Psychological Methods
2014, Vol. 19, No. 2, 211–229
© 2013 American Psychological Association
1082-989X/14/$12.00 DOI: 10.1037/a0032968

Modeling Dependent Effect Sizes With Three-Level Meta-Analyses: A Structural Equation Modeling Approach

Mike W.-L. Cheung
National University of Singapore

Meta-analysis is an indispensable tool used to synthesize research findings in the social, educational, medical, management, and behavioral sciences. Most meta-analytic models assume independence among effect sizes. However, effect sizes can be dependent for various reasons. For example, studies might report multiple effect sizes on the same construct, and effect sizes reported by participants from the same cultural group are likely to be more similar than those reported by other cultural groups. This article reviews the problems and common methods to handle dependent effect sizes. The objective of this article is to demonstrate how 3-level meta-analyses can be used to model dependent effect sizes. The advantages of the structural equation modeling approach over the multilevel approach with regard to conducting a 3-level meta-analysis are discussed. This article also seeks to extend the key concepts of Q statistics, I^2, and R^2 from 2-level meta-analyses to 3-level meta-analyses. The proposed procedures are implemented using the open source metaSEM package for the R statistical environment. Two real data sets are used to illustrate these procedures. New research directions related to 3-level meta-analyses are discussed.

Keywords: meta-analysis, structural equation modeling, three-level model, maximum likelihood, R statistical environment

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Glass (1976) coined the term “meta-analysis” to describe “the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings” (p. 3). Reviewers who conduct meta-analyses are able to (a) address whether the effect sizes from a pool of empirical studies are consistent, (b) determine the average effect size, (c) establish the degree of heterogeneity due to study-specific random effects, and (d) specify whether the study characteristics explain the variability of the effect sizes. After Glass was introduced to the social sciences, meta-analysis has gradually become an indispensable tool across many disciplines including psychology, education, management, and the medical sciences.

Most meta-analytic procedures assume independence among the effect sizes. This assumption might not be realistic for many research settings. In general, the degree of dependence can be classified as either known (i.e., it can be estimated via summary statistics) or unknown. When the degree of dependence is known, multivariate meta-analyses can be used to model this dependence (e.g., Becker, 1992; M. W.-L. Cheung, in press; Gleser & Olkin, 1994; Jackson, Riley, & White, 2011; Nam, Mengersen, & Garthwaite, 2003; Raudenbush, Becker, & Kalaian, 1988; Stevens & Taylor, 2009). This type of dependence is not discussed further because it is not the focus of this article. Several procedures have been suggested to address unknown dependence (e.g., Borenstein, Hedges, Higgins, & Rothstein, 2009; Cooper, 2010). This article uses the three-level meta-analysis (e.g., Konstantopoulos, 2011; Van den Noortgate, López-López, Marín-Martínez, & Sánchez-Meca, 2012) to model unknown dependence.

Traditional meta-analyses can be considered two-level models with participants at Level 1 and the studies at Level 2. The intra-class correlation (ICC) indicates the degree of between-study variation (i.e., dependence) with regard to the total variance. When dependence does not exist, the ICC is 0. Three-level meta-analyses add another level that allows the effect sizes to be correlated within a cluster. Rather than one ICC, Level 2 and Level 3 ICCs indicate the variations at their respective levels. Suppose that multiple effect sizes are reported for each study; thus, there are within- (Level 2) and between-study (Level 3) variations. A three-level model is equivalent to a two-level model when the ICC at Level 3 is 0. The literature on three-level meta-analyses primarily focuses on parameter estimates and hypothesis testing (e.g., Konstantopoulos, 2011; Marsh, Bornmann, Mutz, Daniel, & O'Mara, 2009; Van den Noortgate et al., 2012). Indices such as the Q statistics, I^2, and R^2, which are crucial to a meta-analysis, are missing in the literature. This article fills this gap by extending these indices to three-level meta-analyses.

This article had several objectives. The first goal was to review the problems and common methods to handle dependent effect sizes in traditional meta-analyses. The second objective was to introduce how the three-level meta-analysis addresses dependent effect sizes using structural equation modeling (SEM) perspective. This article also highlights the advantages of SEM over multilevel modeling with regard to conducting a three-level meta-analysis. For example, SEM places flexible constraints on parameters, con-

This article was published Online First July 8, 2013.

This research was supported by the Academic Research Fund Tier 1 (Grant R581-000-111-112) from the Ministry of Education, Singapore.

Correspondence concerning this article should be addressed to Mike W.-L. Cheung, Department of Psychology, Faculty of Arts and Social Sciences, National University of Singapore, Block AS4, Level 2, 9 Arts Link, Singapore 117570. E-mail: mikewlcheung@nus.edu.sg

structs better confidence intervals (CIs) via the likelihood-based approach and handles missing covariates using the full information maximum likelihood (FIML).

In a series of articles (M. W.-L. Cheung, 2008, 2010, 2013, in press), I proposed a SEM framework to formulate univariate and multivariate meta-analyses using the maximum likelihood (ML) and restricted (or residual) ML (REML) estimation methods. This article extends this line of research to unknown dependence among effect sizes. Mathematical models of two- and three-level meta-analyses can be succinctly represented as graphical models using SEM. The three-level meta-analysis techniques discussed in this article were implemented using the metaSEM package written by the author (M. W.-L. Cheung, 2012). These advanced techniques are readily accessible to applied users.

The rest of the article is organized as follows. In the next section, the traditional meta-analytic models, problems, and common methods to handle dependent effect sizes are reviewed. Next, two-level and three-level formulation of meta-analyses and their SEM implementations are introduced. The issues related to extending two-level to three-level meta-analyses and the advantages of the SEM approach are discussed. Two real data sets are used to illustrate these procedures. The last section discusses future directions with regard to three-level meta-analyses.

Traditional Meta-Analytic Models

Let y_i be a generic effect size in the ith study. y_i can be any effect size, such as an odds ratio, raw mean difference, standardized mean difference, correlation coefficient, or its Fisher's z transformed score. y_i is assumed to be approximately normally distributed with a known variance of v_i (see Borenstein et al., 2009, for the formulas to calculate v_i for various effect sizes). The random-effects model is

$$y_i = \beta_0 + u_i + e_i, \tag{1}$$

where $\mathrm{Var}(e_i) = v_i$ is the known sampling variance in the ith study, β_0 is the average population effect, and $\mathrm{Var}(u_i) = \tau^2$ is the study-specific heterogeneity that must be estimated (Borenstein, Hedges, Higgins, & Rothstein, 2010).

Although fixed-effects models and random-effects models are conceptually different, mathematically, the former is a specific type of the latter by fixing $\tau^2 = 0$. Fixed-effects models are used for conditional inferences based on selected studies. They are intended to draw conclusions only for the studies included in the meta-analysis. In other words, the findings cannot be generalized beyond the studies included in the analysis. On the other hand, random-effects models allow for unconditional inferences to be generalized beyond the studies included under the assumption that the studies were randomly selected (see Borenstein et al., 2010; Hedges & Vevea, 1998). The current article primarily focuses on random-effects models because fixed-effects models can be easily fit by fixing $\tau^2 = 0$.

When the estimated τ^2 is large, the population effect sizes are likely to be heterogeneous. Reviewers might want to explain this heterogeneity using study characteristics as predictors. Suppose covariate x_i exists in the ith study; the mixed-effects model is

$$y_i = \beta_0 + \beta_1 x_i + u_i + e_i, \tag{2}$$

where β_0 and β_1 are the intercept and the regression coefficient, respectively, and $\mathrm{Var}(u_i) = \tau^2$ is the residual heterogeneity after controlling for the covariate.

Handling Dependent Effect Sizes in Traditional Meta-Analysis

In his chapter on statistical considerations in meta-analyses, Hedges (2009) listed dependence among the effect sizes as one of the critical considerations in conducting a meta-analysis. Matt and Cook (2009) also considered a lack of statistical independence on the effect sizes as one of the threats to the validity of meta-analyses. I review the problems and common methods to handle dependent effect sizes in the following section.

Causes of Dependence Among Effect Sizes

Dependence among effect sizes can be introduced by either the researchers who conducted the primary studies or the reviewers who are conducting the meta-analysis. The researchers who conducted the primary studies might have collected data from multiple sites. They might have compared different treatment groups to the same control group or have used multiple measures for the same construct. Single case studies might have reported multiple effect sizes with regard to the same participants (Owens & Ferron, 2012). Studies might also have reported multiple effect sizes for similar constructs, over multiple time points or conditions; meanwhile, the correlations among the effect sizes remain unknown (Van den Bussche, Van den Noortgate, & Reynvoet, 2009). Effect sizes are not independent when summary statistics are extracted for a meta-analysis.

Let us consider Nguyen and Benet-Martínez's (2013) meta-analysis as an example. These authors extracted 935 correlation coefficients between biculturalism (e.g., behavior, values and identity) and adjustment (e.g., life satisfaction and grades) from 141 studies. Because each study reported multiple effect sizes, it is not reasonable to assume that the effect sizes are independent.

On the other hand, the reviewers who conducted the meta-analysis might also have introduced dependence. For example, reviewers who conduct cross-cultural meta-analyses might hypothesize that culture plays an important role on psychological processes. The effect sizes reported by participants from the same cultural group are likely to be more similar than those reported by other cultural groups. Although different researchers independently conducted the primary studies, these studies will be naturally grouped by culture in the meta-analysis (e.g., Fischer & Boer, 2011; Hanke & Fischer, 2012). In another example, the effect sizes reported by the same research team or authors might be more similar compared with those reported by other research teams or authors (Cooper, 2010; Rosenthal & DiMatteo, 2001; Shin, 2009). The reviewers who conduct the meta-analysis might group studies by the research teams.

Importantly, the nested structure depends on the research questions. A reviewer who conducts a cross-cultural meta-analysis might consider studies nested within cultures by ignoring the effect of research teams, whereas a reviewer who conducts a meta-analysis on the effects of research teams might

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

ignore cultural effects. Before choosing a statistical model for the meta-analyses, reviewers should consider how different sources of dependence affect the validity of the model and the research questions.

Common Methods for Handling Dependent Effect Sizes

Several strategies have been suggested for handling dependent effect sizes (see Chapter 24 in Borenstein et al., 2009; Cooper, 2010). These strategies have pros and cons, which I briefly review here.

Ignoring dependence. One incorrect strategy is simply to ignore the dependence and analyze the data as if they were independent. Much work has been conducted on the effects of ignoring this dependence in multilevel modeling (e.g., Hox, 2010; Snijders & Bosker, 2012). First, studies with multiple effect sizes have a larger influence than those with one effect size. Fixed-effects parameter estimates, such as the intercept and regression coefficients, might be biased against studies with fewer reported effect sizes. Second, dependent effect sizes appear more "similar" than independent effect sizes. The associated standard errors (*SE*s) of parameter estimates are likely under-estimated due to an under estimation of uncertainty. Thus, the associated significance tests are inflated. Ignoring dependence is generally not acceptable for meta-analyses (Becker, Hedges, & Pigott, 2004).

Averaging dependent effect sizes within studies. One reasonable approach to removing dependence is averaging the dependent effect sizes into a single effect size and then using this average in subsequent analyses. This approach is also known as "aggregation." Several procedures have been suggested for averaging dependent effect sizes (e.g., S. F. Cheung & Chan, 2004; Marín-Martínez & Sánchez-Meca, 1999; Rosenthal & Rubin, 1986). One simple method is to weigh the dependent effect sizes equally within a study. Suppose two dependent effect sizes y_{1j} and y_{2j} exist with their sampling variances (v_{1j} and v_{2j}) and covariance (v_{12j}) in the *j*th study. If v_{12j} is unknown, an ad hoc estimate based on expert knowledge or other sources could be used. We could compute an average effect size $y_j = 0.5(y_{1j} + y_{2j})$ and its sampling variance $v_j = 0.25(v_{1j} + v_{2j} + 2v_{12j})$ (e.g., Borenstein et al., 2009; Gleser & Olkin, 1994). This procedure has several advantages. The averaged effect sizes and their sampling variances can be used directly in a traditional meta-analysis. Second, the averaged effect sizes can also improve the precision of the estimates when dependent effect sizes represent the same construct.

