

A Multivariate Mixed Linear Model for Meta-Analysis

Hripsime A. Kalaian and Stephen W. Raudenbush
Michigan State University

A multivariate mixed-effects approach for meta-analysis is presented. The approach (a) incorporates as outcomes multiple effect sizes per study; (b) allows different studies to have different subsets of effect sizes; and (c) treats each study's effect sizes as random realizations from a population of possible effect sizes. Application is illustrated via reanalysis of data from studies assessing the effects of coaching on verbal and mathematical subtests of the Scholastic Aptitude Test. Covariance components are estimated via restricted maximum likelihood (REML); inferences about regression coefficients and specific study effect sizes are based on their joint conditional distribution given the REML covariance component estimates. The approach can be implemented via now-standard software for unbalanced hierarchical data.

Meta-analysis (Glass, 1976) or quantitative research synthesis (Hedges & Olkin, 1985) is the statistical analysis of data from a collection of independent studies that test the same hypotheses. A statistical indicator often used for integrating primary study results is an effect size estimate, for example, a standardized mean difference between the experimental and control groups from each individual study. Most meta-analyses to date have treated such effect size estimates as independent, and this is appropriate when each study produces a single effect size estimate. However, studies commonly produce multiple effect size estimates, and methodologists have recently proposed statistical methods for such multivariate effect size data (Gleser & Olkin, 1994; Hedges & Olkin, 1985; Raudenbush, Becker, & Kalaian, 1988; Rosenthal & Rubin, 1986). This article extends these multivariate methods to allow flexible modeling of variation between and within studies using a mixed effects regression model.

Hripsime A. Kalaian and Stephen W. Raudenbush, Department of Counseling, Educational Psychology, and Special Education, Michigan State University.

Research reported here was partially supported by the Project on Human Development in Chicago Neighborhoods with funding from the MacArthur Foundation and the National Institute of Justice.

Correspondence concerning this article should be addressed to Hripsime A. Kalaian, Department of Counseling, Educational Psychology, and Special Education, Michigan State University, East Lansing, Michigan 48824. Electronic mail may be sent via Internet to 15718HAK@msu.edu.

Rationale for a Multivariate Mixed Model

Multivariate Effect Size Data

Gleser and Olkin (1994) considered two sources of multivariate effect size data. First, a study with multiple treatments will report two or more comparisons for a given dependent variable. Second, a study may have multiple dependent variables and report an effect size for each of them. This article focuses on the second case, though the methods we propose readily extend to the case of multiple treatments.

Consider a setting in which each study reports two or more effect sizes, one for each dependent variable. For example, in the illustrative data to be presented later, several studies of the effect of coaching on the Scholastic Aptitude Test (SAT) report treatment-control contrasts for the verbal and math subtests of the SAT. Although the analyst may consider each dependent variable in a separate analysis, there are several good reasons to pursue multivariate analyses, that is, analyses that take into account the correlation between the multiple dependent variables reported in a study.

First, it may be of interest to discover whether an experimental treatment produces a larger effect for one dependent variable than for another. For example, one might ask whether coaching, on average, produces a larger effect on math SAT than on verbal SAT. Testing this hypothesis requires an analysis that takes into account the correlation between the effect size estimates in each study.

Second, one might be interested in whether mod-

erator variables relate differently to different outcomes. In the illustrative example, we test the hypothesis that the duration of coaching relates to the magnitude of the experimental effect, and we ask whether the association between duration and effect size is similar for the verbal and math subtests. Again, a multivariate approach is required.

Third, multivariate tests can protect the investigator against errors of statistical inference that can arise when the investigators compute multiple significance tests, one for each dependent variable (Gleser & Olkin, 1994, sect. 4). For example, we can test the joint null hypothesis that coaching has no relationship to the math or the verbal effect size.

Although multivariate analyses are often desirable in analyzing effect size data, the analyst will rarely encounter a stream of research in which every study reports results for exactly the same set of dependent variables. Rather, different studies will report results for different subsets of dependent variables. For example, in research on teacher expectancy effects, one will encounter as dependent variables teacher behavior, student attitudes, and student IQ. However, not all studies will measure all three of these dependent variables (Raudenbush, 1984). Raudenbush et al. (1988) and Gleser and Olkin (1994) developed fixed-effects regression approaches that allowed different numbers of effect sizes in different studies. Estimation via generalized least squares allowed flexible modeling of features of study design, implementation, and sampling as predictors of effect size.

