PSYCHOMETRIKA—VOL. 70, NO. 1, 147–167
MARCH 2005
DOI: 10.1007/s11336-003-1070-8

ON MUTHÉN'S MAXIMUM LIKELIHOOD FOR TWO-LEVEL COVARIANCE STRUCTURE MODELS

KE-HAI YUAN

UNIVERSITY OF NOTRE DAME

KENTARO HAYASHI

UNIVERSITY OF HAWAII AT MANOA

Data in social and behavioral sciences are often hierarchically organized. Special statistical procedures that take into account the dependence of such observations have been developed. Among procedures for 2-level covariance structure analysis, Muthén's maximum likelihood (MUML) has the advantage of easier computation and faster convergence. When data are balanced, MUML is equivalent to the maximum likelihood procedure. Simulation results in the literature endorse the MUML procedure also for unbalanced data. This paper studies the analytical properties of the MUML procedure in general. The results indicate that the MUML procedure leads to correct model inference asymptotically when level-2 sample size goes to infinity and the coefficient of variation of the level-1 sample sizes goes to zero. The study clearly identifies the impact of level-1 and level-2 sample sizes on the standard errors and test statistic of the MUML procedure. Analytical results explain previous simulation results and will guide the design or data collection for the future applications of MUML.

Key words: asymptotics, likelihood ratio statistic, multilevel covariance structure, standard error estimates

1. Introduction

Data in social and behavioral sciences often exhibit hierarchical structure. For example, households are nested within neighborhoods, neighborhoods are nested within cities, and cities are further nested within countries; students are nested within classes, classes are nested within schools, and schools are further nested within school districts. Cases within a cluster are generally correlated. Analysis of such data has to explicitly account for these correlations. The development of statistical methods for hierarchical data is documented by monographs and edited books (Goldstein, 1995; Heck & Thomas, 2000; Hox, 2002; Kreft & de Leeuw, 1998; Raudenbush & Bryk, 2002; Reise & Duan, 2003; Snijders & Bosker, 1999). Among these are the hierarchical linear model (HLM) and the multilevel structural equation model (SEM) (Bentler & Liang, 2003; du Toit & du Toit, 2004 Goldstein & McDonald, 1988; Lee, 1990; Lee & Poon, 1998; Liang & Bentler, in 2004; Little, Schnabel & Baumert, 2000; Longford, 1993; McArdle & Hamagami, 1996; McDonald & Goldstein, 1989; Muthén, 1994, 1997; Muthén & Satorra, 1995; Poon & Lee, 1994; Yuan & Bentler, 2002, 2003a).

In a multilevel covariance structure model, parameters appear at each level. Parameter estimates can be obtained by maximizing the normal-theory based likelihood function. Iterative procedures are used in this process (Bentler & Liang, 2003; du Toit & du Toit, in press; Goldstein, 1986; Lee & Poon, 1998; Liang & Bentler, 2004; Longford, 1987). As a multilevel SEM typically involves many variables, with unbalanced data many high dimensional matrices have to

This research was supported by NSF Grant DMS04-37167.

We thank Dr. Bengt Muthén for providing key references. We are also grateful to three expert reviewers for their constructive comments that have led the paper to an improvement over the previous version.

Request for reprints should be sent to Ke-Hai Yuan, University of Notre Dame, In 46556, USA. E-mail: kyuan@nd.edu

be inverted at each iteration (see du Toit & du Toit, in press; Liang & Bentler, 2004; McDonald & Goldstein, 1989). When a large number of level-1 units contain small sample sizes, maximizing the likelihood function is not just time consuming but may also put the convergence in jeopardy. Muthén (1989) observed that, for balanced data, a 2-level covariance structure model can be solved by a conventional SEM program with the 2-group option. Muthén (1990) further proposed an ad hoc 2-group procedure to deal with unbalanced data for 2-level covariance structure models. Using real data, Muthén (1990) showed that it yields essentially the same results as the maximum likelihood (ML). In a parallel development with balanced data, McDonald and Goldstein (1989) obtained a compact expression of the likelihood ratio statistic in terms of sufficient statistics (the sample mean and the between and within level sample covariance matrices). Using pseudobalanced sufficient statistics for unbalanced data, McDonald (1994) further illustrated that the McDonald-Goldstein's (1989) balanced data likelihood discrepancy function is equivalent to Muthén's (1990) procedure. Using examples with simulated data, McDonald (1994) also compared the ad hoc and ML procedures. His results indicate that they lead to the same conclusion for model inference. Using simulation and a 2-level factor model, Hox (1993) studied Muthén's ad hoc procedure and found that it recovers the population parameters well when level-1 sample sizes are relatively large. Hox and Maas (2001) further studied standard errors and the test statistic resulting from Muthén's (1990) procedure. They found that the standard error estimates for the between-level parameters are negatively biased; the test statistic for the overall model evaluation tends to over-reject correct models (or positively biased).

Obtaining a set of converged solutions is the prerequisite in evaluating any model. Convergence problems arise with small samples when fitting a conventional covariance structure model (See e.g., Anderson & Gerbing, 1984; Boomsma, 1982; Curran et al., 2002). Convergence is more challenging with 2-level models when many level-1 sample sizes are small and unequal. Due to the computational and convergent advantage of Muthén's(1990) procedure, it has been implemented in popular software (e.g., EQS 6.0, Mplus 2.12) and formally introduced in standard textbooks (e.g., Duncan, Durcan, strycker, Li, & Alpert, 1999; Hox, 2002; Kano & Miura, 2002). This procedure is now commonly called MUML. Simulation or numerical results about MUML are important but have limitations. For example, the biases in standard errors or test statistic of MUML may come from two sources, one is related to sample sizes and gradually disappears as sample sizes increase, the other remains as sample sizes get larger. Analytical results are necessary to distinguish the two sources. Because of the computational advantage of the MUML procedure, it is becoming increasingly popular. A systematic analytical study is needed to better guide its application. We are especially interested in finding conditions under which standard errors and the test statistic in the MUML procedure are statistically valid. We will study the MUML estimators and their standard errors in Section 2, and the test statistic of the MUML procedure in Section 3. Illustrations of biases in standard errors and the test statistic are provided in Section 4, verifying the effect of sample size and model conditions identified in Sections 2 and 3. A brief discussion of correcting the biases in MUML and its extension to non-normal data are given in Section 5. Technical details are provided in two appendices.

2. The Asymptotic Distribution of the MUML Estimator

Let the $p \times 1$ vectors $\mathbf{y}_{ij}$, $i = 1, \ldots, n_j$ be observations from cluster j with $j = 1, \ldots, J$. The 2-level structure of $\mathbf{y}_{ij}$ can be described by

$$\mathbf{y}_{ij} = \boldsymbol{\mu} + \mathbf{v}_j + \mathbf{u}_{ij}, \tag{1}$$

where $\boldsymbol{\mu}$ is a mean vector, $\mathbf{v}_j$ and $\mathbf{u}_{ij}$ are independent with $E(\mathbf{v}_j) = E(\mathbf{u}_{ij}) = \mathbf{0}$, $\mathrm{Cov}(\mathbf{v}_j) = \boldsymbol{\Sigma}_b$ and $\mathrm{Cov}(\mathbf{u}_{ij}) = \boldsymbol{\Sigma}_w$. Let $\boldsymbol{\theta}$ denote the vector of parameters in the structural models $\boldsymbol{\Sigma}_b(\boldsymbol{\theta})$ and

$\boldsymbol{\Sigma}_w(\boldsymbol{\theta})$. Assume $\mathbf{v}_j$ and $\mathbf{u}_{ij}$ following multivariate normal distributions, then so does $\mathbf{y}_{ij}$. Consequently, parameter estimation and model testing can proceed with the normal-theory based ML and the likelihood ratio statistic. With unbalanced data, the ML procedure needs special programs and will take longer to converge. Muthén (1989, 1990) proposed a procedure by which parameter estimation and model testing can proceed with the 2-group option of a conventional SEM program. Let

$$\bar{\mathbf{y}}_{\cdot j} = \frac{1}{n_j}\sum_{i=1}^{n_j}\mathbf{y}_{ij}, \quad \bar{\mathbf{y}} = \frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\mathbf{y}_{ij},$$

$$\mathbf{S}_w = \frac{1}{N-J}\sum_{j=1}^{J}\sum_{i=1}^{n_j}(\mathbf{y}_{ij}-\bar{\mathbf{y}}_{\cdot j})(\mathbf{y}_{ij}-\bar{\mathbf{y}}_{\cdot j})',$$

$$\mathbf{S}_b = \frac{1}{J-1}\sum_{j=1}^{J}n_j(\bar{\mathbf{y}}_{\cdot j}-\bar{\mathbf{y}})(\bar{\mathbf{y}}_{\cdot j}-\bar{\mathbf{y}})',$$

where $N = n_1 + \cdots + n_J$. Muthén's ML[1] estimator $\hat{\boldsymbol{\theta}}$ minimizes

$$\begin{aligned} F_{\mathrm{MUML}}(\boldsymbol{\theta}) = {} & J\{\mathrm{tr}[(\boldsymbol{\Sigma}_w + c\boldsymbol{\Sigma}_b)^{-1}\mathbf{S}_b] - \log|(\boldsymbol{\Sigma}_w + c\boldsymbol{\Sigma}_b)^{-1}\mathbf{S}_b| - p\} \\ & +(N-J)\{\mathrm{tr}(\boldsymbol{\Sigma}_w^{-1}\mathbf{S}_w) - \log|\boldsymbol{\Sigma}_w^{-1}\mathbf{S}_w| - p\}, \end{aligned} \tag{2}$$

where $c = (N^2 - \sum_{j=1}^{J} n_j^2)/[N(J-1)]$. Obviously $E(\mathbf{S}_w) = \boldsymbol{\Sigma}_w$, Muthén (1990) has showed that $E(\mathbf{S}_b) = \boldsymbol{\Sigma}_w + c\boldsymbol{\Sigma}_b$. When data are balanced and the denominator $J-1$ in $\mathbf{S}_b$ is replaced by J, F_{MUML} is -2 multiple of the log likelihood ratio of the structural model and the saturated model. So, MUML is equivalent to the normal-theory ML for balanced data (Muthén, 1990). We are interested in the properties of the MUML estimator $\hat{\boldsymbol{\theta}}$ and the associated test statistic $T_{\mathrm{MUML}} = F_{\mathrm{MUML}}(\hat{\boldsymbol{\theta}})$ when data are unbalanced.