However, this approach has certain associated limitations. Using Nguyen and Benet-Martínez's (2013) meta-analysis as an example, their 935 effect sizes were reduced to 141 effect sizes in the analysis. The statistical power of the tests might have been lowered because some information in the effect sizes was lost (Hedges & Pigott, 2001, 2004). Although some reviewers (e.g., Hunter & Schmidt, 2004) do not rely on the use of significance tests in meta-analyses, less data availability also means that the parameter estimates will be less precise with larger *SE*s. Next, this procedure might limit the research questions that can be addressed in the meta-analysis. For example, reviewers might want to investigate the effectiveness of a teaching program on students, and studies often report teaching effectiveness based on student reports y_{1j} and teacher self-evaluations y_{2j}. If the reviewers average the effect sizes reported by students and teachers, then the differences between student reports and teacher evaluations cannot be compared. Lastly, a fixed-effects model is used when averaging effect sizes within studies. These effect sizes are assumed to be homogeneous within studies, but this assumption has been questioned (e.g., Marín-Martínez & Sánchez-Meca, 1999).

Selecting one effect size per study. Another approach is to select only one effect size per study. This method is also known as "elimination." Because only one effect size per study is chosen, the effect sizes are independent. This approach is useful when the reviewers believe that selecting one effect size per study is better than using all the available effect sizes. For example, suppose certain findings suggest that student reports of teaching effectiveness are inaccurate due to their biased attitudes toward teachers. In this case, a meta-analysis of the effect sizes of only teacher evaluations might be best.

This approach has several limitations. Similar to the problems listed in the averaging approach, this method affects the statistical power of the meta-analysis and the precision of the parameter estimates because some effect sizes have been excluded. Using the previous example of the teaching program, the reviewers might need to choose between effect sizes of teaching effectiveness based on student reports y_{1j} or teacher evaluations y_{2j}. This forced choice limits the reviewers' ability to test study characteristics using a mixed-effects model because not all effect sizes were included in the meta-analysis.

From the above descriptions, it is clear that selecting or excluding certain effect sizes depends on the research questions. If the reviewers believe that not all of the effect sizes are equally good or relevant to the research questions, then excluding certain effect sizes is appropriate. If the goal is to remove the dependence among the effect sizes, however, this method is not the best approach to handling dependence among effect sizes.

Shifting the unit of analysis. Cooper (2010) proposed an approach called "shifting the unit of analysis." The basic idea of this method is to select a unit of analysis and then average the effect sizes within those units. For example, when an overall estimate of the pooled effect size is required, the unit of analysis is the studies. The dependent effect sizes are averaged within their respective studies before the analysis. Because each study contributes only one effect size, the average effect is calculated using independent samples. When examining study characteristics such as the effect due to gender, the dependent effect sizes are averaged within each gender. This approach can minimize the effect of dependence. This approach might be useful when reviewers' research questions must be addressed using different sets of effect sizes. However, this method inherits some of the limitations from the aggregation approach. For example, it does not resolve issues such as the loss of information, and it assumes homogeneity within the studies that are averaged.

Ahn, Ames, and Myers (2012) reviewed 56 meta-analyses published across eight education journals between 2000 and 2010. Among other findings, they found that 28 of these meta-analyses encountered issues related to dependent effect sizes due to multiple measures of the same construct, multiple outcomes or interventions, multiple time points, or multiple comparison groups. The methods that were used to address the dependence were (1) averaging the dependent effect sizes within studies (18 studies), (2) shifting the unit of analysis (eight studies), and (3) selecting one effect size per study or combining with other methods (five studies). For the studies that

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

did not mention dependence issues, it is unclear whether dependence was not found or ignored. It is clear that the issue of dependence among effect sizes is the rule rather than the exception in meta-analyses.

Meta-Analytic Models as Multilevel Models

In this section, I briefly review how traditional meta-analyses can be written as a two-level multilevel model when the Level 1 variances are treated as fixed and known values (e.g., Goldstein, 2011; Raudenbush & Bryk, 2002; Snijders & Bosker, 2012). The two-level model is then extended to three-level meta-analyses, which are used to model the dependent effect sizes in this article.

Traditional Meta-Analyses as Two-Level Models

The model for a two-level random-effects meta-analysis is

$$\text{Level 1 model: } y_i = \lambda_i + e_i$$

and

$$\text{Level 2 model: } \lambda_i = \beta_0 + u_i, \tag{3}$$

where λ_i is the "true" effect size in the ith study and the other notations are defined in the same way as those in Equation (1). Level 1 refers to the participants, whereas Level 2 refers to the studies. A participant subscript does not exist because the raw data are usually unavailable for meta-analyses. By combining these two levels, the model becomes the random-effects meta-analytic model listed in Equation (1).

Similarly, the mixed-effects meta-analytic model in Equation (2) can be written as a two-level multilevel model:

$$\text{Level 1 model: } y_i = \lambda_i + e_i$$

and

$$\text{Level 2 model: } \lambda_i = \beta_0 + \beta_1 x_i + u_i. \tag{4}$$

Participant-level covariates are rarely available for meta-analyses except in individual patient data meta-analyses in medicine (e.g., Riley, Lambert, & Abo-Zaid, 2010). This method is also known as "integrative data analysis" in psychology (see Curran, 2009, in the special issue of *Psychological Methods*). Thus, only study-level covariates are used in mixed-effects meta-analyses. By formulating traditional meta-analytic models as multilevel models, meta-analyses are able to be conducted using multilevel modeling packages (Hox, 2010; Raudenbush & Bryk, 2002).

Three-Level Meta-Analytic Models

The two-level meta-analytic model can be extended to a three-level model by adding a cluster effect. Let y_{ij} be the ith effect size in the jth cluster. The definition of clusters depends on the data structure and research questions. For example, y_{ij} might represent one of the multiple effect sizes in the jth study (e.g., Nguyen & Benet-Martínez, 2013), or it might represent one of the studies in the jth cultural group of a cross-cultural meta-analysis (e.g., Fischer & Boer, 2011). In single-case studies, this value represents one of the measures in jth participant (Owens & Ferron, 2012). The model for a three-level random-effect meta-analysis is

$$\text{Level 1 model: } y_{ij} = \lambda_{ij} + e_{ij},$$

$$\text{Level 2 model: } \lambda_{ij} = \kappa_j + u_{(2)ij},$$

and

$$\text{Level 3 model: } \kappa_j = \beta_0 + u_{(3)j}, \tag{5}$$

where λ_{ij} is the "true" effect size and $\text{Var}(e_{ij}) = v_{ij}$ is the known sampling variance in the ith effect size in the jth cluster, κ_j is the average effect in the jth cluster, β_0 is the average population effect, and $\text{Var}(u_{(2)ij}) = \tau^2_{(2)}$ and $\text{Var}(u_{(3)j}) = \tau^2_{(3)}$ are the study-specific Level 2 and Level 3 heterogeneity, respectively.

$\tau^2_{(2)}$ might be of substantive interest depending on how Level 2 is defined. For example, if Level 2 represents the effect sizes using multiple subscales, then $\tau^2_{(2)}$ indicates the heterogeneity of the effect sizes due to the subscales. A small $\tau^2_{(2)}$ indicates that the effect sizes are similar at Level 2. Reviewers might safely conclude that the choice of subscales does not affect the magnitude of the effect sizes. A large $\tau^2_{(2)}$, however, suggests that the magnitude of the effect sizes varies across the different subscales. Investigating how the subscales predict the effect sizes is crucial. Types of subscales can be used as Level 2 covariates to model the heterogeneity. $\tau^2_{(3)}$ indicates the dissimilarity among the true effect sizes across the studies at Level 3 after controlling for the multiple subscales at Level 2. A large $\tau^2_{(3)}$ suggests that the population effect sizes vary across Level 3. Study characteristics can be included to explain the heterogeneity at Level 3.

Similar to the two-level model, the equations are often combined into a single equation:

$$y_{ij} = \beta_0 + u_{(2)ij} + u_{(3)j} + e_{ij}. \tag{6}$$

The standard assumptions of multilevel modeling are made. The random effects at different levels and the sampling error are assumed to be independent, that is, $\text{Cov}(u_{(2)ij}, u_{(3)j}) = \text{Cov}(u_{(2)ij}, e_{ij}) = \text{Cov}(u_{(3)j}, e_{ij}) = 0$. Based on the above assumptions, we have $\mathbf{E}(y_{ij}) = \beta_0$, where $\mathbf{E}(.)$ is the expected value. Thus, (1) $\text{Var}(y_{ij}) = \tau^2_{(3)} + \tau^2_{(2)} + v_{ij}$: The unconditional sampling variance of the effect size is the sum of the Level 2 and Level 3 heterogeneity as well as the known sampling variance; (2) $\text{Cov}(y_{ij}, y_{kj}) = \tau^2_{(3)}$: Effect sizes in the same cluster share the same covariance; and (3) $\text{Cov}(y_{ij}, y_{mn}) = 0$: Effect sizes in different clusters are independent.

The random-effects model can be extended to a mixed-effects model by including study characteristics as covariates. Let x be a covariate that can be either x_{ij} for a Level 2 covariate or x_j for a Level 3 covariate. x_j indicates that the value is the same for all effect sizes in the jth cluster, whereas x_{ij} indicates that the value might vary across the effect sizes in the jth cluster. So as not to lose generality, I used x_{ij} in the model. The mixed-effect model with one covariate is

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_{(2)ij} + u_{(3)j} + e_{ij}. \tag{7}$$

The conditional mean, variance, and covariances are $\mathbf{E}(y_{ij} \mid x_{ij}) = \beta_0 + \beta_1 x_{ij}$, $\text{Var}(y_{ij} \mid x_{ij}) = \tau^2_{(3)} + \tau^2_{(2)} + v_{ij}$, $\text{Cov}(y_{ij}, y_{kj} \mid x_{ij}) = \tau^2_{(3)}$, and $\text{Cov}(y_{ij}, y_{mn} \mid x_{ij}) = 0$. The interpretations of these terms are similar to those in Equation (6), except that $\tau^2_{(2)}$ and $\tau^2_{(3)}$ are now the level-two and level-three residual heterogeneity after controlling for the covariate. We may fit the three-level meta-analysis using the

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

above model-implied means and covariance matrix. I discuss how to fit this model under the SEM approach later.

Extensions to and Issues Associated With Three-Level Meta-Analyses

Because the methodological development of the three-level meta-analysis is not as comprehensive as that of traditional two-level meta-analyses, this section addresses extensions to and issues associated with three-level meta-analyses.

Testing Effect Size Homogeneity

Similar to traditional meta-analyses, reviewers might want to test whether the effect sizes are homogeneous. Suppose that m clusters exist; the null hypothesis of the homogeneity of effect sizes μ_{ij} is

$$H_0: \mu_{11} = \mu_{21} = ... = \mu_{1m} = ... = \mu_{km}. \tag{8}$$

Cluster effects do not exist when the effect sizes are the same under the null hypothesis. Thus, the conventional Q statistic proposed by Cochran (1954) can be applied directly. This statistic is defined as

$$Q = \sum_{i=1}^{n} w_i (y_i - \hat{\beta}_{\text{fixed}})^2, \tag{9}$$

where $w_i = 1/v_i$, $\hat{\beta}_{\text{fixed}} = \sum_{i=1}^{n} w_i y_i / \sum_{i=1}^{n} w_i$ is the fixed-effects estimate and n is the number of effect sizes. Under the null hypothesis, the Q statistic has an approximate chi-square distribution with $(n - 1)$ degrees of freedom (dfs).

Quantifying Effect Size Heterogeneity

Although the above Q statistic can be used to test the hypothesis of effect size homogeneity, it does not determine the magnitude of heterogeneity. A common index of the degree of heterogeneity for traditional meta-analyses is I^2 (Higgins & Thompson, 2002). The general formula is

$$I^2 = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \tilde{v}}, \tag{10}$$

where $\hat{\tau}^2$ is the estimated heterogeneity, and $\tilde{v}$ is the "typical" within-study variance in Equation (1). I^2 can be interpreted as the proportion of the total variation of the effect size due to between-study heterogeneity.