Mixed Regression Models for Effect Size Data

Meta-analysts typically ask whether results from a stream of research are consistent. If the results are consistent, a reasonable summary will often be an estimated mean effect size and its standard error. However, if study results are heterogeneous, a more elaborate summary is needed. One might use a random effects model to estimate the grand mean effect size and the standard deviation of the true effect sizes (Hedges, 1983). The random effects are the deviations of a study's true effect sizes from grand mean effect size. Alternatively, one might adopt a fixed effects regression model that uses information about differences between studies to account for variation between studies. (See Hedges, 1994, for a recent review.) A third approach combines the random and fixed approaches in a mixed model (Raudenbush & Bryk, 1985). The mixed model includes regression coefficients that link study characteristics to study

outcomes. However, unlike the fixed effects approach, the mixed model allows for the possibility that the regression model is not completely successful in accounting for variation between studies. The random effects in this model are residuals, deviations between the true effect size and the effect size predicted by the model. The mixed model is most useful when the number of studies in a meta-analysis is large and when study outcomes are determined by numerous influences, not all of which are measurable (see Raudenbush, 1994, for a review).

Below, we consider a mixed model for multivariate effect size data. Each study is conceived to produce two or more true effect sizes, and these vary according to a multivariate normal distribution defined on a universe of studies. The model is flexible in allowing different numbers of effect sizes per study. Thus the model combines the advantages of the multivariate fixed effects models and the univariate mixed models reviewed above. Estimation of between-studies variance and covariance components is based on restricted maximum likelihood (REML). Estimation of regression coefficients and particular effect sizes is based on their conditional distributions given the REML estimates and the data, leading to generalized least squares estimates of the regression coefficients and empirical Bayes estimates of individual study effect sizes. Thus, the estimation approach is a multivariate extension of the univariate approach of Raudenbush and Bryk (1985) and can be accomplished using now-standard software for hierarchical data. Emphasis in this article is therefore on application rather than estimation theory. The illustrative example shows how the investigator can use potentially different sets of study characteristics to predict different outcomes, ask whether a predictor is more strongly related to one outcome than to another, and test the fit of alternative variance-covariance models.

A: Multivariate Mixed-Effects Linear Model

Our model builds on the mixed model with univariate empirical Bayes estimation (Raudenbush & Bryk, 1985), the multivariate fixed-effects model with generalized least squares (Raudenbush et al., 1988), and the mixed-effects regression model with deficient rank data (Braun, Jones, Rubin, & Thayer, 1983). We apply the resulting model to multivariate random effect size data, using empirical Bayes estimation.

Presentation of the model in two stages clarifies its logic. At the first stage, a "within-study" model

specifies which effect sizes are present and which are absent in each study and represents the estimation error within each study. Thus, the first-stage model may be viewed as a measurement model relating the estimated effect sizes from each study to the "true" effect sizes. The second-stage, or "between-studies" model, specifies the distribution of these true effect sizes across a universe of studies.

Within-Study Model

We associate with each study i a complete vector of M true effect sizes, $\delta_i = (\delta_{1i}, \dots, \delta_{Mi})^T$. In our illustrative example, two effect sizes are possible; thus, $M = 2$ and $\delta_i = (\delta_{1i}, \delta_{2i})^T$, where δ_{1i} is the true effect size for SAT-Verbal, and δ_{2i} is the true effect size for SAT-Math. Although M true effect sizes are associated with study i , only P_i effect size estimates are reported by study i , $P_i \leq M$. The vector of effect size estimates for study i is $d_i = (d_{1i}, \dots, d_{P_i})^T$. The effect size estimate d_{pi} , where $p = 1, \dots, P_i$, is linked to the true effect size, δ_{mi} , where $m = 1, \dots, M$, by indicator variable X_{pmi} via the linear model

$$d_{pi} = \sum_{m=1}^{M_i} \delta_{mi} X_{pmi} + e_{pi} \tag{1}$$

where $X_{pmi} = 1$ if d_{pi} estimates δ_{mi} and $X_{pmi} = 0$ otherwise; and e_{pi} denotes the error with which d_{pi} estimates the relevant δ_{mi} .

To illustrate the within-study model, suppose the reviewer has K studies and that some of these report the maximum $M = 2$ effect size estimates (one for SAT-Math and one for SAT-Verbal). However, some of these studies report only the SAT-Verbal effect size, and some report only the SAT-Math effect size. How would Equation 1 represent each of these possibilities?