We introduce some notation for the technical development. For a $p \times p$ symmetric matrix $\mathbf{A}$, $\mathrm{vec}(\mathbf{A})$ is the p^2-dimensional vector formed by stacking the columns of $\mathbf{A}$. The $p^* = p(p+1)/2$-dimensional vector $\mathrm{vech}(\mathbf{A})$ is obtained by removing the elements above the diagonal of $\mathbf{A}$ from $\mathrm{vec}(\mathbf{A})$. Consequently, there exists a unique $p^2 \times p^*$ matrix $\mathbf{D}_p$ (see Magnus & Neudecker, 1999, p. 49) such that $\mathrm{vec}(\mathbf{A}) = \mathbf{D}_p\mathrm{vech}(\mathbf{A})$ and $\mathrm{vech}(\mathbf{A}) = \mathbf{D}_p^+\mathrm{vec}(\mathbf{A})$, where $\mathbf{D}_p^+ = (\mathbf{D}_p'\mathbf{D}_p)^{-1}\mathbf{D}_p'$ is the generalized inverse of $\mathbf{D}_p$. We will use $\boldsymbol{\sigma}_b = \mathrm{vech}(\boldsymbol{\Sigma}_b)$, $\boldsymbol{\sigma}_w = \mathrm{vech}(\boldsymbol{\Sigma}_w)$, and $\boldsymbol{\sigma}_c = \mathrm{vech}(\boldsymbol{\Sigma}_c)$, where $\boldsymbol{\Sigma}_c = \boldsymbol{\Sigma}_w + c\boldsymbol{\Sigma}_b$. A function with a dot on top means derivative, e.g., $\dot{\boldsymbol{\sigma}}_w(\boldsymbol{\theta}) = d\boldsymbol{\sigma}_w(\boldsymbol{\theta})/d\boldsymbol{\theta}$. When a function is evaluated at the true value of the parameter, we often omit the argument. Further, $t_n = o_p(a_n)$ means that t_n/a_n approaches zero in probability as n approaches infinity, $t_n = O_p(a_n)$ means that t_n/a_n is bounded in probability, and $\xrightarrow{\mathcal{L}}$ denotes converging in distribution. We will denote by the boldface counterpart $\mathbf{o}_p$ or $\mathbf{O}_p$ when every element in a vector or a matrix is of order o_p or O_p. Parallel nonstochastic notation omits the subscript p.

Let $\boldsymbol{\theta}_w$ be the parameters in $\boldsymbol{\Sigma}_w(\boldsymbol{\theta})$ and the remaining parameters of $\boldsymbol{\theta}$ be $\boldsymbol{\theta}_b$. Yuan and Bentler (2002) argued that the normal-theory MLE $\hat{\boldsymbol{\theta}}_w$ and $\hat{\boldsymbol{\theta}}_b$ converge to their population values at different speeds as characterized by

$$\hat{\boldsymbol{\theta}}_w - \boldsymbol{\theta}_{w0} = \mathbf{O}_p\left(1/\sqrt{N-J}\right) \quad \text{and} \quad \hat{\boldsymbol{\theta}}_b - \boldsymbol{\theta}_{b0} = \mathbf{O}_p(1/\sqrt{J}). \tag{3}$$

[1]Muthén (1997, (9)) has extended MUML to include mean structures.

Using essentially the same argument as in Yuan and Bentler (2002) one can show that the MUML estimator $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_b', \hat{\boldsymbol{\theta}}_w')'$ also satisfies (3). Thus, J has to be large in order for $\hat{\boldsymbol{\theta}}_b$ to be near its population value (see also Hox & Maas, 2002). It is not necessary for n_j to approach infinity in order for $\hat{\boldsymbol{\theta}}_b$ or $\hat{\boldsymbol{\theta}}_w$ to be near their population values in $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_{b0}', \boldsymbol{\theta}_{w0}')'$. So we will implicitly assume $J \to \infty$ when we say $\hat{\boldsymbol{\theta}}$ is consistent or asymptotically normally distributed. We will explicitly mention n_j or its average approaching infinity when needed for special results. Because F_{MUML} is not a likelihood function for unbalanced data, standard result for ML cannot be applied to obtain the property of the MUML estimator $\hat{\boldsymbol{\theta}}$. We will characterize the distribution of $\hat{\boldsymbol{\theta}}$ by the theory of estimating equations (see Liang & Zeger, 1986; Yuan & Jennrich, 1998). In this approach, the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}$ is of a sandwich-type. Our effort below mainly involves obtaining and simplifying this covariance matrix for the MUML estimator.

Denote

$$\mathbf{W}_c = 2^{-1}\mathbf{D}_p'(\boldsymbol{\Sigma}_c^{-1} \otimes \boldsymbol{\Sigma}_c^{-1})\mathbf{D}_p, \quad \mathbf{W}_w = 2^{-1}\mathbf{D}_p'(\boldsymbol{\Sigma}_w^{-1} \otimes \boldsymbol{\Sigma}_w^{-1})\mathbf{D}_p.$$

Taking the derivative of F_{MUML} with respect to $\boldsymbol{\theta}$ we get the estimating function

$$\mathbf{g}(\boldsymbol{\theta}) = \frac{N-J}{J}\dot{\boldsymbol{\sigma}}_w'(\boldsymbol{\theta})\mathbf{W}_w(\boldsymbol{\theta})[\mathbf{s}_w - \boldsymbol{\sigma}_w(\boldsymbol{\theta})] + \dot{\boldsymbol{\sigma}}_c'(\boldsymbol{\theta})\mathbf{W}_c(\boldsymbol{\theta})[\mathbf{s}_b - \boldsymbol{\sigma}_c(\boldsymbol{\theta})]. \tag{4}$$

Because $E[\mathbf{g}(\boldsymbol{\theta}_0)] = \mathbf{0}$, under standard regularity conditions (see Yuan & Jennrich, 1998), the MUML estimator $\hat{\boldsymbol{\theta}}$ satisfies the estimating equation $\mathbf{g}(\hat{\boldsymbol{\theta}}) = \mathbf{0}$ and is consistent for $\boldsymbol{\theta}_0$. So, as an estimator for $\boldsymbol{\theta}_0$, $\hat{\boldsymbol{\theta}}$ does not contain any asymptotic biases. Applying the Taylor expansion on $\mathbf{g}(\hat{\boldsymbol{\theta}}) = \mathbf{0}$, we obtain

$$\begin{aligned}\sqrt{J}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) &= -\dot{\mathbf{g}}^{-1}(\bar{\boldsymbol{\theta}})\sqrt{J}\mathbf{g}(\boldsymbol{\theta}_0) \\ &= -\dot{\mathbf{g}}^{-1}(\boldsymbol{\theta}_0)\sqrt{J}\mathbf{g}(\boldsymbol{\theta}_0) + o_p(1),\end{aligned} \tag{5}$$

where $\bar{\boldsymbol{\theta}}$ is a vector between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$. Hence

$$\sqrt{J}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{L}} N(\mathbf{0}, \boldsymbol{\Omega}), \tag{6}$$

where $\boldsymbol{\Omega} = \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$ with $\mathbf{A} = -E(\dot{\mathbf{g}})$ and $\mathbf{B} = J\text{Var}(\mathbf{g}) = JE(\mathbf{g}\mathbf{g}')$. The matrix $\mathbf{A}$ is the "information matrix" associated with minimizing (2). In the default MUML, the covariance matrix of $\sqrt{J}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ is $\mathbf{A}^{-1}$. Standard errors of $\hat{\boldsymbol{\theta}}$ based on

$$\sqrt{J}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{L}} N(\mathbf{0}, \mathbf{A}^{-1}) \tag{6a}$$

are asymptotically correct when $\mathbf{A} = \mathbf{B}$. To study the conditions under which $\mathbf{A} \approx \mathbf{B}$ in the following, we will obtain the $\mathbf{B}$ matrix and then relate $\boldsymbol{\Omega}$ to $\mathbf{A}^{-1}$ through level-1 and level-2 sample sizes. Conditions will be identified when $\boldsymbol{\Omega} = \mathbf{A}^{-1}$.

Using $E(\mathbf{s}_b) = \boldsymbol{\sigma}_w$ and $E(\mathbf{s}_c) = \boldsymbol{\sigma}_c$ we obtain

$$\mathbf{A} = \left(\frac{N}{J} - 1\right)\dot{\boldsymbol{\sigma}}_w'\mathbf{W}_w\dot{\boldsymbol{\sigma}}_w + \dot{\boldsymbol{\sigma}}_c'\mathbf{W}_c\dot{\boldsymbol{\sigma}}_c. \tag{7}$$

Notice that the $\mathbf{g}$ in (4) involves the random vectors $\mathbf{s}_w$ and $\mathbf{s}_b$, and

$$\begin{aligned}\text{Var}(\mathbf{g}) \;=\; &\left(\tfrac{N}{J} - 1\right)^2 \dot{\boldsymbol{\sigma}}_w'\mathbf{W}_w\text{Var}(\mathbf{s}_w)\mathbf{W}_w\dot{\boldsymbol{\sigma}}_w + \dot{\boldsymbol{\sigma}}_c'\mathbf{W}_c\text{Var}(\mathbf{s}_b)\mathbf{W}_c\dot{\boldsymbol{\sigma}}_c \\ &+ \left(\tfrac{N}{J} - 1\right)\dot{\boldsymbol{\sigma}}_w'\mathbf{W}_w\text{Cov}(\mathbf{s}_w, \mathbf{s}_b)\mathbf{W}_c\dot{\boldsymbol{\sigma}}_c + \left(\tfrac{N}{J} - 1\right)\dot{\boldsymbol{\sigma}}_c'\mathbf{W}_c\text{Cov}(\mathbf{s}_b, \mathbf{s}_w)\mathbf{W}_w\dot{\boldsymbol{\sigma}}_w.\end{aligned} \tag{8}$$