Based on the above definition, we can define the Level 2 $I^2_{(2)}$ and Level 3 $I^2_{(3)}$ using the model in Equation (6) as

$$I^2_{(2)} = \frac{\hat{\tau}^2_{(2)}}{\hat{\tau}^2_{(2)} + \hat{\tau}^2_{(3)} + \tilde{v}}$$

and

$$I^2_{(3)} = \frac{\hat{\tau}^2_{(3)}}{\hat{\tau}^2_{(2)} + \hat{\tau}^2_{(3)} + \tilde{v}}. \tag{11}$$

$I^2_{(2)}$ and $I^2_{(3)}$ can be interpreted as the proportions of the total variation of the effect size due to Level 2 and Level 3 between-study heterogeneity. For instance, suppose that Level 2 refers to types of subscales and Level 3 refers to studies; $I^2_{(2)}$ and $I^2_{(3)}$ can be interpreted as the proportions of the total variation attributed to the types subscales within and between studies, respectively. In cross-cultural meta-analyses that use cultural groups as a cluster effect, $I^2_{(2)}$ and $I^2_{(3)}$ are interpreted as the proportions of within cultural variation and between cultural variation to the total variation, respectively.

Importantly, the generic I^2 does not estimate any population quantity because it contains $\tilde{v}$, which is treated as a constant. If a different set of studies is selected, then $\tilde{v}$ likely differs. Both $\tilde{v}$ and I^2 are dependent on sample size. As the sample size increases, $\tilde{v}$ approaches 0. Therefore, I^2 is used as a descriptive rather than an inferential statistic. Although we are not testing any population parameter, the CIs of I^2 remain useful to quantify the precision of I^2. Higgins and Thompson (2002) discussed several methods of constructing CIs for I^2. Because $I^2_{(2)}$ and $I^2_{(3)}$ are functions of $\hat{\tau}^2_{(2)}$ and $\hat{\tau}^2_{(3)}$, by treating $\tilde{v}$ as a constant, the metaSEM package implements likelihood-based CIs (LBCIs) on $I^2_{(2)}$ and $I^2_{(3)}$ (see the section on Constructing Likelihood-Based Confidence Intervals below).

Conceptually, we can also define two ICCs that involve only the Level 2 and Level 3 quantities (e.g., Snijders & Bosker, 2012). These ICCs are simply one-to-one transformations of $I^2_{(2)}$ and $I^2_{(3)}$:

$$ICC_{(2)} = \frac{\hat{\tau}^2_{(2)}}{\hat{\tau}^2_{(2)} + \hat{\tau}^2_{(3)}} = \frac{I^2_{(2)}}{I^2_{(2)} + I^2_{(3)}}$$

and

$$ICC_{(3)} = \frac{\hat{\tau}^2_{(3)}}{\hat{\tau}^2_{(2)} + \hat{\tau}^2_{(3)}} = \frac{I^2_{(3)}}{I^2_{(2)} + I^2_{(3)}}. \tag{12}$$

The above indices can be interpreted as the proportions of the total between-study heterogeneity of the effect size due to the Level 2 and Level 3 units. One advantage of ICCs over I^2 is that their values do not depend on sample size. Therefore, they estimate the population quantities $\tau^2_{(2)}/(\tau^2_{(2)} + \tau^2_{(3)})$ and $\tau^2_{(3)}/(\tau^2_{(2)} + \tau^2_{(3)})$. Moreover, $I^2_{(2)}$ and $I^2_{(3)}$ approach $ICC_{(2)}$ and $ICC_{(3)}$, respectively, as $\tilde{v}$ decreases.

Because v_i varies across studies, no single *correct* "typical" within-study variance exists. Several definitions of "typical" within-study variance have been proposed. Takkouche, Cadarso-Suárez, and Spiegelman (1999) suggested using the harmonic mean of v_i,

$$\tilde{v}_{\text{HM}} = \frac{n}{\sum_{i=1}^{n} 1/v_i}, \tag{13}$$

whereas Higgins and Thompson (2002) proposed defining "typical" within-study variance using the Q statistic via

$$\tilde{v}_{\text{Q}} = \frac{(n-1)\sum_{i=1}^{n} 1/v_i}{\left(\sum_{i=1}^{n} 1/v_i\right)^2 - \sum_{i=1}^{n} 1/v_i^2}. \tag{14}$$

In addition to these estimators, Xiong, Miller, and Morris (2010) discussed the use of the arithmetic mean of v_i:

$$\tilde{v}_{\text{AM}} = \sum_{i=1}^{n} v_i/n. \tag{15}$$

Of the above estimators, Higgins and Thompson's (2002) statistic is the most popular definition among meta-analyses. One reason to prefer this estimator is that I^2 can be simplified to $I^2_{\text{Q}} = 1 - (k - 1)/Q$ in two-level meta-analyses. However, this relationship does not

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

apply to $I^2_{(2)}$ or $I^2_{(3)}$. The default estimator in the metaSEM package is based on Equation (14), although users can switch to other estimators.

Explained Variance in Mixed-Effects Models

When a mixed-effects meta-analysis is conducted, it is of interest to quantify the percentage of variation explained by the predictors. Raudenbush (2009) defined R^2 in two-level meta-analyses using the model in Equation (2) as

$$R^2 = 1 - \frac{\hat{\tau}^2_1}{\hat{\tau}^2_0}, \qquad (16)$$

where $\hat{\tau}^2_1$ and $\hat{\tau}^2_0$ are the estimated heterogeneity with and without predictors, respectively. The above logic can be extended to three-level meta-analyses in Equation (7). We can define the Level 2 $R^2_{(2)}$ and the Level 3 $R^2_{(3)}$ as

$$R^2_{(2)} = 1 - \frac{\hat{\tau}^2_{(2)1}}{\hat{\tau}^2_{(2)0}}$$

and

$$R^2_{(3)} = 1 - \frac{\hat{\tau}^2_{(3)1}}{\hat{\tau}^2_{(3)0}}. \qquad (17)$$

$R^2_{(2)}$ and $R^2_{(3)}$ can be interpreted as the percentages of estimated heterogeneity at Level 2 and Level 3 that are explained by the predictors. However, after adding the predictors into Equations (16) and (17), $\hat{\tau}^2_1$ can sometimes be smaller than $\hat{\tau}^2_0$. Because R^2, $R^2_{(2)}$, and $R^2_{(3)}$ are non-negative by definition, negative values are truncated to zero.

A Structural Equation Modeling Approach to Multilevel Meta-Analytic Models

Many multivariate statistics such as factor analyses, path analyses, growth models, multilevel models, item response theory and mixture models have been integrated into the general SEM framework (Muthén & Muthén, 2012; Skrondal & Rabe-Hesketh, 2004). A unified model helps methodological advances by integrating various techniques into a single framework. For example, the use of FIML to handling missing data is well developed in SEM (e.g., Enders, 2010). These techniques are directly available to researchers who want to handle missing data in path analyses, item response theory and even multilevel models using an SEM approach. Similarly, FIML is also available for meta-analyses by conceptualizing these models as structural equation models (M. W.-L. Cheung, 2008, in press). In this section, I review how two-level meta-analyses can be formulated as structural equation models. This logic is then extended to three-level meta-analytic models. The advantages of the SEM approach are addressed in the next section.

The SEM approach presented here relies on the use of definition variables (Mehta & Neale, 2005) and phantom variables (Rindskopf, 1984). Definition variables are often used to fix participant-specific values to any model parameter. For example, consider a latent growth model. A confirmatory factor analysis is often used to represent a latent growth model using the factor loadings as the observed time points (e.g., 0, 1, 2, and 3). When observation times vary across participants (e.g., 0, 1, 2, and 3 for one participant, and 0, 1, 5, and 7 for another participant), definition variables can be used to fix participant-specific factor loadings. Different participants may have different sets of factor loadings.

Phantom variables are another building block of this SEM approach. Phantom variables can be used to impose equality or inequality constraints on the parameters. They can also be used to "store" new parameters. Let us consider a simple regression example in which Y is regressed on X and b is the regression coefficient. If we want to ensure that b is non-negative, then we might introduce a phantom variable P, a latent variable with a variance of 0, in the model so that $X \rightarrow P \rightarrow Y$ (with d as the regression coefficient for both paths). The effect from X to Y becomes d^2, which is always non-negative. Readers may refer to M. W.-L. Cheung (2008, 2010) for more details concerning how to apply definition and phantom variables in meta-analysis.

Two-Level Meta-Analyses

To formulate a two-level meta-analytic model as a structural equation model, we must specify the model-implied means $\mu_i(\theta)$ and the model-implied covariance matrix $\Sigma_i(\theta)$ for the ith study. These model-implied moments are the hypothesized values when the proposed model is correct. The model-implied means and covariance matrix for the random-effects meta-analysis listed in Equation (1) are

$$\mu_i(\theta) = \beta_0$$

and

$$\Sigma_i(\theta) = \tau^2 + v_i. \qquad (18)$$

Figure 1 shows a structural equation of the random-effects meta-analysis. Rectangles, circles, and triangles are used to represent observed variables, latent variables, and a constant vector of ones, respectively. Definition variables are used to impose v_i on y_i, so that v_i is able to vary across studies. Suppose that 10 studies are included in a meta-analysis; one variable exists with 10 participants. This model presentation is similar to the "long format" of handling latent growth model data using a multilevel modeling approach.[1]

Alternatively, studies can be treated as variables (see Mehta & Neale, 2005). Figure 2, which is equivalent to the model shown in Figure 1, displays the model using two studies. The effect sizes of these studies are represented by variables y_1 and y_2. This model representation has three key features. First, all expected means labeled β_0 are the same. Next, the studies are independent. Therefore, the model-implied covariance between y_1 and y_2 is zero. Third, one row is used to represent the entire data set because the

[1] For example, suppose two participants have variable X measured at $T1$ and $T2$. The so-called "wide format" treats X measured at $T1$ and $T2$ as two variables (X_{T1} and X_{T2}). The data frame is $\begin{bmatrix} Subject & X_{T1} & X_{T2} \\ 1 & X_{11} & X_{12} \\ 2 & X_{21} & X_{22} \end{bmatrix}$. When missing values exist in X, they are labeled (e.g., "NA") in the data frame. The "long format" treats X as a single variable by stacking the measurements together. The data frame is $\begin{bmatrix} Subject & X & Time \\ 1 & X_{11} & 1 \\ 1 & X_{12} & 2 \\ 2 & X_{21} & 1 \\ 2 & X_{22} & 2 \end{bmatrix}$. When missing values exist in X, the rows with missing values are excluded from the data frame.

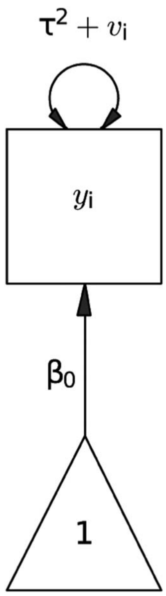

Figure 1. A structural equation model with long format for random-effects meta-analysis.

number of studies is represented by the variables. If 10 studies are included in a meta-analysis, then 10 variables will represent the data. This model representation is similar to the “wide format” of latent growth model data using an SEM approach (see the Footnote 1). By using variables as studies, we are able to convert a two-level model into a single-level model. I apply this technique to model three-level meta-analyses below.

Three-Level Meta-Analyses

To model three-level effect sizes in SEM, we must treat each cluster as a participant and the effect sizes within clusters as variables (e.g., Mehta & Neale, 2005; Preacher, 2011). The basic idea is to treat the dependent effect sizes within a cluster as multiple variables. This technique ensures that the clusters are independent: SEM can be used to model the data.

The following steps outline how to conduct the analysis. First, we must create as many variables as the maximum number of studies per cluster. This step ensures that we have sufficient variables to represent the dependent effect sizes. For example, if the maximum number of effect sizes per cluster is 5, then we must create 5 variables. For clusters with fewer than 5 effect sizes, the incomplete effect sizes are treated as missing values and handled using FIML (Mehta & Neale, 2005).