If study i reports both effect size estimates, we have $P_i = M = 2$. Assuming the first effect size estimate to be for SAT-Verbal, Equation 1 becomes, for $p = 1$,

$$\begin{aligned} d_{1i} &= \delta_{1i} X_{11i} + \delta_{2i} X_{12i} + e_{1i} \\ &= \delta_{1i} * (1) + \delta_{2i} * (0) + e_{1i} \\ &= \delta_{1i} + e_{1i} \end{aligned} \tag{2}$$

For $p = 2$,

$$\begin{aligned} d_{2i} &= \delta_{1i} X_{21i} + \delta_{2i} X_{22i} + e_{2i} \\ &= \delta_{1i} * (0) + \delta_{2i} * (1) + e_{2i} \\ &= \delta_{2i} + e_{2i} \end{aligned} \tag{3}$$

In matrix notation, we have

$$d_i = X_i \delta_i + e_i \tag{4a}$$

or

$$\begin{bmatrix} d_{1i} \\ d_{2i} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \delta_{1i} \\ \delta_{2i} \end{bmatrix} + \begin{bmatrix} e_{1i} \\ e_{2i} \end{bmatrix} \tag{4b}$$

Suppose, however, that study i reports only the SAT-Verbal effect size. In this case, $p = 1$, and we have

$$\begin{aligned} d_{1i} &= \delta_{1i} * (1) + \delta_{2i} * (0) + e_{1i} \\ &= \delta_{1i} + e_{1i} \end{aligned} \tag{5}$$

or in matrix notation,

$$\begin{bmatrix} d_{1i} \end{bmatrix} = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} \delta_{1i} \\ \delta_{2i} \end{bmatrix} + \begin{bmatrix} e_{1i} \end{bmatrix} \tag{6}$$

If study i reports only the SAT-Math effect size, again $p = 2$, but now we have

$$\begin{aligned} d_{2i} &= \delta_{2i} * (0) + \delta_{1i} * (1) + e_{2i} \\ &= \delta_{2i} + e_{2i} \end{aligned} \tag{7}$$

or in matrix notation,

$$\begin{bmatrix} d_{2i} \end{bmatrix} = \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} \delta_{1i} \\ \delta_{2i} \end{bmatrix} + \begin{bmatrix} e_{2i} \end{bmatrix} \tag{8}$$

An essential feature of the model is that regardless of how many effect sizes are actually estimated in a given study, we conceive of that study as having M latent true effect sizes, one for each outcome variable. The problem of estimation then becomes an incomplete data problem and can be handled readily within the framework of maximum likelihood.

The variances and covariances of the within-study errors— e_{pi} , where $p = 1, \dots, P_i$ in Equation 1—will depend on the problem at hand. If the estimated effect size, d_{pi} , is a standardized mean difference (e.g., the difference between the experimental and control means divided by a pooled standard deviation), we have the following consistent estimator of the error variance (Hedges, 1981):

$$\text{Var}(e_{pi}) = V_{pi} = \frac{n_i^E + n_i^C}{n_i^E n_i^C} + \frac{d_{pi}^2}{2(n_i^E + n_i^C)} \tag{9}$$

where n_i^E, n_i^C are the experimental and control sample sizes, respectively, in study i .¹

¹ Hedges (1981) has found a correction that improves the sample estimate of the standardized mean difference between two groups. The sample standardized mean difference is multiplied by $[1 - 3/4(m_i - 1)]$, where $m_i = n_i^E + n_i^C - 2$. This correction was used in the illustrative example.

The covariance between a pair of within-study errors, e_{pi} and $e_{p'i}$, may similarly be consistently estimated as follows (Gleser & Olkin, 1994):

$$V_{pp'i} = \text{Cov}(e_{pi}, e_{p'i}) = \left(\frac{1}{n_i^E} + \frac{1}{n_i^C} \right) \rho_{pp'i} + \frac{1}{2} \frac{\delta_{pi} \delta_{p'i}}{n_i^E + n_i^C}, \quad (10)$$

where $\rho_{pp'i}$ is the correlation between the dependent variables associated with estimated effect sizes p and p' . This correlation may be estimated from the sample, deduced from published test information or imputed on the basis of past research. Gleser and Olkin (1994) provided the needed variance and covariance expressions for the case of multiple treatments per study. In general, we assume the Level 1 errors to be multivariate normal in distribution, that is $e_i \sim N(0, V_i)$, where V_i is a $P_i \times P_i$ covariance matrix constituted of elements computed with Equations 9 and 10.