We need to obtain the variance–covariance matrices $\mathrm{Var}(\mathbf{s}_b)$, $\mathrm{Var}(\mathbf{s}_w)$ and $\mathrm{Cov}(\mathbf{s}_b, \mathbf{s}_w)$ before obtaining the $\mathbf{B}$ matrix. Let

$$\mathbf{h}_{j2} = \mathrm{vech}[(\bar{\mathbf{y}}_{.j} - \boldsymbol{\mu})(\bar{\mathbf{y}}_{.j} - \boldsymbol{\mu})'] - \left(\boldsymbol{\sigma}_b + \frac{1}{n_j}\boldsymbol{\sigma}_w\right),$$

$$\mathbf{h}_{j3} = \mathrm{vech}\left[\sum_{i=1}^{n_j}(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{.j})(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{.j})'\right] - (n_j - 1)\boldsymbol{\sigma}_w,$$

and denote $\mathbf{C}_{j22} = E(\mathbf{h}_{j2}\mathbf{h}'_{j2})$, $\mathbf{C}_{j33} = E(\mathbf{h}_{j3}\mathbf{h}'_{j3})$, $\mathbf{C}_{j23} = E(\mathbf{h}_{j2}\mathbf{h}'_{j3})$. The Appendix A provides details leading to

$$\begin{aligned}\mathrm{Var}(\mathbf{s}_b) = {} & \frac{1}{(J-1)^2}\sum_{j=1}^{J}\left(n_j^2 + \frac{n_j^4}{N^2} - \frac{2n_j^3}{N}\right)\mathbf{C}_{j22} \\ & + \frac{2}{N^2(J-1)^2}\mathbf{D}_p^+\left\{\left[\sum_{j=1}^{J} n_j^2\left(\boldsymbol{\Sigma}_b + \frac{1}{n_j}\boldsymbol{\Sigma}_w\right)\right] \otimes \left[\sum_{j=1}^{J} n_j^2\left(\boldsymbol{\Sigma}_b + \frac{1}{n_j}\boldsymbol{\Sigma}_w\right)\right]\right\}\mathbf{D}_p^{+\prime} \\ & - \frac{2}{N^2(J-1)^2}\sum_{j=1}^{J} n_j^4 \mathbf{D}_p^+\left[\left(\boldsymbol{\Sigma}_b + \frac{1}{n_j}\boldsymbol{\Sigma}_w\right) \otimes \left(\boldsymbol{\Sigma}_b + \frac{1}{n_j}\boldsymbol{\Sigma}_w\right)\right]\mathbf{D}_p^{+\prime}, \end{aligned} \tag{9}$$

$$\mathrm{Var}(\mathbf{s}_w) = \frac{1}{(N-J)^2}\sum_{j=1}^{J}\mathbf{C}_{j33}, \tag{10}$$

and

$$\mathrm{Cov}(\mathbf{s}_b, \mathbf{s}_w) = \frac{1}{(J-1)(N-J)}\sum_{j=1}^{J}\left(n_j - \frac{n_j^2}{N}\right)\mathbf{C}_{j23}. \tag{11}$$

We are mainly interested in the property of MUML when data are normal. With normal data,

$$\mathbf{C}_{j22} = 2\mathbf{D}_p^+\left[\left(\boldsymbol{\Sigma}_b + \frac{1}{n_j}\boldsymbol{\Sigma}_w\right) \otimes \left(\boldsymbol{\Sigma}_b + \frac{1}{n_j}\boldsymbol{\Sigma}_w\right)\right]\mathbf{D}_p^{+\prime}, \quad \mathbf{C}_{j23} = \mathbf{0}, \tag{12}$$

and

$$\mathbf{C}_{j33} = 2(n_j - 1)\mathbf{D}_p^+(\boldsymbol{\Sigma}_w \otimes \boldsymbol{\Sigma}_w)\mathbf{D}_p^{+\prime}. \tag{13}$$

Denote $N_2 = \sum_{j=1}^J n_j^2$, $N_3 = \sum_{j=1}^J n_j^3$, $N_4 = \sum_{j=1}^J n_j^4$, it follows from (9) and (12) that

$$\begin{aligned}\mathrm{Var}(\mathbf{s}_b) = {} & \frac{2}{(J-1)^2}\mathbf{D}_p^+\left\{\left(N_2 + \frac{N_4}{N^2} - \frac{2N_3}{N}\right)\boldsymbol{\Sigma}_b \otimes \boldsymbol{\Sigma}_b\right. \\ & \left. + \left(N + \frac{N_3}{N^2} - \frac{2N_2}{N}\right)(\boldsymbol{\Sigma}_b \otimes \boldsymbol{\Sigma}_w + \boldsymbol{\Sigma}_w \otimes \boldsymbol{\Sigma}_b) + \left(J + \frac{N_2}{N^2} - 2\right)\boldsymbol{\Sigma}_w \otimes \boldsymbol{\Sigma}_w\right\}\mathbf{D}_p^{+\prime} \\ & + \frac{2}{N^2(J-1)^2}\mathbf{D}_p^+\left\{N_2^2\boldsymbol{\Sigma}_b \otimes \boldsymbol{\Sigma}_b + NN_2(\boldsymbol{\Sigma}_b \otimes \boldsymbol{\Sigma}_w + \boldsymbol{\Sigma}_w \otimes \boldsymbol{\Sigma}_b) + N^2\boldsymbol{\Sigma}_w \otimes \boldsymbol{\Sigma}_w\right\}\mathbf{D}_p^{+\prime}\end{aligned}$$

$$
-\frac{2}{N^2(J-1)^2}\mathbf{D}_p^+\{N_4\boldsymbol{\Sigma}_b\otimes\boldsymbol{\Sigma}_b+N_3(\boldsymbol{\Sigma}_b\otimes\boldsymbol{\Sigma}_w+\boldsymbol{\Sigma}_w\otimes\boldsymbol{\Sigma}_b)+N_2\boldsymbol{\Sigma}_w\otimes\boldsymbol{\Sigma}_w\}\mathbf{D}_p^{+\prime}
$$

$$
=\frac{2}{(J-1)^2}\mathbf{D}_p^+\left\{\left(N_2+\frac{N_2^2}{N^2}-\frac{2N_3}{N}\right)\boldsymbol{\Sigma}_b\otimes\boldsymbol{\Sigma}_b\right.
$$

$$
\left.+\left(N-\frac{N_2}{N}\right)(\boldsymbol{\Sigma}_b\otimes\boldsymbol{\Sigma}_w+\boldsymbol{\Sigma}_w\otimes\boldsymbol{\Sigma}_b)+(J-1)\boldsymbol{\Sigma}_w\otimes\boldsymbol{\Sigma}_w\right\}\mathbf{D}_p^{+\prime}.
$$

Let $\bar{n}_1=N/J$, $\bar{n}_2=N_2/J$, $\bar{n}_3=N_3/J$. To simplify the expression for matrix $\mathbf{B}=JE(\mathbf{gg}')$, we rewrite $\mathrm{Var}(\mathbf{s}_b)$ as

$$
\mathrm{Var}(\mathbf{s}_b)=\frac{2}{(J-1)^2}\mathbf{D}_p^+\left\{\left(J\bar{n}_2+\frac{\bar{n}_2^2}{\bar{n}_1^2}-\frac{2\bar{n}_3}{\bar{n}_1}\right)\boldsymbol{\Sigma}_b\otimes\boldsymbol{\Sigma}_b\right.
$$

$$
\left.+\left(J\bar{n}_1-\frac{\bar{n}_2}{\bar{n}_1}\right)(\boldsymbol{\Sigma}_b\otimes\boldsymbol{\Sigma}_w+\boldsymbol{\Sigma}_w\otimes\boldsymbol{\Sigma}_b)+(J-1)\boldsymbol{\Sigma}_w\otimes\boldsymbol{\Sigma}_w\right\}\mathbf{D}_p^{+\prime}
$$

$$
=\frac{2}{(J-1)}\mathbf{D}_p^+(\boldsymbol{\Sigma}_c\otimes\boldsymbol{\Sigma}_c)\mathbf{D}_p^{+\prime}+2a\mathbf{D}_p^+(\boldsymbol{\Sigma}_b\otimes\boldsymbol{\Sigma}_b)\mathbf{D}_p^{+\prime}, \tag{14a}
$$

where

$$
a=\frac{1}{(J-1)^3}\left[J^2(\bar{n}_2-\bar{n}_1^2)+J\left(\frac{\bar{n}_2^2}{\bar{n}_1^2}-\frac{2\bar{n}_3}{\bar{n}_1}+\bar{n}_2\right)+2\left(\frac{\bar{n}_3}{\bar{n}_1}-\frac{\bar{n}_2^2}{\bar{n}_1^2}\right)\right]. \tag{14b}
$$

It follows from (10) and (13) that

$$
\mathrm{Var}(\mathbf{s}_w)=\frac{2}{N-J}\mathbf{D}_p^+(\boldsymbol{\Sigma}_w\otimes\boldsymbol{\Sigma}_w)\mathbf{D}_p^{+\prime}. \tag{15}
$$

Equations (11) and (12) imply

$$
\mathrm{Cov}(\mathbf{s}_b,\mathbf{s}_w)=\mathbf{0}. \tag{16}
$$

Note that

$$
\mathbf{W}_c^{-1}=2\mathbf{D}_p^+[(\boldsymbol{\Sigma}_w+c\boldsymbol{\Sigma}_b)\otimes(\boldsymbol{\Sigma}_w+c\boldsymbol{\Sigma}_b)]\mathbf{D}_p^{+\prime}\quad\text{and}\quad\mathbf{W}_w^{-1}=2\mathbf{D}_p^+(\boldsymbol{\Sigma}_w\otimes\boldsymbol{\Sigma}_w)\mathbf{D}_p^{+\prime}.
$$

Combining (8), (14a), (15) and (16) we obtain

$$
\mathbf{B}=\left(\frac{N}{J}-1\right)\dot{\boldsymbol{\sigma}}_w'\mathbf{W}_w\dot{\boldsymbol{\sigma}}_w+\frac{J}{J-1}\dot{\boldsymbol{\sigma}}_c'\mathbf{W}_c\dot{\boldsymbol{\sigma}}_c+Ja\boldsymbol{\Delta}, \tag{17}
$$

where

$$
\boldsymbol{\Delta}=\dot{\boldsymbol{\sigma}}_c'\mathbf{W}_c\mathbf{W}_b^{-1}\mathbf{W}_c\dot{\boldsymbol{\sigma}}_c \tag{18}
$$

with

$$
\mathbf{W}_b=2^{-1}\mathbf{D}_p'(\boldsymbol{\Sigma}_b^{-1}\otimes\boldsymbol{\Sigma}_b^{-1})\mathbf{D}_p.
$$