Assume that the maximum number of effect sizes per cluster is k; the model-implied means for the jth cluster is a $k \times 1$ vector of β_0, that is,

$$\boldsymbol{\mu}_j(\boldsymbol{\theta}) = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \beta_0. \tag{19}$$

The model-implied covariance matrix for the jth cluster is a $k \times k$ matrix of the sum of three matrices: a $k \times k$ matrix $\mathrm{T}^2_{(3)}$ with all the elements of $\tau^2_{(3)}$ that represent the Level 3 heterogeneity, a $k \times k$ diagonal matrix $\mathrm{T}^2_{(2)}$ with the elements of $\tau^2_{(2)}$ that represent the Level 2 heterogeneity, and a $k \times k$ diagonal matrix $\mathbf{V}$ with the elements of v_{ij} (Konstantopoulos, 2011), that is,

$$\begin{aligned} \Sigma_j(\boldsymbol{\theta}) = {} & \mathrm{T}^2_{(3)} + \mathrm{T}^2_{(2)} + \mathbf{V} \\ = {} & \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 1 & \cdots & 1 & 1 \end{bmatrix} \tau^2_{(3)} + \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{bmatrix} \tau^2_{(2)} \\ & + \begin{bmatrix} v_{1j} & 0 & \cdots & 0 \\ 0 & v_{2j} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & v_{11j} \end{bmatrix}. \end{aligned} \tag{20}$$

For instance, if 10 clusters have a maximum of 5 effect sizes per cluster, then the structural equation model includes 5 variables with 10 participants. Figure 3 shows a graphical model of a three-level meta-analytic model with 2 effect sizes per cluster. As this figure illustrates, the covariance matrix is a structured matrix with the elements listed in Equation (20).

When there is a high degree of heterogeneity at Levels 2 or 3, reviewers might want to explain this heterogeneity by conducting a mixed-effects meta-analysis. Suppose that one covariate x_{ij} exists; the conditional model-implied mean vector for the jth cluster is

$$\boldsymbol{\mu}_j(\boldsymbol{\theta} \mid x_{ij}) = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \beta_0 + \begin{bmatrix} x_{1j} \\ \vdots \\ x_{kj} \end{bmatrix} \beta_1, \tag{21}$$

whereas the conditional model-implied covariance matrix $\Sigma_j(\boldsymbol{\theta} \mid x_{ij})$ is the same as that of the random-effects model in Equation (20).

There are two approaches to implement covariates in a meta-analysis. The first approach, which is commonly used in SEM, is to treat the covariates as random variables (e.g., M. W.-L. Cheung, 2008). By doing so, we must estimate their means and covariance matrix. The second approach, which is usually used in regression and meta-analyses, is to treat the covariates as a design matrix. A design matrix, $[x_{1j} \ \cdots \ x_{kj}]^{\mathrm{T}}$ in Equation (21), represents the known values of the covariate. The values of the design matrix are treated as fixed. No distribution assumptions are made on the

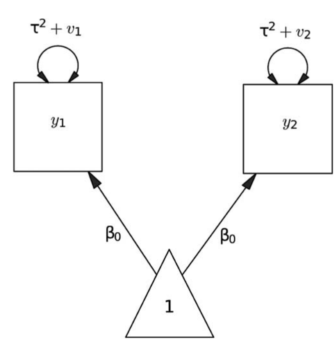

Figure 2. A structural equation model with wide format for random-effects meta-analysis with two studies.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

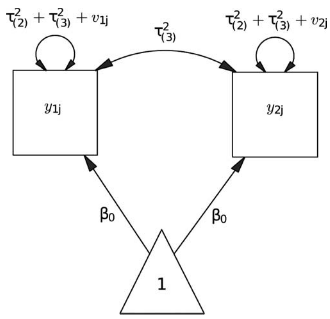

Figure 3. A structural equation model on three-level random-effects meta-analysis with two studies in the *j*th cluster.

values of the design matrix. This article uses the latter approach because it is more consistent with traditional meta-analyses.

Figure 4 shows a model with covariate x_{ij} for the *j*th cluster. The phantom variable P_1 (with zero variance) is created to store the values of the covariate. The specific values of x_{ij} are imposed via definition variables, whereas β_1 is the regression coefficient. Let us consider $\begin{bmatrix} y_{1j} \\ y_{2j} \end{bmatrix}$ in the figure as an example. By using the tracing rules, the model-implied mean is $\begin{bmatrix} 1 \\ 1 \end{bmatrix}\beta_0 + \begin{bmatrix} x_{1j} \\ x_{2j} \end{bmatrix}\beta_1$, which follows Equation (21).

Restricted (or Residual) Maximum Likelihood Estimation

ML estimation method is usually used as an estimation method in the aforementioned SEM procedures. The fixed effects and the variance components of the random effects are estimated simultaneously. The variance components of the random effects are slightly under-estimated because the uncertainty in the fixed effects estimation has not been taken into account (e.g., Harville, 1977; Patterson & Thompson, 1971). An REML estimation method can minimize this bias. REML estimation removes the fixed-effects parameters before estimating the variance components. A contrast matrix is chosen such that the fixed-effects parameters are not involved in the model. Thus, the estimated variance components are not biased by treating the fixed-effects estimates as known.

The REML model for two-level meta-analyses (M. W.-L. Cheung, 2013) can be extended to three-level meta-analyses by constructing a contrast to remove the fixed-effects parameters ($\boldsymbol{\beta}$ in Equation [21]). Let us stack all y_{ij} into a single column vector $\mathbf{y}$. The three-level meta-analytic model is

$$\mathbf{y} = X\boldsymbol{\beta} + Z(\mathbf{u}_{(2)} + \mathbf{u}_{(3)}) + \mathbf{e}, \tag{22}$$

where X, $\mathbf{u}_{(2)}$, $\mathbf{u}_{(3)}$ and $\mathbf{e}$ are stacked over all the units, and $Z = \text{Diag}(Z_1, Z_2, \ldots, Z_k)$ is a diagonal selection matrix that selects the random effects. The structure of $\text{Cov}(Z(\mathbf{u}_{(2)} + \mathbf{u}_{(3)}) + \mathbf{e}, Z(\mathbf{u}_{(2)} + \mathbf{u}_{(3)}) + \mathbf{e})$ is similar to that represented by Equation (20). Instead of analyzing $\mathbf{y}$, we can analyze its residuals $\tilde{\mathbf{y}}$:

$$\tilde{\mathbf{y}} = A\mathbf{y} = AX\boldsymbol{\beta} + A(Z(\mathbf{u}_{(2)} + \mathbf{u}_{(3)}) + \mathbf{e}), \tag{23}$$

where $A = I - X(X^TX)^{-1}X^T$ with p arbitrary rows is removed, I is an identity matrix and p is the number of columns of X (Harville, 1977; Patterson & Thompson, 1971). Because the rank of A is not full, p redundant rows are linearly dependent on the other rows. p rows can be arbitrarily selected without affecting the results (Harville, 1977). The common practice is to delete the last p rows in A. Importantly, the mean of $\tilde{\mathbf{y}}$ is always zero because these data are residuals. Thus, the fixed effects ($\boldsymbol{\beta}$ in Equation [22]) are not estimable. After estimating the variance component, we can compute the fixed effects by

$$\hat{\boldsymbol{\beta}} = (X^{\mathrm{T}}V^{-1}X)^{-1}X^{\mathrm{T}}V^{-1}\mathbf{y}. \tag{24}$$

In the context of SEM, we can fit $\boldsymbol{\mu}(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ by treating $\tau^2_{(3)}$ and $\tau^2_{(2)}$ as fixed values obtained from the REML estimation (see M. W.-L. Cheung, 2013). The estimates of β under these constraints are the same as those in Equation (24). Three-level meta-analyses, with either ML or REML estimation methods, are implemented in the metaSEM package.

Benefits of Using Structural Equation Modeling to Conduct Three-Level Meta-Analyses

Three-level meta-analyses can be conducted in multilevel modeling packages such as HLM (Raudenbush & Bryk, 2002) and SAS (SAS Institute, 2003). Readers might wonder why an SEM approach is preferable. As discussed in M. W.-L. Cheung (2008), many state-of-the-art techniques are available to reviewers who conduct a meta-analysis using an SEM perspective. This section highlights the benefits of conceptualizing three-level meta-analyses as an SEM.

Comparisons Between Two- and Three-Level Models With Constraints

Many research hypotheses can be formulated as comparisons of nested models. By comparing these nested models, reviewers can test whether the imposed constraints are appropriate. SEM provides a flexible framework for comparing models with and without constraints.

Testing $H_0: \tau^2_{(3)} = 0$. By comparing the two-level meta-analysis in Figure 2 with the three-level meta-analysis in Figure 3,

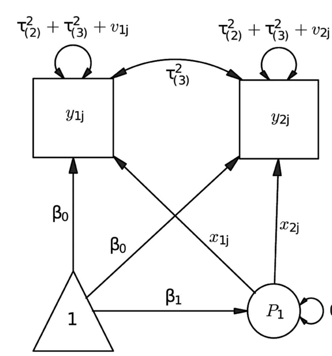

Figure 4. A structural equation model on three-level mixed-effects meta-analysis with two studies and one predictor in the *j*th cluster.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

it is clear that the two-level meta-analysis is a special case of the three-level meta-analysis after fixing $\tau^2_{(3)} = 0$. Because these models are nested, we can test the null hypothesis $H_0 : \tau^2_{(3)} = 0$ by using a likelihood ratio or chi-square difference test. If this test statistic is non-significant, then the three-level model is not statistically better than the two-level model.

However, reviewers should be cautious of two items. First, the test statistic is not distributed with a chi-square distribution with 1 *df* even when the null hypothesis is true. Because $H_0 : \tau^2_{(3)} = 0$ is tested on the boundary, $\tau^2_{(3)}$ cannot be negative. The test statistic is distributed as a 50:50 mixture of a degenerate random variable with all of its probability mass concentrated at zero and a chi-square random variable with 1 *df* (Self & Liang, 1987; Stoel, Garre, Dolan, & van den Wittenboer, 2006). One simple strategy to correct for this bias is to use 2α instead of α as the alpha level. That is, we reject the null hypothesis at $\alpha = .05$ when the observed *p* value is larger than .10 (Pinheiro & Bates, 2000).

Second, it is inadvisable to decide between two- and three-level models by testing $H_0 : \tau^2_{(3)} = 0$. Similar to the choice between a fixed- and a random-effects model in a traditional meta-analysis, this decision should be based on whether a conditional or unconditional inference is required (Hedges & Vevea, 1998). If the reviewers want to generalize their findings to Levels 2 and 3, then a three-level meta-analysis should be used regardless of whether or not $H_0 : \tau^2_{(3)} = 0$ is rejected.

Testing $H_0 : \tau^2_{(3)} = \tau^2_{(2)}$. Frequently, reviewers might want to compare the amounts of heterogeneity at Levels 2 and 3. For example, in the case of cross-cultural research, most researchers are primarily interested in between-cultural variation. However, some researchers argue that intra-cultural variation is also meaningful (e.g., Au, 1999; Au & Cheung, 2004). If a cross-cultural meta-analysis is conducted by treating studies as Level 2 and cultural groups as Level 3, then reviewers might want to test $H_0 : \tau^2_{(3)} = \tau^2_{(2)}$. Under this null hypothesis, the difference of the test statistic between the models with and without the constraint has a chi-square distribution with 1 *df*. Three-level meta-analyses provide a statistical approach to testing the presence of intra-cultural variation. Importantly, the above research hypothesis cannot be tested using common methods to handle dependent effect sizes such as averaging the dependent effect sizes because these methods cannot disentangle Level 2 and Level 3 heterogeneity.

Flexible Constraints Using the SEM Approach

Occasionally, reviewers might want to test whether certain regression coefficients are equal. For example, clinical psychologists might want to compare the treatment effectiveness of cognitive behavioral therapy (CBT), medication (Med) and a waiting list (WL). Depending on their theories, the reviewers might only want to compare CBT with Med. Using a SEM approach, reviewers can compare two models: $H_0{:}y_{ij} = \beta_1 CBT_{ij} + \beta_1 Med_{ij} + \beta_3 WL_{ij} + u_{(2)ij} + u_{(3)j} + e_{ij}$ versus $H_1{:}y_{ij} = \beta_1 CBT_{ij} + \beta_2 Med_{ij} + \beta_3 WL_{ij} + u_{(2)ij} + u_{(3)j} + e_{ij}$, where the covariates are indicator coded, and an intercept does not exist.[2] Under the null hypothesis $H_0{:}\beta_1 = \beta_2$, the difference between the test statistics of these two models follows a chi-square distribution with 1 *df*. Although the above covariates are Level 2 variables, the same applies to Level 3 covariates or a mix of Level 2 and Level 3 covariates.