Between-Studies Model

At the second stage, the M latent true effect sizes for each study become the outcome variables in a linear regression model:

$$\delta_{mi} = \gamma_{m0} + \sum_{q=1}^Q \gamma_{mq} W_{qi} + u_{mi}, \quad (11)$$

where the M residuals u_{mi} , $m = 1, \dots, M$, are assumed M -variate normal with null means and covariance matrix τ . In matrix notation, the model can be written

$$\delta_{\tau} = W_{\tau} \gamma + u_{\tau}, \quad u_{\tau} \sim N(0, \tau). \quad (12)$$

Consider, for example, the SAT coaching example with a single W variable (duration of coaching). Equation 12 could then be written as

$$\begin{bmatrix} \delta_{1i} \\ \delta_{2i} \end{bmatrix} = \begin{bmatrix} 1 & W_{1i} & 0 & 0 \\ 0 & 0 & 1 & W_{2i} \end{bmatrix} \begin{bmatrix} \gamma_{10} \\ \gamma_{11} \\ \gamma_{20} \\ \gamma_{21} \end{bmatrix} + \begin{bmatrix} u_{1i} \\ u_{2i} \end{bmatrix}, \quad (13)$$

where

$$\begin{bmatrix} u_{1i} \\ u_{2i} \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{11}^2 & \tau_{12} \\ 0 & \tau_{21} & \tau_{22}^2 \end{bmatrix} \right). \quad (14)$$

Here W_{1i} and W_{2i} represent hours of coaching for the SAT-Verbal and SAT-Math outcome measures in

study i , respectively; γ_{10} and γ_{20} are the intercepts for the SAT-Verbal and SAT-Math effects; and γ_{11} and γ_{21} are the effects of the hours of coaching on SAT-Verbal and SAT-Math, respectively.

Estimation Theory

Substituting the between-studies model (Equation 12) into the within-study model (Equation 4) yields the combined model

$$d_i = X_i W_i \gamma + X_i \mu_i + e_i, \quad (15)$$

a special case of a two-level hierarchical model (Raudenbush, 1988, Case 2) that is itself a special case of the general mixed model of Dempster, Rubin, and Tsutakawa (1981), who derived parameter estimates using the expectation-maximization (EM) algorithm. Inferences about the variance components in τ are based on REML. Inferences about the fixed effects, γ , are based on generalized least squares, given the REML estimates of τ ; and inferences about δ_{pi} , $i = 1, \dots, K$, are based on their conditional distributions given REML estimates of τ . The estimates of δ_i are known as "empirical Bayes" estimates (see Morris, 1983).²

Application to SAT Coaching Data

SAT coaching studies (Kalaian, 1994; Kalaian & Raudenbush, 1994) illustrate application of the multivariate mixed-effects linear model for meta-analysis to educational research. These coaching studies are analyzed using the Hierarchical Linear Model (HLM) software of Bryk, Raudenbush, and Congdon (1993; see the Appendix for details), which is designed for analyzing multilevel data sets.

Twenty samples in this review were coached for both SAT-Verbal and SAT-Math subtests. Another 18 samples were coached for only the SAT-Verbal subtest, while the other 9 samples were coached for SAT-Math.

The values of the SAT-Verbal effect sizes for the 38 samples ranged from -0.35 to 0.74 in standard deviation units with an overall weighted average

² We note that empirical Bayes provides a full complement of M estimated δ s for every study, even for those studies that report estimates of just a subset of the possible effect sizes. In fact, the analysis described here can readily be used to provide model-based imputations for the "missing" effect sizes.

coaching effect of 0.12 and standard deviation 0.22, while the 29 SAT-Math effect sizes ranged from -0.53 to 0.60 with an overall weighted average of 0.11 and standard deviation 0.28 (Table 1). Thus, the average effect of coaching on SAT-Verbal and SAT-Math gains appear to be quite similar. Note that the SAT-Math average effects are a bit smaller than those reported in previous reviews but that the SAT-Verbal effects are about the same. Although most of the coaching effect sizes are positive, the magnitudes of the coaching effects appear quite variable for both subtests.

We illustrate the modeling of the bivariate outcome using the model of Equation 12. For the 29 SAT-Math and the 38 SAT-Verbal data points in this review, the student contact hours ranged from 4 to 63 hours for both subtests with average coaching hours of 15 for SAT-Math and 17 for SAT-Verbal. Most of the data points are clustered at the low end of the number of hours dimension (Kalaian & Raudenbush, 1994, Table 3). For this reason and the fact that there are likely diminishing returns of increased hours in both SAT subtests' scores (Messick & Jungeblut, 1981), we used in the analysis the natural logarithmic transformation of the hours of coaching dimension. The scatterplot (see Figure 1) shows that SAT coaching effect size is moderately related to logarithmically transformed contact hours ($r = .5$ for SAT-Verbal and $r = .4$ for SAT-Math). Thus, both SAT-Verbal and SAT-Math with logarithmically transformed hours of coaching (see Hours column in Table 1) are modeled jointly.

Results are displayed in Table 2 (*Unconstrained Model*). There is some evidence that after controlling for hours of coaching, residual variation remains to be explained in both SAT-Verbal ($\hat{\tau}_1^2 = 0.0077$) and SAT-Math ($\hat{\tau}_1^2 = 0.0285$). We also see some evidence of positive effects of coaching hours for SAT-Verbal ($\hat{\gamma}_{11} = 0.058, t = 1.74, p = .09$), and for SAT-Math ($\hat{\gamma}_{21} = 0.149, t = 2.42, p = .02$). However, several questions arise regarding the adequacy of the model, the variance-covariance structure, and the fixed effects structure; these can be addressed by multivariate tests.