When data are balanced, $a=0$. It follows from (17) that the matrix $\boldsymbol{\Omega}$ in (6) is given by

$$
\boldsymbol{\Omega}=\mathbf{A}^{-1}\left[\left(\frac{N}{J}-1\right)\dot{\boldsymbol{\sigma}}_w'\mathbf{W}_w\dot{\boldsymbol{\sigma}}_w+\frac{J}{J-1}\dot{\boldsymbol{\sigma}}_c'\mathbf{W}_c\dot{\boldsymbol{\sigma}}_c\right]\mathbf{A}^{-1}. \tag{19a}
$$

When J is large, $\mathbf{\Omega} \approx \mathbf{A}^{-1}$. For unbalanced data we have

$$\mathbf{\Omega} = \mathbf{A}^{-1} + Ja\mathbf{A}^{-1}\mathbf{\Delta}\mathbf{A}^{-1}. \tag{19b}$$

It follows from (19b) that Ja plays an important role in whether $\mathbf{\Omega} \approx \mathbf{A}^{-1}$. Let

$$v_n^2 = \sum_{j=1}^{J} \frac{(n_j - \bar{n}_1)^2}{J} = \bar{n}_2 - \bar{n}_1^2.$$

When J is large and the n_js are bounded, it follows from (14b) that $Ja \approx v_n^2$. For more insight into the magnitude of a, suppose the n_js are uniformly distributed on an interval $[n_a + 1, n_b]$ with $n_b = n_a + n$. Then the variance of the n_js is $v_n^2 = (n^2 - 1)/12$. The Appendix B gives the outline leading to

$$a = \frac{(J-2)}{(J-1)^2} v_n^2 + \frac{(J-2)}{(J-1)^3} \frac{v_n^4}{[n_a + (n+1)/2]^2}. \tag{20}$$

Because $v_n^2 < [n_a + (n+1)/2]^2/3$,

$$\frac{(J-2)}{(J-1)^2} v_n^2 < a < \frac{v_n^2}{(J-1)}.$$

So, regardless of the range of the n_js, $\lim_{J\to\infty} Ja = v_n^2$ when n_js are uniformly distributed.

Notice that

$$c = \frac{N^2 - N_2}{N(J-1)} = \frac{J\bar{n}_1 - \bar{n}_2/\bar{n}_1}{J-1} = \bar{n}_1 - \frac{v_n^2}{\bar{n}_1(J-1)}. \tag{21}$$

So $c \to \infty$ when the average sample size goes to infinity and $v_n^2/(J-1)$ is bounded. Assume the elements of $\mathbf{\Sigma}_w$ are uniformly bounded, we have

$$\mathbf{W}_c = c^{-2}\mathbf{W}_b + \mathbf{O}(c^{-3}) \tag{22}$$

and

$$\mathbf{\Delta} = c^{-2}\dot{\boldsymbol{\sigma}}_b'\mathbf{W}_b\dot{\boldsymbol{\sigma}}_b + \mathbf{O}(c^{-3}).$$

It follows from (7) and (22) that the submatrix of $\mathbf{A}$ corresponding to $\boldsymbol{\theta}_b$ is of order $\mathbf{O}(1)$. Because $Ja = O(v_n^2)$, $\dot{\boldsymbol{\sigma}}_b'\mathbf{W}_b\dot{\boldsymbol{\sigma}}_b = \mathbf{O}(1)$, (19b) implies $\mathbf{\Omega} \to \mathbf{A}^{-1}$ as $J \to \infty$ and $v_n^2/c^2 \to 0$. It follows from (21) that

$$\frac{v_n}{c} = \frac{v_n}{\bar{n}_1\{1 - v_n^2/[\bar{n}_1^2(J-1)]\}} = \frac{v_n}{\bar{n}_1}\left\{1 + O\left(\frac{v_n^2}{[\bar{n}_1^2(J-1)]}\right)\right\} = \mathrm{CV}(n) + O\left[\frac{\mathrm{CV}^3(n)}{(J-1)}\right],$$

where $\mathrm{CV}(n) = v_n/\bar{n}_1$ is the coefficient of variation of the level-1 sample sizes $n_1, \ldots, n_J$. Hence, for large J, $v_n/c \approx \mathrm{CV}(n)$. Consequently, standard errors by MUML are asymptotically correct as $J \to \infty$ and $\mathrm{CV}(n) \to 0$.

Hox and Maas (2001) studied standard errors of $\hat{\boldsymbol{\theta}}$ by simulation. Their results imply that there are little or no biases for the standard errors of $\hat{\boldsymbol{\theta}}_w$ while the standard errors of $\hat{\boldsymbol{\theta}}_b$ possess substantial biases. We next look at the effect of $\mathbf{\Delta}$ on standard errors of $\hat{\boldsymbol{\theta}}_b$ and $\hat{\boldsymbol{\theta}}_w$ separately. For this purpose, we assume that $\mathbf{\Sigma}_b(\boldsymbol{\theta})$ and $\mathbf{\Sigma}_w(\boldsymbol{\theta})$ have separate and unrelated parameters.

That is $\boldsymbol{\theta} = (\boldsymbol{\theta}_b', \boldsymbol{\theta}_w')'$ with $\boldsymbol{\Sigma}_b = \boldsymbol{\Sigma}_b(\boldsymbol{\theta}_b)$ and $\boldsymbol{\Sigma}_w = \boldsymbol{\Sigma}_w(\boldsymbol{\theta}_w)$. Denote $\dot{\boldsymbol{\sigma}}_{bb} = d\boldsymbol{\sigma}_b/d\boldsymbol{\theta}_b$ and $\dot{\boldsymbol{\sigma}}_{ww} = d\boldsymbol{\sigma}_w/d\boldsymbol{\theta}_w$, then $\dot{\boldsymbol{\sigma}}_w = (\mathbf{0}, \dot{\boldsymbol{\sigma}}_{ww})$ and $\dot{\boldsymbol{\sigma}}_c = (c\dot{\boldsymbol{\sigma}}_{bb}, \dot{\boldsymbol{\sigma}}_{ww})$. It follows from (7) and (18) that

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix} = \begin{pmatrix} c^2\dot{\boldsymbol{\sigma}}_{bb}'\mathbf{W}_c\dot{\boldsymbol{\sigma}}_{bb} & c\dot{\boldsymbol{\sigma}}_{bb}'\mathbf{W}_c\dot{\boldsymbol{\sigma}}_{ww} \\ c\dot{\boldsymbol{\sigma}}_{ww}'\mathbf{W}_c\dot{\boldsymbol{\sigma}}_{bb} & \dot{\boldsymbol{\sigma}}_{ww}'\mathbf{W}_c\dot{\boldsymbol{\sigma}}_{ww} + (\bar{n}_1 - 1)\dot{\boldsymbol{\sigma}}_{ww}'\mathbf{W}_w\dot{\boldsymbol{\sigma}}_{ww} \end{pmatrix},$$

and

$$\boldsymbol{\Delta} = \begin{pmatrix} \boldsymbol{\Delta}_{11} & \boldsymbol{\Delta}_{12} \\ \boldsymbol{\Delta}_{21} & \boldsymbol{\Delta}_{22} \end{pmatrix} = \begin{pmatrix} c^2\dot{\boldsymbol{\sigma}}_{bb}'\mathbf{W}_c\mathbf{W}_b^{-1}\mathbf{W}_c\dot{\boldsymbol{\sigma}}_{bb} & c\dot{\boldsymbol{\sigma}}_{bb}'\mathbf{W}_c\mathbf{W}_b^{-1}\mathbf{W}_c\dot{\boldsymbol{\sigma}}_{ww} \\ c\dot{\boldsymbol{\sigma}}_{ww}'\mathbf{W}_c\mathbf{W}_b^{-1}\mathbf{W}_c\dot{\boldsymbol{\sigma}}_{bb} & \dot{\boldsymbol{\sigma}}_{ww}'\mathbf{W}_c\mathbf{W}_b^{-1}\mathbf{W}_c\dot{\boldsymbol{\sigma}}_{ww} \end{pmatrix}.$$

Denote

$$\mathbf{A}^{-1} = \begin{pmatrix} \mathbf{A}^{11} & \mathbf{A}^{12} \\ \mathbf{A}^{21} & \mathbf{A}^{22} \end{pmatrix},$$

standard formula (e.g., Magnus & Neudecker, 1999, p. 11) gives

$$\mathbf{A}^{11} = (\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1}, \quad \mathbf{A}^{12} = -\mathbf{A}^{11}\mathbf{A}_{12}\mathbf{A}_{22}^{-1},$$
$$\mathbf{A}^{21} = -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}^{11}, \quad \mathbf{A}^{22} = (\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1}.$$

Let

$$\mathbf{D} = \mathbf{A}^{-1}\boldsymbol{\Delta}\mathbf{A}^{-1} = \begin{pmatrix} \mathbf{D}_{11} & \mathbf{D}_{12} \\ \mathbf{D}_{21} & \mathbf{D}_{22} \end{pmatrix},$$

then

$$\mathbf{D}_{11} = \mathbf{A}^{11}(\boldsymbol{\Delta}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\boldsymbol{\Delta}_{21} - \boldsymbol{\Delta}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} + \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\boldsymbol{\Delta}_{22}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})\mathbf{A}^{11} \tag{23}$$

and

$$\begin{aligned}\mathbf{D}_{22} &= \mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}^{11}\boldsymbol{\Delta}_{11}\mathbf{A}^{11}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} - \mathbf{A}^{22}\boldsymbol{\Delta}_{21}\mathbf{A}^{11}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ &\quad -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}^{11}\boldsymbol{\Delta}_{12}\mathbf{A}^{22} + \mathbf{A}^{22}\boldsymbol{\Delta}_{22}\mathbf{A}^{22}.\end{aligned} \tag{24}$$