Constructing Likelihood-Based Confidence Intervals

It is always recommended to report CIs to assist the interpretation of parameter estimates and effect sizes (American Psychological Association, 2010; Wilkinson & Task Force on Statistical Inference, 1999). One common method of constructing CIs is based on the Wald test (or *SE*). Thus, these statistics are called the Wald CIs. The $100(1 - \alpha)\%$ Wald CI for parameter θ can be computed by $\hat{\theta} \pm z_{1-\alpha/2} SE(\hat{\theta})$, where $SE(\hat{\theta})$ is the estimated *SE* for θ, and $z_{1-\alpha/2}$ is the $100(1 - \alpha/2)$th percentile of the standard normal distribution. Because *SE* is routinely available in data analyses, Wald CIs are widely reported.

However, several criticisms have been made with regard to the Wald CIs. Specifically, Wald CIs are always symmetrical around $\hat{\theta}$; however, the sampling distributions of the estimates can be asymmetrical. For example, the sampling distributions of the variance parameters, correlation coefficients that are close to ± 1, and the product term of two random variables with regard to indirect effects are usually asymmetrical. The accuracy of Wald CIs is questionable in these settings (e.g., MacKinnon, 2008; Searle, Casella, & McCulloch, 1992). A related issue of symmetrical CIs is that they can fall outside of meaningful boundaries; for example, a negative lower limit for variance or CIs beyond ± 1 for correlation coefficients. The standard practice is to truncate these CIs to their meaningful boundaries (e.g., zero for variance and ± 1 for correlation). However, Steiger and Fouladi (1997) warned that the width of these CIs might be suspicious as an indicator for the precision of the measurement because of the truncation.

LBCI is an alternative to Wald CIs (Casella & Berger, 2002). Suppose that we want to construct a $100(1 - \alpha)\%$ LBCI ($\hat{\theta}_{\text{Lower}}$, $\hat{\theta}_{\text{Upper}}$) on θ; we move the parameter estimate (treated as variable) as far away as possible to the right for $\hat{\theta}_{\text{Upper}}$ (or to the left for $\hat{\theta}_{\text{Lower}}$) from its maximum likelihood estimate (MLE), such that it is just significant at the desired α level (for more details, see M. W.-L. Cheung, 2009; Neale & Miller, 1997). LBCIs are asymmetrical and able to capture the distributions of the parameter estimates. Moreover, they always fall inside meaningful boundaries because the parameter estimates are meaningful. Although LBCIs are not guaranteed to be optimum, they will seldom be too bad (Casella & Berger, 2002). Despite this, Casella and Berger (2002) nevertheless "recommend constructing a confidence set based on inverting an LRT [likelihood ratio test], if possible" (p. 430). LBCIs have been used to quantify the precision of the heterogeneity variances in meta-analyses (Hardy & Thompson, 1996; Viechtbauer, 2005). Because the sampling distributions of $\hat{\tau}^2_{(2)}$, $\hat{\tau}^2_{(3)}$, $I^2_{(2)}$, $I^2_{(3)}$, $ICC_{(2)}$, and $ICC_{(3)}$ are asymmetrical (i.e., they are based on variances or variance functions), LBCIs are preferred to Wald CIs.

Handling Missing Covariates With Full-Information Maximum Likelihood Estimation

Cooper and Hedges (2009) argued that missing data are the most pervasive practical problem in the synthesis of research. Publica-

[2] The covariates are coded as $CBT_{ij} = 1$, $Med_{ij} = 0$, and $WL_{ij} = 0$ for the CBT condition; $CBT_{ij} = 0$, $Med_{ij} = 1$, and $WL_{ij} = 0$ for the Med condition; and $CBT_{ij} = 0$, $Med_{ij} = 0$, and $WL_{ij} = 1$ for the WL condition.

tion bias refers to when missing values on the effect sizes are related to whether the studies have been published, which is beyond the scope of this article (see Rothstein, Sutton, & Borenstein, 2005). However, another type of missing data is missing covariates in mixed-effects meta-analyses. Cooper and Hedges (2009) commented that "the prevalence of missing data on moderator . . . influences the degree to which the problems investigated by the synthesis can be formulated" (p. 565).

Three types of missingness mechanisms exist. Data are missing completely at random (MCAR) when the probability of their missingness on *Y* is unrelated to the value of *Y* itself or the values of any other variable in the data set. Data on *Y* are missing at random (MAR) when the probability of their missingness on *Y* is unrelated to the value of *Y* after controlling for other variables in the analysis. When the missingness is neither MCAR nor MAR, the data are missing not at random (MNAR).

Several approaches have been proposed to handle missing data (see Enders, 2010, for a general review, and Pigott, 2012, for a review in the context of meta-analyses). These include complete-case analyses, single value imputations, EM algorithms, multiple imputation (MI) and ML, which is also widely known as FIML in the SEM literature. MI and FIML are generally recommended as the best approaches to handle MCAR or MAR data (Enders, 2010; Schafer & Graham, 2002).

Compared with MI, FIML has at least three advantages. First, the results based on MI are asymptotically equivalent to those based on FIML. That is, they are equal when the number of imputations in MI approaches infinity. In general, 3–5 imputations are considered sufficient to obtain excellent results using MI (e.g., Schafer & Olsen, 1998). However, Graham, Olchowski, and Gilreath (2007) showed that many more imputations are required. They suggested requiring *m* imputations based on the fraction of missing information (γ). They recommended that researchers use $m = 20$, 20, 40, 100, and >100 imputations for data with $\gamma = 0.1$, 0.3, 0.5, 0.7, and 0.9. Based on their simulations results, they concluded that FIML is superior to MI in terms of power for testing small effect sizes (unless one has a sufficient number of imputations).

Second, conducting MI for multilevel data is challenging (cf. van Buuren & Groothuis-Oudshoorn, 2011). More importantly, how to apply MI on three-level meta-analysis data remains unclear. Lastly, FIML is more robust than MI when the data are non-normal (Yuan, Yang-Wallentin, & Bentler, 2012). Specifically, the performance of FIML is less biased and more efficient than that of MI. By formulating three-level meta-analyses as single-level structural equation models, it is straightforward to apply FIML to handling missing covariates. Moreover, auxiliary variables at Level 2 and Level 3 are allowed. Auxiliary variables are those that predict the missing values or are correlated with the variables that contain missing values. The inclusion of auxiliary variables can improve the efficiency of the estimation and the parameter estimates (Enders, 2010).

Illustrations Using Two Published Data Sets

This section demonstrates how to conduct a three-level meta-analysis using the metaSEM package, that was based on the OpenMx package (Boker et al., 2011) for the R statistical environment (R Development Core Team, 2013). The metaSEM package was written to provide a general framework for conducting meta-analyses using an SEM approach. There are several functions to conduct meta-analyses. meta() implements the univariate and multivariate meta-analyses (M. W.-L. Cheung, 2008, in press), whereas reml() implements the same analysis using the REML estimation method (M. W.-L. Cheung, 2013). tssem1() and tssem2() conduct meta-analytic structural equation modeling using the two-stage SEM approach (M. W.-L. Cheung & Chan, 2005, 2009). meta3(), meta3X(), and reml3() implement the three-level meta-analysis discussed in this article. The metaSEM package also provides functions to calculate the *Q* statistics, I^2, R^2 indices, LBCIs, and constraints on parameter estimates. The R code for the analyses is listed in the Appendix. The explanations of the R code and the output are available at the author's website (http://courses.nus.edu.sg/course/psycwlm/internet/metaSEM/3level.html). This website also includes detailed steps on how to implement the three-level meta-analyses as structural equation models.

This section provides two examples. The first example (Cooper, Valentine, Charlton, & Melson, 2003; Konstantopoulos, 2011) compares the results between two- and three-level meta-analyses. The second example (Bornmann, Mutz, & Daniel, 2007; Marsh et al., 2009) is used to demonstrate more advanced three-level meta-analysis techniques.

Three-Level Meta-Analysis With Cooper et al.'s (2003) Data Set

Data description. Cooper et al. (2003) conducted a meta-analysis on the effectiveness of modified school calendars with regard to student achievement as well as school and community attitudes. Using a modified calendar, children can be placed in groups with alternating vacation sequences. This arrangement increases the number of students that a particular school can accommodate. Because modified calendars are more popular in school districts in which great needs for additional schools and classrooms exist, the effect sizes are likely to be more similar for schools in the same district than those between districts. Thus, Cooper et al. considered schools to be nested within districts. To address the dependence among the effect sizes, Cooper et al. used district as the unit of analysis. The effect sizes were averaged within districts before estimating the average effect.

Konstantopoulos (2011, Table 1) illustrated a three-level meta-analytic procedure with selected cases from Cooper et al. (2003). This data set included 56 effect sizes nested within 11 districts. The effect size employed was the *d* index (standardized mean difference). Positive values of *d* denoted modified school calendar effectiveness. Year of publication was included as a covariate. This data set was stored as Cooper03 in the metaSEM package. The effect size and its sampling variance, year of publication, and district are labeled as y, v, Year, and District in the data set.

Results and discussion. In this section, I present the results of a traditional (two-level) meta-analysis. The results of a three-level meta-analysis and the comparisons between these analyses are then presented.

The *Q* statistic was $\chi^2\,(df = 55) = 578.87$, $p < .001$, which indicates that the null hypothesis of effect size homogeneity was rejected. Thus, a fixed-effects model that assumes effect size homogeneity cannot sufficiently describe the data variability. The average effect size (and its 95% Wald CIs) based on a two-level

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

random-effects model was 0.1280 (0.0428, 0.2132). The estimated amount of heterogeneity $\hat{\tau}^2$ and the I^2 based on the Q statistic were .0865 and .9459, respectively. These results indicate that the study level explained 95% of the total variation using a two-level model. When year of publication was included as a covariate, the estimated coefficient (and its 95% Wald CIs) was 0.0051 (−0.0033, 0.0136) with $R^2 = .0164$. Because the CIs included 0, this result was not significant at $\alpha = .05$, which suggests that the year of publication does not statistically explain variation in the effect size.

The average effect size (and its 95% Wald CIs) for a three-level meta-analysis was 0.1845 (0.0266, 0.3423). The estimated amount of heterogeneity at Level 2 $\hat{\tau}^2_{(2)}$ and Level 3 $\hat{\tau}^2_{(3)}$ were 0.0329 and 0.0578, respectively. The Level 2 $I^2_{(2)}$ and the Level 3 $I^2_{(3)}$ based on the Q statistic were .3440 and .6043, respectively. These values indicate that schools (Level 2) and districts (Level 3) explained 34% and 60% of the total variation, respectively. The remaining 6% of the total variation was due to known sampling error. It appears that Level 3 (the district effect) explains much more of the variance than Level 2 (the school effect). I tested the null hypothesis $H_0 : \tau^2_{(3)} = \tau^2_{(2)}$. The chi-square difference test statistic was $\chi^2(df = 1) = 0.6871$, $p = .4072$, which indicates that we cannot reject the null hypothesis of equal variances. In other words, the amount of heterogeneity is not significantly different at both levels.

The two-level model with the $\tau^2_{(3)} = 0$ constraint was fit and tested against the three-level model. The test statistic was significant ($\chi^2(df = 1) = 16.50$, $p < .001$) after adjusting for the testing on the boundary condition. Therefore, the three-level model provided a better fit than the two-level model. If a two-level meta-analysis is used, then the model is misspecified. The parameter estimates and their associated *SE*s are likely incorrect. When year of publication was included as a covariate, the estimated coefficient (and its 95% Wald CIs) was 0.0051 (−0.0116, 0.0218) with $R^2_{(2)} = .0000$ and $R^2_{(3)} = .0221$. Because the CIs included 0, this result was not significant at $\alpha = .05$.