Variance-Covariance Structure

Is it necessary to view both the SAT-Verbal and SAT-Math effect sizes as significantly varying, with hours of coaching controlled? In particular, the variance component for SAT-Verbal appears small. To examine the possibility that hours of coaching com-

pletely accounts for the between-studies variation in multivariate effect sizes, we estimate a model that constrains the variance of the SAT-Verbal effects to be null (also implying that the covariance between SAT-Verbal and SAT-Math is null). Formally, we are testing the joint hypothesis

$$H_0: \tau_1^2 = 0; \tau_{12} = 0. \tag{16}$$

Results are given in Table 2 (*Constrained Model*). To evaluate the fit of the constrained model, we compare its deviance (-2 times the log-likelihood evaluated at the maximum) to the deviance associated with the unconstrained model. The difference between these deviances is distributed asymptotically as chi-square with degrees of freedom equal to the difference in the number of variance-covariance parameters estimated. As Table 2 shows, this difference is 6.93 ($df = 2$), suggesting that the simpler model is unjustified at $p = .03$. Thus, the model with nonzero conditional variation for both SAT-Verbal and SAT-Math appears to provide a better fit than does the constrained model.

Fixed Effects

Is there evidence that hours of coaching is required in the model? Separate univariate tests give a nonsignificant result for SAT-Verbal and significant result for SAT-Math. However, one may prefer an omnibus test of the joint hypothesis that the effect of hours of coaching is null, that is

$$H_0: \gamma_{11} = \gamma_{21} = 0, \tag{17}$$

or

$$H_0: \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \gamma_{10} \\ \gamma_{11} \\ \gamma_{20} \\ \gamma_{21} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{18}$$

This is a multivariate hypothesis of the form $C^T \gamma = 0$. Under this hypothesis, the statistic $C^T \hat{\gamma} [C^T \text{Var}(\hat{\gamma}) C]^{-1} \hat{\gamma}^T C$ has a large sample chi-square distribution with degrees of freedom equal to the number of rows of C (Bryk & Raudenbush, 1992, chap. 3). In our case, this statistic takes on a value of 9.30 ($df = 2, p < .01$), implying that hours of coaching are needed in the model. The omnibus test leaves unexamined the question of whether each coefficient is different from zero. Univariate t ratios (Table 2, unconstrained model) give contradictory results for SAT-Verbal ($p = .09$) and for SAT-Math ($p = .02$). Does this imply