Approximations are needed to simplify (23) and (24). Actually, we only need to identify the leading terms in $\mathbf{D}_{11}$ and $\mathbf{D}_{22}$. It follows from (22) that

$$\mathbf{A}_{22}^{-1} = [\dot{\boldsymbol{\sigma}}_{ww}'\mathbf{W}_c\dot{\boldsymbol{\sigma}}_{ww} + (\bar{n}_1 - 1)\dot{\boldsymbol{\sigma}}_{ww}'\mathbf{W}_w\dot{\boldsymbol{\sigma}}_{ww}]^{-1} = \bar{n}_1^{-1}(\dot{\boldsymbol{\sigma}}_{ww}'\mathbf{W}_w\dot{\boldsymbol{\sigma}}_{ww})^{-1} + \mathbf{O}(\bar{n}_1^{-2}), \tag{25}$$

and

$$\mathbf{A}^{11} = (c^2\dot{\boldsymbol{\sigma}}_{bb}'\mathbf{W}_c\dot{\boldsymbol{\sigma}}_{bb} - c^2\dot{\boldsymbol{\sigma}}_{bb}'\mathbf{W}_c\dot{\boldsymbol{\sigma}}_{ww}\mathbf{A}_{22}^{-1}\dot{\boldsymbol{\sigma}}_{ww}'\mathbf{W}_c\dot{\boldsymbol{\sigma}}_{bb})^{-1} = (\dot{\boldsymbol{\sigma}}_{bb}'\mathbf{W}_b\dot{\boldsymbol{\sigma}}_{bb})^{-1} + \mathbf{O}(c^{-1}). \tag{26}$$

Substituting (25) and (26) in (23), and using (22), we have

$$\begin{aligned}\mathbf{D}_{11} &= c^{-2}(\dot{\boldsymbol{\sigma}}_{bb}'\mathbf{W}_b\dot{\boldsymbol{\sigma}}_{bb})^{-1}\Big[\dot{\boldsymbol{\sigma}}_{bb}'\mathbf{W}_b\dot{\boldsymbol{\sigma}}_{bb} - \frac{2}{\bar{n}_1c^2}\dot{\boldsymbol{\sigma}}_{bb}'\mathbf{W}_b\dot{\boldsymbol{\sigma}}_{ww}(\dot{\boldsymbol{\sigma}}_{ww}'\mathbf{W}_w\dot{\boldsymbol{\sigma}}_{ww})^{-1}\dot{\boldsymbol{\sigma}}_{ww}'\mathbf{W}_b\dot{\boldsymbol{\sigma}}_{bb} \\ &\quad + \frac{1}{\bar{n}_1^2c^4}\dot{\boldsymbol{\sigma}}_{bb}'\mathbf{W}_b\dot{\boldsymbol{\sigma}}_{ww}(\dot{\boldsymbol{\sigma}}_{ww}'\mathbf{W}_w\dot{\boldsymbol{\sigma}}_{ww})^{-1}\dot{\boldsymbol{\sigma}}_{ww}'\mathbf{W}_b\dot{\boldsymbol{\sigma}}_{ww}(\dot{\boldsymbol{\sigma}}_{ww}'\mathbf{W}_w\dot{\boldsymbol{\sigma}}_{ww})^{-1}\dot{\boldsymbol{\sigma}}_{ww}'\mathbf{W}_b\dot{\boldsymbol{\sigma}}_{bb}\Big] \\ &\quad (\dot{\boldsymbol{\sigma}}_{bb}'\mathbf{W}_b\dot{\boldsymbol{\sigma}}_{bb})^{-1} + \mathbf{O}(c^{-3}).\end{aligned} \tag{27}$$

The leading term in (27) is $c^{-2}(\dot{\boldsymbol{\sigma}}'_{bb}\mathbf{W}_b\dot{\boldsymbol{\sigma}}_{bb})^{-1}$. Because $Ja \approx v_n^2$, the variance of $\hat{\boldsymbol{\theta}}_b$ is inflated by approximately $c^{-2}v_n^2(\dot{\boldsymbol{\sigma}}'_{bb}\mathbf{W}_b\dot{\boldsymbol{\sigma}}_{bb})^{-1}$. Using (26), the relative inflation in the standard errors of $\hat{\boldsymbol{\theta}}_b$ is approximately $(1 + v_n^2/c^2)^{1/2} - 1 \approx v_n^2/(2c^2) \approx \mathrm{CV}^2(n)/2$ when $\mathrm{CV}(n)$ is small. The number of groups J will reduce the absolute biases in the standard errors of $\hat{\boldsymbol{\theta}}_b$ but not their relative biases.

Using (22), (24), (25), (26) and

$$\mathbf{A}^{22} = [\dot{\boldsymbol{\sigma}}'_{ww}\mathbf{W}_c\dot{\boldsymbol{\sigma}}_{ww} + (\bar{n}_1 - 1)\dot{\boldsymbol{\sigma}}'_{ww}\mathbf{W}_w\dot{\boldsymbol{\sigma}}_{ww} - \dot{\boldsymbol{\sigma}}'_{ww}\mathbf{W}_c\dot{\boldsymbol{\sigma}}_{bb}(\dot{\boldsymbol{\sigma}}'_{bb}\mathbf{W}_c\dot{\boldsymbol{\sigma}}_{bb})^{-1}\dot{\boldsymbol{\sigma}}'_{bb}\mathbf{W}_c\dot{\boldsymbol{\sigma}}_{ww}]^{-1}$$

$$= \bar{n}_1^{-1}(\dot{\boldsymbol{\sigma}}'_{ww}\mathbf{W}_w\dot{\boldsymbol{\sigma}}_{ww})^{-1} + \mathbf{O}(\bar{n}_1^{-2}),$$

we have

$$\mathbf{D}_{22} = \frac{1}{\bar{n}_1^2 c^4}(\dot{\boldsymbol{\sigma}}'_{ww}\mathbf{W}_w\dot{\boldsymbol{\sigma}}_{ww})^{-1}[\dot{\boldsymbol{\sigma}}'_{ww}\mathbf{W}_b\dot{\boldsymbol{\sigma}}_{ww} - \dot{\boldsymbol{\sigma}}'_{ww}\mathbf{W}_b\dot{\boldsymbol{\sigma}}_{bb}(\dot{\boldsymbol{\sigma}}'_{bb}\mathbf{W}_b\dot{\boldsymbol{\sigma}}_{bb})^{-1}\dot{\boldsymbol{\sigma}}'_{bb}\mathbf{W}_b\dot{\boldsymbol{\sigma}}_{ww}]$$

$$(\dot{\boldsymbol{\sigma}}'_{ww}\mathbf{W}_w\dot{\boldsymbol{\sigma}}_{ww})^{-1} + \mathbf{o}(\bar{n}_1^2 c^4).$$

When

$$\dot{\boldsymbol{\sigma}}'_{bb}\mathbf{W}_b\dot{\boldsymbol{\sigma}}_{bb} = \dot{\boldsymbol{\sigma}}'_{bb}\mathbf{W}_b\dot{\boldsymbol{\sigma}}_{ww} = \dot{\boldsymbol{\sigma}}'_{ww}\mathbf{W}_w\dot{\boldsymbol{\sigma}}_{ww} = \dot{\boldsymbol{\sigma}}'_{ww}\mathbf{W}_b\dot{\boldsymbol{\sigma}}_{ww} \tag{28}$$

holds, the leading term of $\mathbf{D}_{22}$ vanishes. Even when (28) does not hold, the magnitude of $\mathbf{D}_{22}$ is of order $\mathbf{O}(\bar{n}_1^{-6})$, which will be tiny even when $\bar{n}_1$ is small. Because $\mathbf{A}^{22}$ is of order $\mathbf{O}(\bar{n}_1^{-1})$, the relative inflation in the standard errors of $\hat{\boldsymbol{\theta}}_w$ is approximately $\mathrm{CV}^2(n)/\bar{n}_1^3$. This explains why Hox and Maas (2001) found little or no biases in the standard errors of $\hat{\boldsymbol{\theta}}_w$.

3. The Asymptotic Distribution of the MUML Statistic

We will study the distribution of the MUML statistic $T_{\mathrm{MUML}} = F_{\mathrm{MUML}}(\hat{\boldsymbol{\theta}})$. The default MUML procedure refers T_{MUML} to a chi-square distribution. We will analytically compare T_{MUML} to this reference distribution. By decomposing T_{MUML} into several terms, we will show that one term asymptotically follows the reference distribution. The other terms contribute to the discrepancy between T_{MUML} and the reference chi-square distribution. We will also relate the discrepancy to level-1 and level-2 sample sizes and model structures.

Notice that $\hat{\boldsymbol{\theta}}$ is consistent for $\boldsymbol{\theta}_0$ as $J \to \infty$. Using (3) and the appendix of Yuan, Marshall, and Bentler (2002) we obtain

$$T_{\mathrm{MUML}} = J[\mathbf{s}_b - \boldsymbol{\sigma}_c(\hat{\boldsymbol{\theta}})]'\mathbf{W}_c[\mathbf{s}_b - \boldsymbol{\sigma}_c(\hat{\boldsymbol{\theta}})] + (N - J)[\mathbf{s}_w - \boldsymbol{\sigma}_w(\hat{\boldsymbol{\theta}})]'\mathbf{W}_w[\mathbf{s}_w - \boldsymbol{\sigma}_w(\hat{\boldsymbol{\theta}})] + o_p(1). \tag{29}$$

Let

$$\mathbf{W} = \begin{pmatrix} \mathbf{W}_c & \mathbf{0} \\ \mathbf{0} & \left(\frac{N}{J-1}\right)\mathbf{W}_w \end{pmatrix}, \quad \boldsymbol{\xi} = (\boldsymbol{\sigma}'_c, \boldsymbol{\sigma}'_w)', \quad \mathbf{s} = (\mathbf{s}'_b, \mathbf{s}'_w)'.$$

Equation (29) can be rewritten as

$$T_{\mathrm{MUML}} = J[\mathbf{s} - \boldsymbol{\xi}(\hat{\boldsymbol{\theta}})]'\mathbf{W}[\mathbf{s} - \boldsymbol{\xi}(\hat{\boldsymbol{\theta}})] + o_p(1). \tag{30}$$