Having presented both the traditional (two-level) and three-level meta-analyses, I highlight the major differences between them here. Table 1 lists the results between the two- and the three-level models. Several observations can be made. First, the width of the CIs (or the *SE*s) of the fixed-effect parameters (the intercept and regression coefficient) was usually smaller in the two-level model. This finding is typical when a higher level is incorrectly ignored by a multilevel analysis (e.g., Snijders & Bosker, 2012).

Second, τ^2 and I^2 in the two-level meta-analysis were overestimated. These values were approximately equal to the sum of $\tau^2_{(2)}$ and $\tau^2_{(3)}$ in the three-level meta-analysis. When Level 3 was ignored, the between-study heterogeneity was incorrectly attributed to Level 2. If a two-level meta-analysis was conducted and reported, then the reviewers might draw the incorrect conclusion that a large amount of heterogeneity ($I^2 = .95$) exists at the school level. In fact, the school level can only explain 34% of the total variation, whereas the other 60% of the total variation is attributed by the district level.

Third, the Wald CIs on the variance components can have non-meaningful bounds. For example, the 95% lower bounds of the Wald CIs on $\tau^2_{(3)}$ in the three-level models were negative. When LBCIs were used, the lower bounds on $\tau^2_{(3)}$ stay within a meaningful range. LBCIs are preferred to quantify precise variances.

Table 1
Results of the Two- and Three-Level Meta-Analyses and Their 95% Wald Confidence Intervals (CIs)

Model type	Two-level meta-analysis	Three-level meta-analysis
Random-effects model		
β_0	0.1280 (0.0428, 0.2132)	0.1845 (0.0266, 0.3423)
$\tau^2_{(2)}$	0.0865 (0.0483, 0.1247)	0.0329 (0.0110, 0.0547)
$\tau^2_{(3)}$	0[a]	0.0577 (−0.0025, 0.1180)[b]
$I^2_{(2)}$	.9460 (.9194, .9651)	.3440 (.1274, .6573)
$I^2_{(3)}$	0[a]	.6043 (.2794, .8454)
$ICC^2_{(2)}$	Not defined	.36273 (.1310, .7015)
$ICC^2_{(3)}$	Not defined	.63727 (.2985, .8690)
Mixed-effects model		
β_0	0.1259 (0.0412, 0.2106)	0.1780 (0.0202, 0.3358)
β_1	0.0051 (−0.0033, 0.0136)	0.0051 (−0.0116, 0.0218)
$\tau^2_{(2)}$	0.0851 (0.0478, 0.1224)	0.0329 (0.0111, 0.0548)
$\tau^2_{(3)}$	0[a]	0.0565 (−0.0024, 0.1153)[c]
$R^2_{(2)}$	.0164	.0000
$R^2_{(3)}$	0[a]	.0221

[a] Fixed at 0 in the two-level meta-analysis. [b] 95% likelihood-based CIs (LBCIs): 0.0198–0.1763. [c] 95% LBCIs: 0.0192–0.1720.

More Complex Three-Level Meta-Analysis With Bornmann et al.'s (2007) Data Set

Data description. The second data set was based on Bornmann et al. (2007) and Marsh et al. (2009). Bornmann et al. extracted 66 effect sizes from 21 studies of gender differences in the peer reviews of grant and fellowship applications. The effect size was an odds ratio; specifically, the odds of being approved among the female applicants divided by the odds of being approved among the male applicants. To improve the normality assumption with regard to the effect sizes, a logarithm was applied to the odds ratio. Bornmann et al. and Marsh et al. used a three-level meta-analysis to handle dependent effect sizes. Level 1 was the authors, whereas Levels 2 and 3 were the proposals and studies, respectively.

This data set was stored as Bornmann07 in the metaSEM package. The effect size and its sampling variance are logOR and v, respectively, whereas Cluster is the cluster effect at Level 3. The covariates are Year (Year of publication), Type (Fellowship vs. Grants), Discipline (Physical sciences, Life sciences/biology, Social sciences/humanities, or Multidisciplinary), and Country (United States, Canada, Australia, United Kingdom, or Europe).

Results and discussion. I first present the results for the three-level meta-analysis. More advanced data analyses, such as those that include covariates and missing covariates, are addressed next.

The Q statistic was 221.2809 ($df = 65$), $p < .001$. This value rejected the null hypothesis that all effect sizes were homogeneous. The estimated average effect (and its 95% Wald CIs) under a three-level model was −0.1008 (−0.1794, −0.0221). The test statistic on $H_0{:}\tau^2_{(3)} = 0$ was $\chi^2(df = 1) = 10.2202$, $p < .01$, after adjusting for the testing on the boundary condition. This result

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

shows that the three-level model was a better fit than the two-level model.

$\hat{\tau}^2_{(2)}$ and $\hat{\tau}^2_{(3)}$ were 0.0038 and 0.0141, respectively. $I^2_{(2)}$ and $I^2_{(3)}$, which are based on the Q statistics, were .1568 and .5839, respectively. These values suggest that Level 3 (studies) had a higher degree of heterogeneity than Level 2 (proposals). The test statistic on H_0:$\tau^2_{(2)} = \tau^2_{(3)}$ was $\chi^2(df = 1) = 1.3591$, $p = .2437$, however, which was not significant. Thus, the amount of Level 2 (proposals) and Level 3 (studies) heterogeneity did not significantly differ.

Type indicates whether the proposals were grants or fellowships. A dummy variable for Type was created where 0 denoted grants, and 1 represented fellowships. The estimated intercept and slope (and their 95% Wald CIs) were −0.0066 (−0.0793, 0.0661) and −0.1956 (−0.3017, −0.0894), respectively, with $R^2_{(2)}$ = .0693 and $R^2_{(3)}$ = .7943. Thus, the estimated effect size for grants was −0.0066, whereas the difference between grants and fellowships was −0.1956, which is significant at $\alpha = .05$. Type primarily explains the variation at Level 3 (studies), rather than at Level 2 (proposals). For ease of interpretation, an equivalent model would be $y_{ij} = \beta_1 Grants_{ij} + \beta_2 Fellowship_{ij} + u_{(2)ij} + u_{(3)j} + e_{ij}$. The advantage of this parameterization is that the estimated effect sizes and their CIs for both grants and fellowships are available. After rerunning the analysis, the estimated effect sizes (and their 95% Wald CIs) for the grants and fellowships were −0.0066 (−0.0793, 0.0661) and −0.2022 (−0.2805, −0.1239), respectively.

Marsh et al. (2009) hypothesized a quadratic relationship exists between year of publication and effect sizes. Following Marsh et al., Year was standardized prior to creating a quadratic term. This analysis revealed that both Year and Year^2 were non-significant, $\chi^2(df = 2) = 3.4190$, $p = .1810$. Therefore, the evidence does not support this quadratic relationship.

To illustrate how an SEM approach can be used to handle missing covariates with FIML in the metaSEM package, I created missing values in Type. The purpose of this exercise was to illustrate the procedures rather than compare the results between a complete-case analysis and FIML. Readers may refer to Enders (2010) for the findings that compare different methods of handling missing data.

Approximately 25% of the data in Type were randomly deleted. Because this missingness was random, it is considered MCAR. The estimated intercept and slope (and their 95% Wald CIs) based on a complete case analysis were −0.0048 (−0.0820, 0.0723) and −0.2109 (−0.3157, −0.1061), respectively. $\hat{\tau}^2_{(2)}$ and $\hat{\tau}^2_{(3)}$ were 0.0045 and 0.0009, respectively. FIML was used to handle the missing covariate via the meta3X() function. The estimated intercept and slope (and their 95% Wald CIs) were −0.0106 (−0.0886, 0.0673) and −0.1753 (−0.2895, −0.0611), respectively. $\hat{\tau}^2_{(2)}$ and $\hat{\tau}^2_{(3)}$ were 0.0037 and 0.0037, respectively.

To create MAR data, values in Type were treated as missing when Year fell below 1996. Approximately 40% of the data in Type was missing. Because the missingness in Type depended on Year, it is MAR when Year is included in the model. The estimated intercept and slope (and their 95% Wald CIs) based on a complete case analysis were −0.0159 (−0.0933, 0.0616) and −0.1757 (−0.2998, −0.0517), respectively. $\hat{\tau}^2_{(2)}$ and $\hat{\tau}^2_{(3)}$ were 0.0026 and 0.0027, respectively. FIML was used to handle the missing covariate by including Year as a Level 2 auxiliary variable. The estimated intercept and slope (and their 95% Wald CIs) were −0.0264 (−0.1385, 0.0857) and −0.2004 (−0.3359, −0.0649), respectively. $\hat{\tau}^2_{(2)}$ and $\hat{\tau}^2_{(3)}$ were 0.0039 and 0.0030, respectively.

General Discussion and Future Directions

This article reviewed how three-level meta-analyses might address the dependence among effect sizes. The advantages of formulating three-level meta-analyses as structural equation models were discussed. Several examples illustrated how the metaSEM package can be used to conduct analyses. Because three-level meta-analytic applications are new, several questions remain unanswered and require additional study.

Empirical Comparisons Between Two- and Three-Level Meta-Analyses

This study demonstrated that the *SE*s (or CIs) of the fixed effects are likely inflated, whereas the heterogeneity of the random effects is overestimated using a traditional meta-analysis. Although three-level meta-analyses are theoretically preferable and methodologically apt for nested data, some researchers may still want to use two-level meta-analyses if they primarily focus on fixed-effects estimates.

Traditional two-level meta-analyses can be considered as three-level meta-analyses with misspecified covariance structures where $\tau^2_{(3)} = 0$. The use of robust statistics to address model misspecification is popular in the SEM literature (e.g., Yuan & Bentler, 2007). Many SEM packages have implemented types of robust statistics and *SE*s.

Similar ideas have also been suggested in the meta-analysis literature. Hedges, Tipton, and Johnson (2010) showed that robust (sandwich) *SE*s perform well in terms of coverage probability in traditional meta-analyses when testing fixed effects with dependent effect sizes. Importantly, however, τ^2 and I^2 are likely to be overestimated when one level is ignored. Because both three-level meta-analyses and robust *SE*s for dependent effect sizes are new, additional studies should address their pros and cons by analyzing dependent effect sizes.

Two-, Three-, or Even More Levels?

This article argued that three-level meta-analyses are the preferred way of handling dependent effect sizes in meta-analyses. In reality, data structures might be more complicated. Recall Nguyen and Benet-Martínez (2013), who analyzed 935 effect sizes nested within 141 studies. Because these authors were interested in culture, their analysis involved four levels: individuals (Level 1), multiple effect sizes within the study (Level 2), studies (Level 3), and cultures (Level 4). Rather than considering culture as a level, these authors treated cultural groups as a fixed effect and handled them using dummy variables. The authors considered studies to be the unit of analysis and averaged multiple effect sizes within studies.

The situation becomes even more complicated when researchers want to address the effect of publications by the same research teams or the same data sets (e.g., Shin, 2009) because the same authors might be involved in several different research teams and publications. Thus, the effect of authors is usually cross-classified

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

rather than nested. The current meta-analytic methodology is limited to three levels without cross-classified effects. Additional studies should address whether more levels (e.g., four) and cross-classified effects are necessary to better handle the data structure and address more complex research questions.

The Empirical Performance of I^2 and R^2 in Three-Level Meta-Analyses

In this article, the I^2 and R^2 statistics of traditional two-level meta-analyses were extended to three-level meta-analyses. Although these definitions appear to be intuitive, the empirical performance of these indices is unclear. With regard to I^2, comparing different estimates of the "typical" within-study variance is of interest. v_i was treated as a constant in this article, and $\tilde{v}$ was calculated without considering the cluster effect. Additional studies should address whether $\tilde{v}$ must be adjusted for the cluster effect.

In a traditional meta-analysis, it was found that different estimation methods can result in different R^2s (Aloe, Becker, & Pigott, 2010; López-López, Marín-Martínez, Sánchez-Meca, Van den Noortgate, & Viechtbauer, 2013). Three-level meta-analyses can only be analyzed with either ML or REML estimation methods. Further studies may address how these estimation methods affect R^2. Moreover, additional research should address the likelihood of obtaining negative values before truncation and how the inclusion of predictors at one level (e.g., Level 2) affects the R^2 at another level (e.g., Level 3).