Table 1
Effect Sizes of SAT Coaching Studies

Study	Year	n^E	n^C	d_1 SAT V	d_1 SAT M	Hr	ETS	Study type	Home work
Randomized studies									
Alderman & Powers (A)	1980	28	22	0.22	—	7	1	1	1
Alderman & Powers (B)	1980	39	40	0.09	—	10	1	1	1
Alderman & Powers (C)	1980	22	17	0.14	—	10.5	1	1	1
Alderman & Powers (D)	1980	48	43	0.14	—	10	1	1	1
Alderman & Powers (E)	1980	25	74	-0.01	—	6	1	1	1
Alderman & Powers (F)	1980	37	35	0.14	—	5	1	1	1
Alderman & Powers (G)	1980	24	70	0.18	—	11	1	1	1
Alderman & Powers (H)	1980	16	19	0.01	—	45	1	1	1
Evans & Pike (A)	1973	145	129	0.13	0.12	21	1	1	1
Evans & Pike (B)	1973	72	129	0.25	0.06	21	1	1	1
Evans & Pike (C)	1973	71	129	0.31	0.09	21	1	1	1
Laschewer	1986	13	14	0.00	0.07	8.9	0	1	0
Roberts & Oppenheim (A)	1966	43	37	0.01	—	7.5	1	1	0
Roberts & Oppenheim (B)	1966	19	13	0.67	—	7.5	1	1	0
Roberts & Oppenheim (D)	1966	16	11	-0.38	—	7.5	1	1	0
Roberts & Oppenheim (E)	1966	20	12	-0.24	—	7.5	1	1	0
Roberts & Oppenheim (F)	1966	39	28	0.29	—	7.5	1	1	0
Roberts & Oppenheim (G)	1966	38	25	—	0.26	7.5	1	1	0
Roberts & Oppenheim (H)	1966	18	13	—	-0.41	7.5	1	1	0
Roberts & Oppenheim (I)	1966	19	13	—	0.08	7.5	1	1	0
Roberts & Oppenheim (J)	1966	37	22	—	0.30	7.5	1	1	0
Roberts & Oppenheim (K)	1966	19	11	—	-0.53	7.5	1	1	0
Roberts & Oppenheim (L)	1966	17	13	—	0.13	7.5	1	1	0
Roberts & Oppenheim (M)	1966	20	12	—	0.26	7.5	1	1	0
Roberts & Oppenheim (N)	1966	20	13	—	0.47	7.5	1	1	0
Zaman (B)	1988	16	17	0.13	0.48	24	0	1	1
Matched studies									
Barke (A)	1986	25	25	0.50	—	50	0	2	1
Barke (B)	1986	25	25	0.74	—	50	0	2	1
Coffin	1987	8	8	-0.23	0.33	18	0	2	0
Davis	1985	22	21	0.13	0.13	15	0	2	0
Frankel	1960	45	45	0.13	0.34	30	0	2	0
Kintisch	1979	38	38	0.06	—	20	0	2	1
Whitla	1962	52 ^a	52 ^a	0.09	-0.11	10	1	2	1
Nonequivalent comparison studies									
Curran (A)	1988	21	17	-0.10	-0.08	6	0	3	0
Curran (B)	1988	24	17	-0.14	-0.29	6	0	3	0
Curran (C)	1988	20	17	-0.16	-0.34	6	0	3	0
Curran (D)	1988	20	17	-0.07	-0.06	6	0	3	0
Dear	1958	60	526	-0.02	0.21	15	1	3	1
Dyer	1953	225	193	0.06	0.17	15	1	3	1
French (B)	1955	110	158	0.06	—	4.5	1	3	1
French (C)	1955	161	158	—	0.20	15	1	3	1
FTC	1978	192	684	0.15	0.03	40	0	3	0
Keefauver	1976	16	25	0.17	-0.19	14	0	3	0
Lass	1961	38	82	0.02	0.10	—	1	3	1
Reynolds & Oberman	1987	93	47	-0.04	0.60	63	0	3	1
Teague	1992	10	15	0.40	—	18	0	3	0
Zaman (A)	1988	21	34	0.54	0.57	27	0	3	1

Note. SAT = Scholastic Aptitude Test; V = verbal; M = math; ETS = Educational Testing Service; FTC = Federal Trade Commission. The letters after the author(s) names refer to the fact that a given author may have replicated the experiment on multiple samples.

^a The sample sizes for SAT-M were $n^E = 50$ and $n^C = 50$.

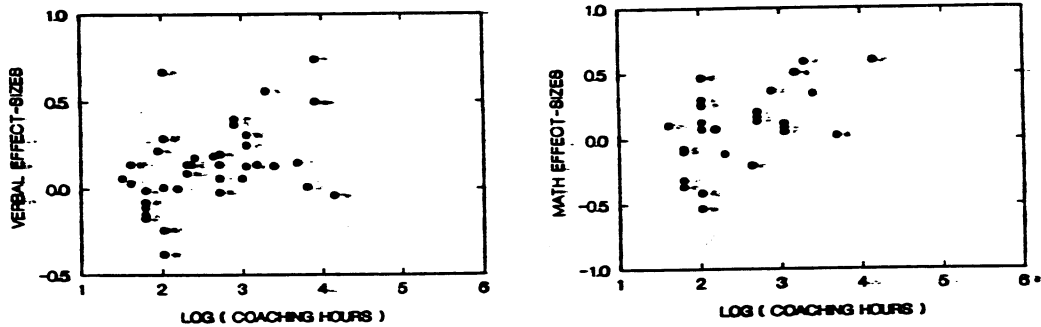


Figure 1. Relationships between Scholastic Aptitude Test effect sizes and log (contact time).

that the effect of coaching is less pronounced for the verbal test than for the math test? To address this question, we pose the null hypothesis

$$H_0: \gamma_{21} - \gamma_{11} = 0, \quad (19)$$

or

$$H_0: [0 \quad -1 \quad 0 \quad 1] \begin{bmatrix} \gamma_{10} \\ \gamma_{11} \\ \gamma_{20} \\ \gamma_{21} \end{bmatrix} = 0, \quad (20)$$

another example of a multivariate hypothesis test. The result in this case is $\chi^2(1, K = 47) = 1.61, p = .20$. Thus, there is little evidence that the effect of hours of coaching differs for SAT-Verbal and SAT-Math. Of course, this test may lack power given the number of studies available. It appears most judicious to allow separate effects of hours to remain in the model with the understanding that these effects may be very simi-

lar. Alternatively, one could constrain these effects to be the same.