Similarly, we can rewrite the $\mathbf{g}$ in (4) as $\mathbf{g}(\boldsymbol{\theta}_0) = \dot{\boldsymbol{\xi}}'\mathbf{W}(\mathbf{s} - \boldsymbol{\xi}_0)$ and the $\mathbf{A}$ in (7) as $\mathbf{A} = \dot{\boldsymbol{\xi}}'\mathbf{W}\dot{\boldsymbol{\xi}}$. It follows from (5) that

$$\sqrt{J}[\boldsymbol{\xi}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\xi}(\boldsymbol{\theta}_0)] = \sqrt{J}\dot{\boldsymbol{\xi}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + \mathbf{o}_p(1) = \dot{\boldsymbol{\xi}}(\dot{\boldsymbol{\xi}}'\mathbf{W}\dot{\boldsymbol{\xi}})^{-1}\dot{\boldsymbol{\xi}}'\mathbf{W}\sqrt{J}(\mathbf{s} - \boldsymbol{\xi}_0) + \mathbf{o}_p(1).$$

Thus,

$$\sqrt{J}[\mathbf{s}-\boldsymbol{\xi}(\hat{\boldsymbol{\theta}})] = \{\mathbf{I}-\dot{\boldsymbol{\xi}}(\dot{\boldsymbol{\xi}}'\mathbf{W}\dot{\boldsymbol{\xi}})^{-1}\dot{\boldsymbol{\xi}}'\mathbf{W}\}\sqrt{J}(\mathbf{s}-\boldsymbol{\xi}_0)+\mathbf{o}_p(1). \tag{31}$$

It follows from (30) and (31) that

$$T_{\text{MUML}} = \sqrt{J}[\mathbf{W}^{1/2}(\mathbf{s}-\boldsymbol{\xi}_0)]'\mathbf{Q}\sqrt{J}[\mathbf{W}^{1/2}(\mathbf{s}-\boldsymbol{\xi}_0)]+o_p(1), \tag{32}$$

where

$$\mathbf{Q} = \mathbf{I}-\mathbf{W}^{1/2}\dot{\boldsymbol{\xi}}(\dot{\boldsymbol{\xi}}'\mathbf{W}\dot{\boldsymbol{\xi}})^{-1}\dot{\boldsymbol{\xi}}'\mathbf{W}^{1/2}$$

is a projection matrix.

It follows from (14a), (15) and (16) that

$$\text{Var}(\mathbf{s}-\boldsymbol{\xi}_0) = \begin{pmatrix} (J-1)^{-1}\mathbf{W}_c^{-1} & \mathbf{0} \\ \mathbf{0} & (N-J)^{-1}\mathbf{W}_w^{-1} \end{pmatrix} + a\begin{pmatrix} \mathbf{W}_b^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}. \tag{33}$$

Let $\mathbf{z}_1 \sim N(\mathbf{0}, \mathbf{I}_{2p^*})$, $\mathbf{z}_2 \sim N(\mathbf{0}, \mathbf{I}_{p^*})$ and $\mathbf{z}_1$ and $\mathbf{z}_2$ be independent. Then it follows from (33) and the central limit theorem that

$$\sqrt{J}\mathbf{W}^{1/2}(\mathbf{s}-\boldsymbol{\xi}_0) = \mathbf{z}_1+\sqrt{aJ}\mathbf{W}_{bc}\mathbf{z}_2+o_p(1), \tag{34}$$

where

$$\mathbf{W}_{bc} = \begin{pmatrix} \mathbf{W}_c^{1/2}\mathbf{W}_b^{-1/2} \\ \mathbf{0} \end{pmatrix}.$$

Combining (32) and (34) leads to

$$T_{\text{MUML}} = \mathbf{z}_1'\mathbf{Q}\mathbf{z}_1+2\sqrt{aJ}\mathbf{z}_1'\mathbf{Q}\mathbf{W}_{bc}\mathbf{z}_2+aJ\mathbf{z}_2'\mathbf{W}_{bc}'\mathbf{Q}\mathbf{W}_{bc}\mathbf{z}_2+o_p(1). \tag{35}$$

The first term on the right of (35) follows $\chi^2_{2p^*-q}$ (see Section 1.4 of Muirhead, 1982), where q is the number of free parameters in $\boldsymbol{\theta}$. When data are balanced, $a=0$. Thus, $T_{\text{MUML}} \xrightarrow{\mathcal{L}} \chi^2_{2p^*-q}$ as $J\to\infty$. Notice that $\mathbf{W}_{bc} = \mathbf{O}(1/c)$, $\mathbf{Q} = \mathbf{O}(1)$, and $aJ = O(v_n^2)$. Substituting them in (35) leads to

$$T_{\text{MUML}} = \mathbf{z}_1'\mathbf{Q}\mathbf{z}_1+O_p\left(\frac{v_n}{c}\right)+o_p(1). \tag{35a}$$

It follows from (35a) that $T_{\text{MUML}} \xrightarrow{\mathcal{L}} \chi^2_{2p^*-q}$ when $J\to\infty$ and v_n/c or $\text{CV}(n)\to 0$. When $\text{CV}(n)$ is substantial, T_{MUML} will not behave like a chi-square variate in general even when $J\to\infty$.

Let's look at the difference in the first-order moments of T_{MUML} and $\chi^2_{2p^*-q}$. Because $\mathbf{z}_1$ and $\mathbf{z}_2$ are independent, $E(\mathbf{z}_1'\mathbf{Q}\mathbf{W}_{bc}\mathbf{z}_2)=0$. It is $E(\mathbf{z}_2'\mathbf{W}_{bc}'\mathbf{Q}\mathbf{W}_{bc}\mathbf{z}_2)$ that makes T_{MUML} stochastically greater than $\chi^2_{2p^*-q}$. For a large J,

$$E(\mathbf{z}_2'\mathbf{W}_{bc}'\mathbf{Q}\mathbf{W}_{bc}\mathbf{z}_2) = \text{tr}(\mathbf{Q}\mathbf{M}),$$

where

$$\mathbf{M} = \begin{pmatrix} \mathbf{W}_c^{1/2}\mathbf{W}_b^{-1}\mathbf{W}_c^{1/2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

Let $\mathbf{P} = \mathbf{W}^{1/2}\dot{\boldsymbol{\xi}}(\dot{\boldsymbol{\xi}}'\mathbf{W}\dot{\boldsymbol{\xi}})^{-1}\dot{\boldsymbol{\xi}}'\mathbf{W}^{1/2}$, then

$$\mathrm{tr}(\mathbf{QM}) = \mathrm{tr}(\mathbf{M}) - \mathrm{tr}(\mathbf{PM}). \tag{36}$$

It follows from (22) that

$$\mathrm{tr}(\mathbf{M}) = \mathrm{tr}(\mathbf{W}_b^{-1}\mathbf{W}_c) = c^{-2}p^* + O(c^{-3}), \tag{37}$$

$$\begin{aligned}\dot{\boldsymbol{\xi}}'\mathbf{W}\dot{\boldsymbol{\xi}} &= c^{-2}\dot{\boldsymbol{\sigma}}_w'\mathbf{W}_b\dot{\boldsymbol{\sigma}}_w + c^{-1}\dot{\boldsymbol{\sigma}}_w'\mathbf{W}_b\dot{\boldsymbol{\sigma}}_b + c^{-1}\dot{\boldsymbol{\sigma}}_b'\mathbf{W}_b\dot{\boldsymbol{\sigma}}_w \\ &\quad + \dot{\boldsymbol{\sigma}}_b'\mathbf{W}_b\dot{\boldsymbol{\sigma}}_b + (\bar{n}_1 - 1)\dot{\boldsymbol{\sigma}}_w'\mathbf{W}_w\dot{\boldsymbol{\sigma}}_w + \mathbf{O}(c^{-1}) \\ &= \dot{\boldsymbol{\sigma}}_b'\mathbf{W}_b\dot{\boldsymbol{\sigma}}_b + (\bar{n}_1 - 1)\dot{\boldsymbol{\sigma}}_w'\mathbf{W}_w\dot{\boldsymbol{\sigma}}_w + \mathbf{O}(c^{-1}),\end{aligned}$$

and

$$\begin{aligned}\dot{\boldsymbol{\sigma}}_c'\mathbf{W}_c\mathbf{W}_b^{-1}\mathbf{W}_c\dot{\boldsymbol{\sigma}}_c &= c^{-4}\dot{\boldsymbol{\sigma}}_w'\mathbf{W}_b\dot{\boldsymbol{\sigma}}_w + c^{-3}\dot{\boldsymbol{\sigma}}_b'\mathbf{W}_b\dot{\boldsymbol{\sigma}}_w + c^{-3}\dot{\boldsymbol{\sigma}}_w'\mathbf{W}_b\dot{\boldsymbol{\sigma}}_b + c^{-2}\dot{\boldsymbol{\sigma}}_b'\mathbf{W}_b\dot{\boldsymbol{\sigma}}_b + \mathbf{O}(c^{-3}) \\ &= c^{-2}\dot{\boldsymbol{\sigma}}_b'\mathbf{W}_b\dot{\boldsymbol{\sigma}}_b + \mathbf{O}(c^{-3}).\end{aligned}$$

Thus,

$$\begin{aligned}\mathrm{tr}(\mathbf{PM}) &= \mathrm{tr}[(\dot{\boldsymbol{\xi}}'\mathbf{W}\dot{\boldsymbol{\xi}})^{-1}(\dot{\boldsymbol{\sigma}}_c'\mathbf{W}_c\mathbf{W}_b^{-1}\mathbf{W}_c\dot{\boldsymbol{\sigma}}_c)] \\ &= \mathrm{tr}\{[\dot{\boldsymbol{\sigma}}_b'\mathbf{W}_b\dot{\boldsymbol{\sigma}}_b + (\bar{n}_1 - 1)\dot{\boldsymbol{\sigma}}_w'\mathbf{W}_w\dot{\boldsymbol{\sigma}}_w + \mathbf{O}(c^{-1})]^{-1}[c^{-2}\dot{\boldsymbol{\sigma}}_b'\mathbf{W}_b\dot{\boldsymbol{\sigma}}_b + \mathbf{O}(c^{-3})]\} \\ &= (\bar{n}_1^{-1}c^{-2})\mathrm{tr}[(\dot{\boldsymbol{\sigma}}_w'\mathbf{W}_w\dot{\boldsymbol{\sigma}}_w)^{-1}(\dot{\boldsymbol{\sigma}}_b'\mathbf{W}_b\dot{\boldsymbol{\sigma}}_b)] + o(\bar{n}_1^{-1}c^{-2}).\end{aligned} \tag{38}$$