Extensions to Multivariate Three-Level Meta-Analyses

This article and the metaSEM package only addressed univariate three-level meta-analyses. When study designs or measures are complicated, studies might report multiple effect sizes. These multivariate effect sizes can be nested within a higher level (e.g., studies or cultural groups). In theory, three-level multivariate meta-analyses can be implemented as structural equation models by extending multivariate meta-analyses (M. W.-L. Cheung, in press) and three-level meta-analyses (this article). Observing how three-level multivariate meta-analyses can be used to address more complex research questions is of interest.

To summarize, SEM is a viable alternative to model meta-analytic data. Showing how three-level meta-analyses can be integrated using an SEM approach further demonstrates the usefulness of SEM in meta-analyses. Many state-of-the-art SEM techniques are available to reviewers who conduct meta-analyses by using the SEM approach. This article also generated many new questions for future research with regard to three-level meta-analyses. Three-level meta-analyses can be a useful tool to address methodologically and substantively important questions in research synthesis. Finally, this article might further help to develop a unified SEM approach for primary, secondary, and meta-analyses.

References

Ahn, S., Ames, A. J., & Myers, N. D. (2012). A review of meta-analyses in education: Methodological strengths and weaknesses. *Review of Educational Research, 82,* 436–476. doi:10.3102/0034654312458162

Aloe, A. M., Becker, B. J., & Pigott, T. D. (2010). An alternative to R2 for assessing linear models of effect size. *Research Synthesis Methods, 1,* 272–283. doi:10.1002/jrsm.23

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Au, K. Y. (1999). Intra-cultural variation: Evidence and implications for international business. *Journal of International Business Studies, 30,* 799–812. doi:10.1057/palgrave.jibs.8490840

Au, K., & Cheung, M. W.-L. (2004). Intra-cultural variation and job autonomy in 42 countries. *Organization Studies, 25,* 1339–1362. doi:10.1177/0170840604046345

Becker, B. J. (1992). Using results from replicated studies to estimate linear models. *Journal of Educational Statistics, 17,* 341–362. doi:10.2307/1165128

Becker, B. J., Hedges, L. V., & Pigott, T. D. (2004). *Campbell Collaboration statistical analysis policy brief*. Retrieved from http://www.campbellcollaboration.org/artman2/uploads/1/C2_Statistical_Analysis_Policy_Brief-2.pdf

Boker, S., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T., . . . Fox, J. (2011). OpenMx: An open source extended structural equation modeling framework. *Psychometrika, 76,* 306–317. doi:10.1007/s11336-010-9200-6

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. doi:10.1002/9780470743386

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods, 1,* 97–111. doi:10.1002/jrsm.12

Bornmann, L., Mutz, R., & Daniel, H.-D. (2007). Gender differences in grant peer review: A meta-analysis. *Journal of Informetrics, 1,* 226–238. doi:10.1016/j.joi.2007.03.001

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Pacific Grove, CA: Brooks/Cole.

Cheung, M. W.-L. (2008). A model for integrating fixed-, random-, and mixed-effects meta-analyses into structural equation modeling. *Psychological Methods, 13,* 182–202. doi:10.1037/a0013163

Cheung, M. W.-L. (2009). Constructing approximate confidence intervals for parameters with structural equation models. *Structural Equation Modeling, 16,* 267–294. doi:10.1080/10705510902751291

Cheung, M. W.-L. (2010). Fixed-effects meta-analyses as multiple-group structural equation models. *Structural Equation Modeling, 17,* 481–509. doi:10.1080/10705511.2010.489367

Cheung, M. W.-L. (2012). *metaSEM: Meta-analysis using structural equation modeling*. Retrieved from http://courses.nus.edu.sg/course/psycwlm/Internet/metaSEM/

Cheung, M. W.-L. (2013). Implementing restricted maximum likelihood estimation in structural equation models. *Structural Equation Modeling, 20,* 157–167. doi:10.1080/10705511.2013.742404

Cheung, M. W.-L. (in press). Multivariate meta-analysis as structural equation models. *Structural Equation Modeling*.

Cheung, M. W.-L., & Chan, W. (2005). Meta-analytic structural equation modeling: A two-stage approach. *Psychological Methods, 10,* 40–64. doi:10.1037/1082-989X.10.1.40

Cheung, M. W.-L., & Chan, W. (2009). A two-stage approach to synthesizing covariance matrices in meta-analytic structural equation modeling. *Structural Equation Modeling, 16,* 28–53. doi:10.1080/10705510802561295

Cheung, S. F., & Chan, D. K. S. (2004). Dependent effect sizes in meta-analysis: Incorporating the degree of interdependence. *Journal of Applied Psychology, 89,* 780–791. doi:10.1037/0021-9010.89.5.780

Cochran, W. (1954). The combination of estimates from different experiments. *Biometrics, 10,* 101–129. doi:10.2307/3001666

Cooper, H. M. (2010). *Research synthesis and meta-analysis: A step-by-step approach* (4th ed.). Los Angeles, CA: Sage.

Cooper, H., & Hedges, L. V. (2009). Potentials and limitations. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of*

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

research synthesis and meta-analysis (2nd ed., pp. 561–572). New York, NY: Russell Sage Foundation.

Cooper, H., Valentine, J. C., Charlton, K., & Melson, A. (2003). The effects of modified school calendars on student achievement and on school and community attitudes. *Review of Educational Research, 73,* 1–52. doi:10.3102/00346543073001001

Curran, P. J. (2009). The seemingly quixotic pursuit of a cumulative psychological science: Introduction to the special issue. *Psychological Methods, 14,* 77–80. doi:10.1037/a0015972

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.

Fischer, R., & Boer, D. (2011). What is more important for national well-being: Money or autonomy? A meta-analysis of well-being, burnout, and anxiety across 63 societies. *Journal of Personality and Social Psychology, 101,* 164–184. doi:10.1037/a0023663

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5,* 3–8. doi:10.3102/0013189X005010003

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, NY: Russell Sage Foundation.

Goldstein, H. (2011). *Multilevel statistical models* (4th ed.). Hoboken, NJ: Wiley.

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science, 8,* 206–213. doi:10.1007/s11121-007-0070-9

Hanke, K., & Fischer, R. (2012). Socioeconomical and sociopolitical correlates of interpersonal forgiveness: A three-level meta-analysis of the Enright Forgiveness Inventory across 13 societies. *International Journal of Psychology*. Advance online publication. doi:10.1080/00207594.2011.651086

Hardy, R. J., & Thompson, S. G. (1996). A likelihood approach to meta-analysis with random effects. *Statistics in Medicine, 15,* 619–629. doi:10.1002/(SICI)1097-0258(19960330)15:6<619::AID-SIM188>3.0.CO;2-A

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association, 72,* 320–338. doi:10.1080/01621459.1977.10480998

Hedges, L. V. (2009). Statistical considerations. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 37–46). New York, NY: Russell Sage Foundation.

Hedges, L. V., & Pigott, T. D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods, 6,* 203–217. doi:10.1037/1082-989X.6.3.203

Hedges, L. V., & Pigott, T. D. (2004). The power of statistical tests for moderators in meta-analysis. *Psychological Methods, 9,* 426–445. doi:10.1037/1082-989X.9.4.426

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods, 1,* 39–65. doi:10.1002/jrsm.5

Hedges, L. V., & Vevea, J. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3,* 486–504. doi:10.1037/1082-989X.3.4.486

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21,* 1539–1558. doi:10.1002/sim.1186

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York, NY: Routledge.

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

Jackson, D., Riley, R., & White, I. R. (2011). Multivariate meta-analysis: Potential and promise. *Statistics in Medicine, 30,* 2481–2498. doi:10.1002/sim.4172

Konstantopoulos, S. (2011). Fixed effects and variance components estimation in three-level meta-analysis. *Research Synthesis Methods, 2,* 61–76. doi:10.1002/jrsm.35

López-López, J. A., Marín-Martínez, F., Sánchez-Meca, J., Van den Noortgate, W., & Viechtbauer, W. (2013). Estimation of the predictive power of the model in mixed-effects meta-regression: A simulation study. *British Journal of Mathematical and Statistical Psychology*. Advance online publication. doi:10.1111/bmsp.12002

MacKinnon, D. (2008). *Introduction to statistical mediation analysis*. New York, NY: Erlbaum.

Marín-Martínez, F., & Sánchez-Meca, J. (1999). Averaging dependent effect sizes in meta-analysis: A cautionary note about procedures. *The Spanish Journal of Psychology, 2,* 32–38.

Marsh, H. W., Bornmann, L., Mutz, R., Daniel, H.-D., & O'Mara, A. (2009). Gender effects in the peer reviews of grant proposals: A comprehensive meta-analysis comparing traditional and multilevel approaches. *Review of Educational Research, 79,* 1290–1326. doi:10.3102/0034654309334143

Matt, G. E., & Cook, T. D. (2009). Threats to the validity of generalized inferences. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 537–560). New York, NY: Russell Sage Foundation.

Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods, 10,* 259–284. doi:10.1037/1082-989X.10.3.259

Muthén, B. O., & Muthén, L. K. (2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Nam, I.-S., Mengersen, K., & Garthwaite, P. (2003). Multivariate meta-analysis. *Statistics in Medicine, 22,* 2309–2333. doi:10.1002/sim.1410

Neale, M. C., & Miller, M. B. (1997). The use of likelihood-based confidence intervals in genetic models. *Behavior Genetics, 27,* 113–120. doi:10.1023/A:1025681223921

Nguyen, A.-M. D., & Benet-Martínez, V. (2013). Biculturalism and adjustment: A meta-analysis. *Journal of Cross-Cultural Psychology, 44,* 122–159. doi:10.1177/0022022111435097

Owens, C. M., & Ferron, J. M. (2012). Synthesizing single-case studies: A Monte Carlo examination of a three-level meta-analytic model. *Behavior Research Methods, 44,* 795–805. doi:10.3758/s13428-011-0180-y

Patterson, H. D., & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika, 58,* 545–554. doi:10.1093/biomet/58.3.545

Pigott, T. D. (2012). *Advances in meta-analysis*. doi:10.1007/978-1-4614-2278-5

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed effects models in S and S-Plus*. doi:10.1007/978-1-4419-0318-1

Preacher, K. J. (2011). Multilevel SEM strategies for evaluating mediation in three-level data. *Multivariate Behavioral Research, 46,* 691–731. doi:10.1080/00273171.2011.589280

Raudenbush, S. W. (2009). Analyzing effect sizes: Random effects models. In H. M. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 295–315). New York, NY: Russell Sage Foundation.

Raudenbush, S. W., Becker, B. J., & Kalaian, H. (1988). Modeling multivariate effect sizes. *Psychological Bulletin, 103,* 111–120. doi:10.1037/0033-2909.103.1.111

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.

R Development Core Team. (2013). *R: A language and environment for statistical computing*. Retrieved from http://www.R-project.org/

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Riley, R. D., Lambert, P. C., & Abo-Zaid, G. (2010). Meta-analysis of individual participant data: Rationale, conduct, and reporting. *British Medical Journal, 340,* c221. doi:10.1136/bmj.c221

Rindskopf, D. (1984). Using phantom and imaginary latent-variables to parameterize constraints in linear structural models. *Psychometrika, 49,* 37–47. doi:10.1007/BF02294204

Rosenthal, R., & DiMatteo, M. R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology, 52,* 59–82. doi:10.1146/annurev.psych.52.1.59

Rosenthal, R., & Rubin, D. B. (1986). Meta-analytic procedures for combining studies with multiple effect sizes. *Psychological Bulletin, 99,* 400–406. doi:10.1037/0033-2909.99.3.400

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.). (2005). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. doi:10.1002/0470870168

SAS Institute. (2003). SAS/STAT software (Version 9.1) [Computer software]. Cary, NC.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7,* 147–177. doi:10.1037/1082-989X.7.2.147

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research, 33,* 545–571. doi:10.1207/s15327906mbr3304_5

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance components*. doi:10.1002/9780470316856

Self, S. G., & Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association, 82,* 605–610. doi:10.1080/01621459.1987.10478472

Shin, I.-S. (2009). *Same author and same data dependence in meta-analysis*. Unpublished doctoral dissertation, Florida State University, Tallahassee.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. doi:10.1201/9780203489437

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Los Angeles, CA: Sage.