A Note on Multivariate Hypothesis Testing

Our analysis is based on REML estimation of the variance-covariance components. Bryk and Raudenbush (1992) described the trade-offs between restricted and full maximum likelihood. When using REML, likelihood ratio tests can be computed for only the variance-covariance components (as in Equation 15). The fixed effects part of the model must be held constant in testing alternative models for the variance-covariance components. Multivariate tests involving the fixed effects can be tested with a Wald test as above (Equations 18 and 20).

Summary of Findings

The findings suggest a positive effect of hours of coaching on the math subtest of the SAT. The effect

Table 2
Fitting Multivariate Mixed Linear Model to SAT Coaching Data

Predictor	Unconstrained model				Constrained model			
	Coefficient	SE	t ratio	p	Coefficient	SE	t ratio	p
Fixed effects of fitting unconstrained and constrained models								
For SAT-Verbal								
Intercept	$\hat{\gamma}_{10} = 0.103$	0.025	4.181	.000	$\hat{\gamma}_{10} = 0.105$	0.0179	5.860	.000
Log (hr)	$\hat{\gamma}_{11} = 0.058$	0.033	1.743	.088	$\hat{\gamma}_{11} = 0.051$	0.023	2.262	.033
For SAT-Math								
Intercept	$\hat{\gamma}_{20} = 0.099$	0.042	2.328	.029	$\hat{\gamma}_{20} = 0.116$	0.046	2.509	.020
Log (hr)	$\hat{\gamma}_{21} = 0.149$	0.061	2.417	.024	$\hat{\gamma}_{21} = 0.175$	0.068	2.576	.017
Variance-covariance results for unconstrained and constrained models								
				$\hat{\gamma}_{12}$ constrained to 0				
$\hat{\gamma}_{12} = 0.00768$				$\hat{\gamma}_{12} = -0.00835$				
$\hat{\gamma}_{22} = 0.02848$				$\hat{\gamma}_{22} = 0.03957$				
Deviance = 275.15				Deviance = 282.08				
df = 3				df = 1				

Note. SAT = Scholastic Aptitude Test.

of hours on SAT-Verbal was less clear; this effect, though not significantly different from the effect of hours on SAT-Math, failed to achieve conventional significance levels in a univariate test. After controlling for hours of coaching, there was evidence of significant variation between studies for both outcomes, implying that study characteristics not specified in the model are related to study differences in both outcomes. We examined the other predictors given in Table 1 for their relationship to effect sizes controlling for hours of coaching. No significant relationships were found. Finally, we note that when hours of coaching are held constant at the mean, the expected effect size is significant both for SAT-Verbal ($\hat{\gamma}_{10} = 0.103$, $t = 4.18$, $p = .00$) and for SAT-Math ($\hat{\gamma}_{20} = 0.099$, $t = 2.33$, $p = .03$).³

Conclusion and Discussion

The model developed and presented in this study represents a quite flexible approach for quantitative meta-analysis and research synthesis. It is designed for synthesizing studies with multiple correlated effect sizes when these effect sizes are assumed to be random realizations from the multivariate effect size population. Thus, this approach includes both multivariate fixed and random effects. An essential feature of the model is that it allows different numbers of multiple effect sizes for each study. The methodology extends readily to multivariate meta-analysis using other outcomes (e.g., correlations or log-odds ratios) or multiple contrasts per study.

The flexibility of the model in handling unequal numbers of outcomes per study is clearly an advantage. However, the investigator must examine whether the patterns of "missing data" are associated with different study results. In our example, we defined a dummy variable to indicate whether a study had measured SAT-Verbal but not SAT-Math and a second dummy variable to indicate whether a study had measured SAT-Math but not SAT-Verbal. Using these dummies in the respective between-studies models for SAT-Verbal and SAT-Math, we found no significant differences in SAT-Verbal outcomes between studies that measured both and those that measured SAT-Verbal only. A similar finding of no difference was manifest for the SAT-Math outcome. On the basis of these results, we viewed as justified the analysis pooling data from all studies.

In the future, the application of the illustrated methodology might be further applied to meta-analyses

with more than two outcomes with or without missing effect sizes. Also, these new applications might consider taking into account the within-study characteristics and incorporating them in the model. An useful extension of the approach would involve Bayesian estimation that fully takes into account the uncertainty about τ , the between-studies variance-covariance matrix (Seltzer, 1993). This extension would be most needed when the number of studies, K , is small.

³ The intercepts of the model are estimated mean effect sizes because *hours* was centered around its mean in estimating the model.