It follows from (36), (37) and (38) that

$$\mathrm{tr}(\mathbf{QM}) = c^{-2}\{p^* - \bar{n}_1^{-1}\mathrm{tr}[(\dot{\boldsymbol{\sigma}}_w'\mathbf{W}_w\dot{\boldsymbol{\sigma}}_w)^{-1}(\dot{\boldsymbol{\sigma}}_b'\mathbf{W}_b\dot{\boldsymbol{\sigma}}_b)]\} + o(c^{-2}).$$

When $\dot{\boldsymbol{\sigma}}_w'\mathbf{W}_w\dot{\boldsymbol{\sigma}}_w = \dot{\boldsymbol{\sigma}}_b'\mathbf{W}_b\dot{\boldsymbol{\sigma}}_b$, $\mathrm{tr}[(\dot{\boldsymbol{\sigma}}_w'\mathbf{W}_w\dot{\boldsymbol{\sigma}}_w)^{-1}(\dot{\boldsymbol{\sigma}}_b'\mathbf{W}_b\dot{\boldsymbol{\sigma}}_b)] = q$. So T_{MUML} is stochastically greater than $\chi^2_{2p^*-q}$ by a random term whose expectation is

$$\frac{v_n^2}{c^2}(p^* - \bar{n}_1^{-1}q) + o\left(\frac{v_n^2}{c^2}\right).$$

This term goes to zero as v_n^2/c^2 or CV(n) goes to zero. When CV(n) is substantial, $v_n^2p^*/c^2$ also becomes substantial because p^* is often a large number.

When models $\boldsymbol{\Sigma}_b(\boldsymbol{\theta}_b)$ and $\boldsymbol{\Sigma}_w(\boldsymbol{\theta}_w)$ do not share any parameters, we have $\mathrm{tr}(\mathbf{PM}) = c^{-2}q_b + O(c^{-3})$, where q_b is the number of free parameters in $\boldsymbol{\theta}_b$. So T_{MUML} is stochastically greater than $\chi^2_{2p^*-q}$ by a random term whose expectation is

$$\frac{v_n^2}{c^2}(p^* - q_b) + o\left(\frac{v_n^2}{c^2}\right).$$

For the given level-1 and level-2 sample sizes, T_{MUML} behaves more like a chi-square when $\boldsymbol{\Sigma}_b(\boldsymbol{\theta}_b)$ is less restricted.

4. Illustrations

We have shown that the biases in standard errors by MUML are closely related to the coefficient of variation $\mathrm{CV}(n)$ of the level-1 sample sizes. We also showed that the difference between the statistic T_{MUML} and $\chi^2_{2p^*-q}$ is related to $\mathrm{CV}(n)$. The biases in the standard errors and in T_{MUML} disappear when $\mathrm{CV}(n)$ goes to zero. In practice, it is of interest to know when the biases are small enough so that they can be ignored. Although it is easy to come up with a threshold number such that biases below this number are regarded as small, any such attempt would involve some arbitrary or subjective decision. Instead, we use numerical examples to provide empirical information about the biases as $\mathrm{CV}(n)$ changes. For such a purpose, we can only illustrate the biases under a few selected conditions.

The population covariance matrices $\mathbf{\Sigma}_b$ and $\mathbf{\Sigma}_w$ are created by factor models

$$\mathbf{\Sigma}_b = \mathbf{\Lambda}_b\mathbf{\Phi}_b\mathbf{\Lambda}_b' + \Psi_b, \quad \mathbf{\Sigma}_w = \mathbf{\Lambda}_w\mathbf{\Phi}_w\mathbf{\Lambda}_w' + \Psi_w, \tag{39}$$

where

$$\mathbf{\Lambda}_b = \mathbf{\Lambda}_w = \begin{pmatrix} 0.5 & 0.6 & 0.7 & 0.8 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.5 & 0.6 & 0.7 & 0.8 \end{pmatrix}',$$

$$\mathbf{\Phi}_b = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}, \quad \mathbf{\Phi}_w = \begin{pmatrix} 1 & 0.3 \\ 0.3 & 1 \end{pmatrix},$$

$\mathbf{\Psi}_b$ is a diagonal matrix such that $\mathbf{\Sigma}_b$ is a correlation matrix, and $\mathbf{\Psi}_w$ is a diagonal matrix such that all the diagonal elements of $\mathbf{\Sigma}_w$ are equal to 2. Because both standard errors and their biases converge to zero as the level-2 sample size $J \to \infty$, while J has little effect on the relative biases, we arbitrarily choose $J = 100$. To contrast the effect of $\mathrm{CV}(n)$ on the biases, we choose level-1 sample sizes uniformly distributed on the interval $[n_a+1, n_a+n]$ with $n = 20$, and $n_a = 5, 50$. The corresponding $\mathrm{CV}(n)$ for $n_a = 5$ is 0.372 and for $n_a = 50$ is 0.095. When $n_a = 5$, level-1 sample sizes range from $n_j = 6$ to $n_j = 25$, which might represent various sizes of classes when $\mathbf{y}_{ij}$ represents observations for student i from class j. When $n_a = 50$, level-1 sample sizes range from $n_j = 51$ to $n_j = 70$, which is to see the effect of a smaller $\mathrm{CV}(n)$ on standard errors and the test statistic. As we shall see, the biases in standard errors and T_{MUML} are quite small when $n_a = 50$, and consequently when $\mathrm{CV}(n) \leq 0.095$. We use three models to fit the population generated above.

The first model is as specified in (39) and the factor loadings satisfy the constraint $\mathbf{\Lambda}_b = \mathbf{\Lambda}_w$. All the factor variances are fixed at 1.0 for identification purposes. So there are a total of $q = 26$ free parameters in the model and eight of them are shared by the between- and within- levels. Table 1 contains standard errors by MUML ($\mathrm{SD}_{\mathrm{MUML}}$) and standard errors based on the sandwich-type covariance matrix $\mathbf{\Omega}$ ($\mathrm{SD}_{\mathrm{SW}}$). The absolute biases in $\mathrm{SD}_{\mathrm{MUML}}$ are $\mathrm{SD}_{\mathrm{SW}} - \mathrm{SD}_{\mathrm{MUML}}$. The more interesting quantity is the relative bias defined as

$$\mathrm{R}_{\mathrm{SD}} = \frac{(\mathrm{SD}_{\mathrm{SW}} - \mathrm{SD}_{\mathrm{MUML}})}{\mathrm{SD}_{\mathrm{SW}}},$$

which does not depend on J. When $\mathrm{CV}(n) = 0.372$, the left panel of Table 1 implies that R_{SD}s are about 1.2% for the common factor loadings, 5% for the between-level parameters, and 0.3% for the within-level parameters. $E(T_{\mathrm{MUML}})$ is greater than $E(\chi^2_{2p^*-q}) = 2p^* - q$ by 2.586. With $2p^* - q = 46$, the relative bias

$$R_T = \frac{aJ\mathrm{tr}(\mathbf{QM})}{[2p^* - q + aJ\mathrm{tr}(\mathbf{QM})]}$$

TABLE 1.
Standard errors and their biases ($\mathbf{\Lambda}_b = \mathbf{\Lambda}_w$)

	CV(n) = 0.372			CV(n) = 0.095		
	SD_{MUML}	SD_{SW}	R^*_{SD}	SD_{MUML}	SD_{SW}	R_{SD}
λ_{11}	0.046	0.046	0.010	0.024	0.024	0.001
λ_{21}	0.046	0.046	0.011	0.024	0.024	0.001
λ_{31}	0.047	0.047	0.012	0.025	0.025	0.001
λ_{41}	0.048	0.049	0.013	0.026	0.026	0.001
λ_{52}	0.046	0.046	0.010	0.024	0.024	0.001
λ_{62}	0.046	0.046	0.011	0.024	0.024	0.001
λ_{72}	0.047	0.047	0.012	0.025	0.025	0.001
λ_{82}	0.048	0.049	0.013	0.026	0.026	0.001
ϕ_{b12}	0.114	0.121	0.054	0.105	0.106	0.009
ψ_{b11}	0.132	0.139	0.052	0.119	0.120	0.009
ψ_{b22}	0.121	0.127	0.050	0.107	0.108	0.009
ψ_{b33}	0.110	0.115	0.048	0.096	0.097	0.009
ψ_{b44}	0.103	0.108	0.045	0.087	0.088	0.009
ψ_{b55}	0.132	0.139	0.052	0.119	0.120	0.009
ψ_{b66}	0.121	0.127	0.050	0.107	0.108	0.009
ψ_{b77}	0.110	0.115	0.048	0.096	0.097	0.009
ψ_{b88}	0.103	0.108	0.045	0.087	0.088	0.009
ϕ_{w12}	0.047	0.047	0.000	0.023	0.023	0.000
ψ_{w11}	0.073	0.073	0.001	0.036	0.036	0.000
ψ_{w22}	0.074	0.074	0.002	0.037	0.037	0.000
ψ_{w33}	0.076	0.076	0.003	0.039	0.039	0.000
ψ_{w44}	0.081	0.081	0.005	0.042	0.042	0.000
ψ_{w55}	0.073	0.073	0.001	0.036	0.036	0.000
ψ_{w66}	0.074	0.074	0.002	0.037	0.037	0.000
ψ_{w77}	0.076	0.076	0.003	0.039	0.039	0.000
ψ_{w88}	0.081	0.081	0.005	0.042	0.042	0.000

Note: $^*R_{SD} = (SD_{SW} - SD_{MUML})/SD_{SW}$

is about 5%. The right panel of Table 1 contains standard errors and their biases when CV(n) = 0.095. The R_{SD}s for the between-level parameters are only about 0.9%, which should be considered ignorable in practice because they might be smaller than sampling errors (see Curran, 1994; Yuan & Bentler, 1997). With $aJ\mathrm{tr}(\mathbf{QM}) = 0.219$, the relative bias in T_{MUML} is only about 0.5%.

