|\text{bias}| < .00128$) and more precise than all fixed-effects meta-analytic methods considered here.

6. Standard error (SE) dominated the subject-level estimator's mean square error (MSE): Under all conditions $SE^2/MSE > .998$. Also, because relative efficiency was always computed against subject-level estimates, greater relative efficiency for one method versus another implies the former is more efficient.

References

Alexander, R. A., Scozzaro, M. J., & Borodkin, L. J. (1989). Statistical and empirical examination of the chi-square test for homogeneity of correlations in meta-analysis. *Psychological Bulletin, 106*, 329-331.

Becker, B. J. (1992a, April). *Missing data and the synthesis of correlation matrices.* Paper presented at the meeting of the American Educational Research Association, San Francisco, CA.

Becker, B. J. (1992b). Using results from replicated studies to estimate linear models. *Journal of Educational Statistics, 17*, 341-362.

Becker, B. J. (1995). Corrections to "Using results from replicated studies to estimate linear models." *Journal of Educational and Behavioral Statistics, 20*, 100-102.

Becker, B. J. (2000). Multivariate meta-analysis. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 499-525). San Diego, CA: Academic Press.

Becker, B. J., & Fahrbach (1994, April). *A comparison of approaches to the synthesis of correlation matrices.* Paper presented at the meeting of the American Educational Research Association, New Orleans, LA.

Becker, B. J., & Schram, C. M. (1994). Examining explanatory models through research synthesis. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 357-381). New York: Russell Sage Foundation.

Brown, S. A., & Hedges, L. V. (1994). Predicting metabolic control in diabetes: A pilot study using meta-analysis to estimate a linear model. *Nursing Research, 43*, 362-368.

Burke, M. J. (1984). Validity generalization: A review and critique of the correlation model. *Personnel Psychology, 37*, 93-115.

Cheung, M. W. L., & Chan, W. (2005). Meta-analytic structural equation modeling: A two-stage approach. *Psychological Methods, 10*, 40-64.

Cheung, S. F. (2000). Examining solutions to two practical issues in meta-analysis: Dependent correlations and missing data in correlation matrices. *Dissertation Abstracts International, 61*(08), 4469B. (UMI No. AAI9984691)

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

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.

Emmers-Sommer, T. M., & Allen, M. (1999). Variables related to sexual coercion: A path model. *Journal of Social and Personal Relationships, 16*, 659-678.

Furlow, C. F., & Beretvas, S. N. (2005). Meta-analytic methods of pooling correlation matrices for structural equation modeling under different patterns of missing data. *Psychological Methods, 10*, 227-254.

Gleser, L. J., & Olkin, I. (1994). Stochastically dependent effect sizes. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 339-355). New York: Russell Sage Foundation.

Grube, B. S., Bilder, R. M., & Goldman, R. S. (1998). Meta-analysis of symptom factors in schizophrenia. *Schizophrenia Research, 31*, 113-120.

Hafdahl, A. R. (2001). *Multivariate meta-analysis for exploratory factor analytic research* (doctoral dissertation, the University of North Carolina at Chapel Hill, 2001). *Dissertation Abstracts International, 62*(08), 3843B.

Hafdahl, A. R. (2004, June). *Refinements for random-effects meta-analysis of correlation matrices.* Paper presented at the meeting of the Psychometric Society, Monterey, CA.

Hamilton, M. A. (1998). Message variables that mediate and moderate the effect of equivocal language on source credibility. *Journal of Language and Social Psychology, 17*, 109-143.

Hedges, L. V. (1983). Combining independent estimators in research synthesis. *British Journal of Mathematical and Statistical Psychology, 36*, 123-131.

Hedges, L. V. (1988). The meta-analysis of test validity studies: Some new approaches. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 191-212). Hillsdale, NJ: Lawrence Erlbaum.

Hedges, L. V. (1989). An unbiased correction for sampling error in validity generalization studies. *Journal of Applied Psychology, 74*, 469-477.

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

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

Higgins, E. T. (1987). Self-discrepancy: A theory relating self and affect. *Psychological Review, 94*, 319-340.

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

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

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

Kalaian, H. A., & Raudenbush, S. W. (1996). A multivariate mixed linear model for meta-analysis. *Psychological Methods, 1*, 227-235.

Kavale, K. (1982). Meta-analysis of the relationship between visual perceptual skills and reading achievement. *Journal of Learning Disabilities, 15*, 42-51.

Klein, H. J., Wesson, M. J., Hollenbeck, J. R., Wright, P. M., & DeShon, R. P. (2001). The assessment of goal commitment: A measurement model meta-analysis. *Organizational Behavior and Human Decision Processes, 85*, 32-55.

Law, K. S. (1995). The use of Fisher's Z in Schmidt-Hunter-type meta-analyses. *Journal of Educational and Behavioral Statistics, 20*, 287-306.

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

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis.* Thousand Oaks, CA: Sage.

Olkin, I., & Finn, J. (1990). Testing correlated correlations. *Psychological Bulletin, 108*, 330-333.

Olkin, I., & Finn, J. D. (1995). Correlations redux. *Psychological Bulletin, 118*, 155-164.

Olkin, I., & Siotani, M. (1976). Asymptotic distribution of functions of a correlation matrix. In S. Ikeda, T. Hayakawa, H. Hudimoto, M. Okamoto, M. Siotani, & S. Yamamoto (Eds.), *Essays in probability and statistics: A volume in honor of Professor Junjiro Ogawa* (pp. 235-251). Wakaba, Tokyo: Shinko Tsusho.

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

Roesch, S. C., & Weiner, B. (2001). A meta-analytic review of coping with illness: Do causal attributions matter? *Journal of Psychosomatic Research, 50*, 205-219.

Schmidt, F. L., Hunter, J. E., & Raju, N. S. (1988). Validity generalization and situational specificity: A second look at the 75% rule and Fisher's *z* transformation. *Journal of Applied Psychology, 73*, 665-672.

Shadish, W. R. (1996). Meta-analysis and the exploration of causal mediating processes: A primer of examples, methods, and issues. *Psychological Methods, 1*, 47-65.

Shadish, W. R., & Haddock, C. K. (1994). Combining estimates of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 261-281). New York: Russell Sage Foundation.

Silver, N. C., & Dunlap, W. P. (1987). Averaging correlation coefficients: Should Fisher's *z* transformation be used? *Journal of Applied Psychology, 72*, 146-148.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87*, 245-251.

Viana, M. A. (1993). On a criterion for combining correlational data. *Journal of Educational Statistics, 18*, 261-270.

Viswesvaran, C., & Ones, D. S. (1995). Theory testing: Combining psychometric meta-analysis and structural equations modeling. *Personnel Psychology, 48*, 865-885.

Author

ADAM R. HAFDAHL is a graduate student, Department of Mathematics, at the Washington University in St. Louis, Campus Box 1146, St. Louis, MO 63130; arhafdah@wustl.edu. His area of specialization is meta-analysis for quantitative research synthesis, especially techniques related to correlations, multivariate indices of effect size, and practical impediments such as aberrant cases and missing or degraded data.

Manuscript received December 10, 2004
Accepted August 27, 2005