References

- Braun, H. L., Jones, D. H., Rubin, D. B., & Thayer, D. T. (1983). Empirical Bayes estimation of coefficients in the general linear model with data of deficient rank. *Psychometrika*, *48*(2), 171-181.
- Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.
- Bryk, A. S., Raudenbush, S. W., & Congdon, R. (1993). *An introduction to HLM: Computer programs and user's manual*. Chicago: University of Chicago, Department of Education.
- Dempster, A., Rubin, D., & Tsutakawa, R. (1981). Estimation in covariance components models. *Journal of the American Statistical Association*, *76*, 341-353.
- Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*, 3-8.
- Gleser, L. J., & Olkin, I. (1994). Stochastically dependent effect sizes. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 339-355). New York: Russell Sage Foundation.
- Hedges, L. V. (1981). Distribution theory for Glass's estimator for effect size and related estimators. *Journal of Educational Statistics*, *6*(2), 107-128.
- Hedges, L. V. (1983). A random effects model for effect size. *Psychological Bulletin*, *93*, 388-395.
- Hedges, L. V. (1994). Fixed effects models. In H. Cooper & L. Hedges (Eds.), *The handbook of research synthesis* (pp. 285-300). New York: Russell Sage Foundation.
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. New York: Academic Press.
- Kalaian, H. A. (1994). *A multivariate mixed linear model for meta-analysis*. Unpublished doctoral dissertation, Michigan State University, East Lansing.
- Kalaian, H. A., & Raudenbush, S. W. (1994, April). *Scho-*

lastic Aptitude Test coaching effectiveness: A multivariate hierarchical linear model meta-analysis approach. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Kreft, G. G., de Leeuw, J. & van der Leeden, R. (1994). Review of five multilevel analysis programs: BMDP-5V, GENMOD, HLM, ML3, and VARCL. *The American Statistician*, 48(4), 324-335.

Messick, S., & Jungeblut, A. (1981). Time and method in coaching for the SAT. *Psychological Bulletin*, 89, 191-216.

Morris, C. (1983). Parametric empirical Bayes inference: Theory and applications. *Journal of the American Statistical Association*, 78(381), 47-65.

Raudenbush, S. W. (1984). Magnitude of teacher expectancy effects on pupil IQ as a function of the credibility of expectancy induction: A synthesis of findings from 18 experiments. *Journal of Educational Psychology*, 76, 85-97.

Raudenbush, S. W. (1988). Estimating change in dispersion. *Journal of Educational Statistics*, 13(2), 148-171.

Raudenbush, S. W. (1994). Random effects models. In H. C. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 301-321). New York: Russell Sage Foundation.

Raudenbush, S. W., Becker, B. J., & Kalaian, H. (1988). Modeling multivariate effect sizes. *Psychological Bulletin*, 103, 111-120.

Raudenbush, S. W., & Bryk, A. S. (1985). Empirical Bayes meta-analysis. *Journal of Educational Statistics*, 10, 75-98.

Rosenthal, R., & Rubin, D. (1986). Meta-analytic procedures for combining studies with multiple effect sizes. *Psychological Bulletin*, 99, 400-406.

Seltzer, M. (1993). Sensitivity analysis for fixed effects in the hierarchical model: A Gibbs sampling approach. *Journal of Educational and Behavioral Statistics*, 18(3), 207-235.

Appendix

Application of Method With Available Software

Kreft, de Leeuw, and van der Leeden (1994) reviewed currently available software for estimation of hierarchical models. In principle, any of these packages can be used to estimate the model presented here, so long as the user is able to constrain the Level 1 variance to a constant. We used the HLM package of Bryk et al. (1993). That package requires independent Level 1 (within-study) errors. We therefore transformed the multivariate within-study model of Equation 4 so that the within-study errors would be orthogonal with unit variance. We chose the Cholesky factorization, $V_i = F_i F_i^T$. Therefore, the transformed within study model becomes

$$F_i^{-1} d_i = F_i^{-1} S_i X_i \delta_i + F_i^{-1} e_i \tag{A1}$$

$$d_i^* = X_i^* \delta_i + e_i^* \tag{A2}$$

Note that the transformation leaves unchanged the true effect size vector, δ_i . The key benefit is that while $e_i \sim N(0,$

$V_i)$, $e_i^* \sim N(0, I_p)$, where I_p is the identity matrix of dimension p .^{A1}

Having transformed the data, we can make application via the HLM software straightforward. The Level 1 (within-study) data are the outcomes d_i^* and the Level 1 predictors X_i^* . The Level 1 variance is constrained to unity. No changes in the Level 2 data are needed.

Received March 22, 1996
Accepted March 26, 1996 ■

^{A1} The assumption that the Level 1 variance matrix V_i is known, when in fact it is consistently estimated, is standard in meta-analysis that uses effect size data. Sensitivity of results to errors in estimation of V_i is small unless sample sizes per study are exceptionally small and effect sizes exceptionally large (Hedges, 1981).