Steiger, J. H., & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 221–257). Mahwah, NJ: Erlbaum.

Stevens, J. R., & Taylor, A. M. (2009). Hierarchical dependence in meta-analysis. *Journal of Educational and Behavioral Statistics, 34,* 46–73. doi:10.3102/1076998607309080

Stoel, R. D., Garre, F. G., Dolan, C., & van den Wittenboer, G. (2006). On the likelihood ratio test in structural equation modeling when parameters are subject to boundary constraints. *Psychological Methods, 11,* 439–455. doi:10.1037/1082-989X.11.4.439

Takkouche, B., Cadarso-Suárez, C., & Spiegelman, D. (1999). Evaluation of old and new tests of heterogeneity in epidemiologic meta-analysis. *American Journal of Epidemiology, 150,* 206–215. doi:10.1093/oxfordjournals.aje.a009981

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45,* 1–67.

Van den Bussche, E., Van den Noortgate, W., & Reynvoet, B. (2009). Mechanisms of masked priming: A meta-analysis. *Psychological Bulletin, 135,* 452–477. doi:10.1037/a0015329

Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2012). Three-level meta-analysis of dependent effect sizes. *Behavior Research Methods*. Advance online publication.. doi:10.3758/s13428-012-0261-6

Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics, 30,* 261–293. doi:10.3102/10769986030003261

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604. doi:10.1037/0003-066X.54.8.594

Xiong, C., Miller, J. P., & Morris, J. C. (2010). Measuring study-specific heterogeneity in meta-analysis: Application to an antecedent biomarker study of Alzheimer's disease. *Statistics in Biopharmaceutical Research, 2,* 300–309. doi:10.1198/sbr.2009.0067

Yuan, K.-H., & Bentler, P. M. (2007). Robust procedures in structural equation modeling. In S.-Y. Lee (Ed.), *Handbook of latent variable and related models* (pp. 367–397). Amsterdam, the Netherlands: Elsevier/North-Holland.

Yuan, K.-H., Yang-Wallentin, F., & Bentler, P. M. (2012). ML versus MI for missing data with violation of distribution conditions. *Sociological Methods & Research, 41,* 598–629. doi:10.1177/0049124112460373

(*Appendix follows*)

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Appendix

R Code for the Analyses

```
#### Example on Cooper et al. (2003)

library("metaSEM")
head(Cooper03)

#### Two-level meta-analysis

## Random-effects model
summary( meta(y=y, v=v, data=Cooper03) )

## Fixed-effects model
summary( meta(y=y, v=v, data=Cooper03, RE.constraints=0) )

## Mixed-effects model
summary( meta(y=y, v=v, x=scale(Year, scale=FALSE), data=Cooper03) )

#### Three-level meta-analysis

## Random-effects model
summary( meta3(y=y, v=v, cluster=District, data=Cooper03) )

## Mixed-effects model
summary( meta3(y=y, v=v, cluster=District, x=scale(Year, scale=FALSE),
               data=Cooper03) )

## Model comparisons

model2 <- meta(y=y, v=v, data=Cooper03, model.name="2 level model", silent=TRUE)
#### An equivalent model by fixing tau2 at level 3=0 in meta3()
## model2 <- meta3(y=y, v=v, cluster=District, data=Cooper03,
##                   model.name="2 level model", RE3.constraints=0)
model3 <- meta3(y=y, v=v, cluster=District, data=Cooper03,
                model.name="3 level model", silent=TRUE)
anova(model3, model2)
```

Testing $\tau^2_{(2)} = \tau^2_{(3)}$

```
modelEqTau2 <- meta3(y=y, v=v, cluster=District, data=Cooper03,
                     model.name="Equal tau2 in both levels",
                     RE2.constraints="0.1*Eq_tau2", RE3.constraints="0.1*Eq_tau2")
anova(model3, modelEqTau2)

## Likelihood-based CI
summary( meta3(y=y, v=v, cluster=District, data=Cooper03,
               I2=c("I2q", "ICC"), intervals.type="LB") )

summary( meta3(y=y, v=v, cluster=District, x=scale(Year, scale=FALSE),
               data=Cooper03, intervals.type="LB") )

## REML
summary( reml1 <- reml3(y=y, v=v, cluster=District, data=Cooper03) )

summary( reml0 <- reml3(y=y, v=v, cluster=District, data=Cooper03,
                        RE.equal=TRUE, model.name="Equal Tau2") )
```

(*Appendix continues*)

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

```
anova(reml1, reml0)

summary( reml3(y=y, v=v, cluster=District, x=scale(Year, scale=FALSE),
               data=Cooper03) )

#### Example on Bornmann et al. (2007)

library("metaSEM")
head(Bornmann07)

## Model 0: Intercept
summary( Model0 <- meta3(y=logOR, v=v, cluster=Cluster, data=Bornmann07,
                         model.name="3 level model") )

## Testing $\tau^2_{(3)} = 0$
Model0a <- meta3(logOR, v, cluster=Cluster, data=Bornmann07,
                 RE3.constraints=0, model.name="2 level model")
anova(Model0, Model0a)

## Testing $\tau^2_{(2)} = \tau^2_{(3)}$
Model0b <- meta3(logOR, v, cluster=Cluster, data=Bornmann07,
                 RE2.constraints="0.1*Eq_tau2", RE3.constraints="0.1*Eq_tau2",
                 model.name="tau2_2 equals tau2_3")
anova(Model0, Model0b)

## Model 1: Type as a predictor
## Convert characters into a dummy variable
## Type2=0 (Grants); Type2=1 (Fellowship)
Type2 <- ifelse(Bornmann07$Type=="Fellowship", yes=1, no=0)
summary( Model1 <- meta3(logOR, v, x=Type2, cluster=Cluster, data=Bornmann07))

## Alternative model: Grants and Fellowship as indicators
## Indicator variables
Grants <- ifelse(Bornmann07$Type=="Grants", yes=1, no=0)
Fellowship <- ifelse(Bornmann07$Type=="Fellowship", yes=1, no=0)

summary(Model1b <- meta3(logOR, v, x=cbind(Grants, Fellowship),
                         cluster=Cluster, data=Bornmann07,
                         intercept.constraints=0, model.name="Model 1"))

## Model 2: Year and Year^2 as predictors
summary( Model2 <- meta3(logOR, v, x=cbind(scale(Year), scale(Year)^2),
                         cluster=Cluster, data=Bornmann07,
                         model.name="Model 2") )

## Testing $\beta_{Year} = \beta_{Year^2} = 0$
anova(Model2, Model0)

## Model 3: Discipline as a predictor
## Create dummy variables using multidisciplinary as the reference group
DisciplinePhy <- ifelse(Bornmann07$Discipline=="Physical sciences", yes=1, no=0)
DisciplineLife <- ifelse(Bornmann07$Discipline=="Life sciences/biology", yes=1, no=0)
DisciplineSoc <- ifelse(Bornmann07$Discipline=="Social sciences/humanities", yes=1,
no=0)
summary( Model3 <- meta3(logOR, v, x=cbind(DisciplinePhy, DisciplineLife,
```

(*Appendix continues*)

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

```
DisciplineSoc),
                         cluster=Cluster, data=Bornmann07,
                         model.name="Model 3") )

## Testing whether Discipline is significant
anova(Model3, Model0)

## Model 4: Country as a predictor
## Create dummy variables using the United States as the reference group
CountryAus <- ifelse(Bornmann07$Country=="Australia", yes=1, no=0)
CountryCan <- ifelse(Bornmann07$Country=="Canada", yes=1, no=0)
CountryEur <- ifelse(Bornmann07$Country=="Europe", yes=1, no=0)
CountryUK <- ifelse(Bornmann07$Country=="United Kingdom", yes=1, no=0)

summary( Model4 <- meta3(logOR, v, x=cbind(CountryAus, CountryCan, CountryEur,
                         CountryUK), cluster=Cluster, data=Bornmann07,
                         model.name="Model 4") )

## Testing whether Discipline is significant
anova(Model4, Model0)

## Model 5: Type and Discipline as predictors
summary( Model5 <- meta3(logOR, v, x=cbind(Type2, DisciplinePhy, DisciplineLife,
                         DisciplineSoc), cluster=Cluster, data=Bornmann07,
                         model.name="Model 5") )

## Testing whether Discipline is significant after controlling for Type
anova(Model5, Model1)

## Model 6: Type and Country as predictors
summary( Model6 <- meta3(logOR, v, x=cbind(Type2, CountryAus, CountryCan,
                         CountryEur, CountryUK), cluster=Cluster, data=Bornmann07,
                         model.name="Model 6") )

## Testing whether Country is significant after controlling for Type
anova(Model6, Model1)

## Model 7: Discipline and Country as predictors
summary( meta3(logOR, v, x=cbind(DisciplinePhy, DisciplineLife, DisciplineSoc,
                         CountryAus, CountryCan, CountryEur, CountryUK),
                         cluster=Cluster, data=Bornmann07,
                         model.name="Model 7") )

## Model 8: Type, Discipline and Country as predictors
Model8 <- meta3(logOR, v, x=cbind(Type2, DisciplinePhy, DisciplineLife, DisciplineSoc,
                           CountryAus, CountryCan, CountryEur, CountryUK),
                           cluster=Cluster, data=Bornmann07,
                           model.name="Model 8")
## There was an estimation error. The model was rerun again.
summary(rerun(Model8))

#### Handling missing covariates with FIML

## MCAR
## Set seed for replication
set.seed(1000000)

## Copy Bornmann07 to my.df
```

(*Appendix continues*)

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

```
my.df <- Bornmann07
## "Fellowship": 1; "Grant": 0
my.df$Type_MCAR <- ifelse(Bornmann07$Type=="Fellowship", yes=1, no=0)

## Create 17 out of 66 missingness with MCAR
my.df$Type_MCAR[sample(1:66, 17)] <- NA

summary(meta3(y=logOR, v=v, cluster=Cluster, x=Type_MCAR, data=my.df))

summary(meta3X(y=logOR, v=v, cluster=Cluster, x2=Type_MCAR, data=my.df))

## MAR
Type_MAR <- ifelse(Bornmann07$Type=="Fellowship", yes=1, no=0)

## Create 27 out of 66 missingness with MAR for cases Year<1996
index_MAR <- ifelse(Bornmann07$Year<1996, yes=TRUE, no=FALSE)
Type_MAR[index_MAR] <- NA
summary(meta3(logOR, v, x=Type_MAR, cluster=Cluster, data=Bornmann07))

## Include auxiliary variable
summary(meta3X(y=logOR, v=v, cluster=Cluster, x2=Type_MAR, av2=Year, data=my.df))
```

Received September 11, 2012
Revision received April 7, 2013
Accepted April 10, 2013 ■

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

ORDER FORM

Start my 2014 subscription to ***Psychological Methods®***
ISSN: 1082-989X

___ $61.00 **APA MEMBER/AFFILIATE** ________
___ $125.00 **INDIVIDUAL NONMEMBER** ________
___ $542.00 **INSTITUTION** ________
In DC and MD add 6% sales tax ________
TOTAL AMOUNT DUE $ ________

Subscription orders must be prepaid. Subscriptions are on a calendar year basis only. Allow 4-6 weeks for delivery of the first issue. Call for international subscription rates.

AMERICAN PSYCHOLOGICAL ASSOCIATION

SEND THIS ORDER FORM TO
American Psychological Association
Subscriptions
750 First Street, NE
Washington, DC 20002-4242

Call **800-374-2721** or 202-336-5600
Fax **202-336-5568** :TDD/TTY **202-336-6123**
For subscription information,
e-mail: **subscriptions@apa.org**

❑ **Check enclosed** (make payable to APA)

Charge my: ❑ Visa ❑ MasterCard ❑ American Express

Cardholder Name ________
Card No. ________ Exp. Date ________

____
Signature *(Required for Charge)*

Billing Address
Street ________
City ________ State ________ Zip ________
Daytime Phone ________
E-mail ________

Mail To
Name ________
Address ________
____
City ________ State ________ Zip ________
APA Member # ________

META14