The second model is also as specified in (39) but there is no constraint on the between- and within-level factor loadings. There are $q = 34$ free parameters in this model. Standard errors and their biases are in Table 2. Even when CV(n) = 0.372, the R_{SD}s at the within-level are zero down to the third decimal place. However, standard errors for the between-level parameters still contain about 5% biases, which are comparable to those for the between-level parameters in Table 1. With 38 degrees of freedom and $aJ\mathrm{tr}(\mathbf{QM}) = 1.898$, the R_T is about 5%, also comparable to that for the first model. When CV(n) = 0.095, the R_{SD}s at the between-level are about 0.9%, again comparable to those in Table 1 while the relative bias in T_{MUML} is only about 0.4%.

When $\mathbf{\Sigma}_b(\boldsymbol{\theta}_b)$ and $\mathbf{\Sigma}_w(\boldsymbol{\theta}_w)$ have no overlap in parameters, results in the previous section indicate that there is little bias in T_{MUML} when $p^* = q_b$. In the third model, we choose a saturated between-level model while the within-level model is specified as in (39). So there are only 19 degrees of freedom in this model. As expected, R_T is zero down to the tenth decimal place even when CV(n) = 0.372. Standard errors and their biases are in Table 3; we only report the

TABLE 2.
Standard errors and their biases $\boldsymbol{\Sigma}_b(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}_w(\boldsymbol{\theta})$ have no overlaping parameters

	$\mathrm{CV}(n) = 0.372$			$\mathrm{CV}(n) = 0.095$		
	$\mathrm{SD}_{\mathrm{MUML}}$	$\mathrm{SD}_{\mathrm{SW}}$	R_{SD}	$\mathrm{SD}_{\mathrm{MUML}}$	$\mathrm{SD}_{\mathrm{SW}}$	R_{SD}
λ_{b11}	0.117	0.124	0.055	0.108	0.109	0.009
λ_{b21}	0.114	0.121	0.055	0.105	0.106	0.009
λ_{b31}	0.112	0.119	0.054	0.103	0.104	0.009
λ_{b41}	0.111	0.117	0.054	0.101	0.102	0.009
λ_{b52}	0.117	0.124	0.055	0.108	0.109	0.009
λ_{b62}	0.114	0.121	0.055	0.105	0.106	0.009
λ_{b72}	0.112	0.119	0.054	0.103	0.104	0.009
λ_{b82}	0.111	0.117	0.054	0.101	0.102	0.009
ϕ_{b12}	0.116	0.123	0.055	0.107	0.108	0.009
ψ_{b11}	0.134	0.142	0.052	0.121	0.122	0.009
ψ_{b22}	0.125	0.132	0.051	0.112	0.113	0.009
ψ_{b33}	0.119	0.125	0.049	0.105	0.106	0.009
ψ_{b44}	0.119	0.125	0.048	0.104	0.105	0.009
ψ_{b55}	0.134	0.142	0.052	0.121	0.122	0.009
ψ_{b66}	0.125	0.132	0.051	0.112	0.113	0.009
ψ_{b77}	0.119	0.125	0.049	0.105	0.106	0.009
ψ_{b88}	0.119	0.125	0.048	0.104	0.105	0.009
λ_{w11}	0.050	0.050	0.000	0.025	0.025	0.000
λ_{w21}	0.051	0.051	0.000	0.025	0.025	0.000
λ_{w31}	0.053	0.053	0.000	0.026	0.026	0.000
λ_{w41}	0.055	0.055	0.000	0.027	0.027	0.000
λ_{w52}	0.050	0.050	0.000	0.025	0.025	0.000
λ_{w62}	0.051	0.051	0.000	0.025	0.025	0.000
λ_{w72}	0.053	0.053	0.000	0.026	0.026	0.000
λ_{w82}	0.055	0.055	0.000	0.027	0.027	0.000
ϕ_{w12}	0.047	0.047	0.000	0.023	0.023	0.000
ψ_{w11}	0.074	0.074	0.000	0.037	0.037	0.000
ψ_{w22}	0.075	0.075	0.000	0.037	0.037	0.000
ψ_{w33}	0.079	0.079	0.000	0.039	0.039	0.000
ψ_{w44}	0.086	0.086	0.000	0.042	0.042	0.000
ψ_{w55}	0.074	0.074	0.000	0.037	0.037	0.000
ψ_{w66}	0.075	0.075	0.000	0.037	0.037	0.000
ψ_{w77}	0.079	0.079	0.000	0.039	0.039	0.000
ψ_{w88}	0.086	0.086	0.000	0.042	0.042	0.000

SDs for the first two columns of parameters in $\boldsymbol{\Sigma}_b$ to save space. There are zero biases for the within-level standard errors up to the third decimal place even when $\mathrm{CV}(n) = 0.372$. The R_{SD}s at the between-level are about 5.4% when $\mathrm{CV}(n) = 0.372$ and 0.9% when $\mathrm{CV}(n) = 0.095$. These are comparable to those in the previous Tables.

In these illustrations, we considered only the expected biases in $\mathrm{SD}_{\mathrm{MUML}}$ with a large J. In practice, both $\mathrm{SD}_{\mathrm{SW}}$ and $\mathrm{SD}_{\mathrm{MUML}}$ also contain sampling errors because they have to be estimated by data with finite sample sizes. The overall relative bias in $\mathrm{SD}_{\mathrm{MUML}}$ at the between-level is about 5% when $\mathrm{CV}(n) = 0.372$, and will be smaller for a smaller $\mathrm{CV}(n)$. Whether 5% of errors are ignorable is subject to judgment. Sampling errors in the commonly used standard error estimates for conventional structural equation models might be greater than 5% under unfavorable conditions (see Yuan & Bentler, 1997).

TABLE 3.
Standard errors and their biases ($\mathbf{\Sigma}_b(\boldsymbol{\theta})$ is saturated)

	CV(n) = 0.372			CV(n) = 0.095		
	SD_{MUML}	SD_{SW}	R_{SD}	SD_{MUML}	SD_{SW}	R_{SD}
σ_{b11}	0.160	0.169	0.054	0.146	0.147	0.009
σ_{b21}	0.117	0.124	0.054	0.108	0.109	0.009
σ_{b31}	0.119	0.126	0.054	0.109	0.110	0.009
σ_{b41}	0.121	0.128	0.055	0.111	0.112	0.009
σ_{b51}	0.114	0.120	0.054	0.104	0.105	0.009
σ_{b61}	0.114	0.120	0.054	0.104	0.105	0.009
σ_{b71}	0.114	0.121	0.054	0.105	0.106	0.009
σ_{b81}	0.115	0.121	0.054	0.105	0.106	0.009
σ_{b22}	0.160	0.169	0.054	0.146	0.147	0.009
σ_{b32}	0.121	0.128	0.055	0.112	0.113	0.009
σ_{b42}	0.124	0.131	0.055	0.114	0.115	0.009
σ_{b52}	0.114	0.120	0.054	0.104	0.105	0.009
σ_{b62}	0.114	0.121	0.054	0.105	0.106	0.009
σ_{b72}	0.115	0.121	0.054	0.105	0.106	0.009
σ_{b82}	0.116	0.122	0.054	0.106	0.107	0.009
λ_{w11}	0.050	0.050	0.000	0.025	0.025	0.000
λ_{w21}	0.051	0.051	0.000	0.025	0.025	0.000
λ_{w31}	0.053	0.053	0.000	0.026	0.026	0.000
λ_{w41}	0.055	0.055	0.000	0.027	0.027	0.000
λ_{w52}	0.050	0.050	0.000	0.025	0.025	0.000
λ_{w62}	0.051	0.051	0.000	0.025	0.025	0.000
λ_{w72}	0.053	0.053	0.000	0.026	0.026	0.000
λ_{w82}	0.055	0.055	0.000	0.027	0.027	0.000
ϕ_{w12}	0.047	0.047	0.000	0.023	0.023	0.000
ψ_{w11}	0.074	0.074	0.000	0.037	0.037	0.000
ψ_{w22}	0.075	0.075	0.000	0.037	0.037	0.000
ψ_{w33}	0.079	0.079	0.000	0.039	0.039	0.000
ψ_{w44}	0.086	0.086	0.000	0.042	0.042	0.000
ψ_{w55}	0.074	0.074	0.000	0.037	0.037	0.000
ψ_{w66}	0.075	0.075	0.000	0.037	0.037	0.000
ψ_{w77}	0.079	0.079	0.000	0.039	0.039	0.000
ψ_{w88}	0.086	0.086	0.000	0.042	0.042	0.000

5. Discussion

Muthén's (1990) maximum likelihood procedure has the advantage of easier calculation and faster convergence than the normal-theory based ML procedure. Consequently, the MUML procedure is becoming increasingly popular and has been implemented in SEM software and introduced in textbooks. This paper studies the analytical statistical properties of the MUML procedure. The results indicate that the last term in (14a) or (33) is responsible for the biases in standard errors of $\hat{\boldsymbol{\theta}}$ and in the test statistic T_{MUML}. When this term vanishes, inference with MUML contains no asymptotic biases. Because this term is generally positive, standard errors based on $\mathbf{\Omega} = \mathbf{A}^{-1}$ in (6a) is underestimated and the test statistic T_{MUML} is stochastically greater than the reference $\chi^2_{2p^*-q}$. This is comparable to the normal-theory based methodology for the conventional covariance structure analysis, where the sample covariance matrix $\mathbf{S}$ is fitted to a structure model $\mathbf{\Sigma}(\boldsymbol{\theta})$. Let $\mathbf{s} = \mathrm{vech}(\mathbf{S})$ and $\mathbf{\Gamma}_0 = \mathrm{Var}(\mathbf{s})$. When $\mathbf{\Gamma}_0 - 2\mathbf{D}_p^+(\mathbf{\Sigma}\otimes\mathbf{\Sigma})\mathbf{D}_p^{+\prime}$ is positive definite, the normal-theory standard errors have negative biases and the normal-theory likelihood ratio statistic has a positive bias (Browne, 1984).

Our results identify the role of the level-1 and level-2 sample sizes in the MUML procedure. The level-2 sample size J is responsible for $\hat{\boldsymbol{\theta}}$ to be near $\boldsymbol{\theta}_0$ and the stability of the distribution of T_{MUML}. The average level-1 sample size $\bar{n}_1$ and the standard deviation v_n are responsible for the validity of statistical inference by MUML. When the coefficient of variation $\mathrm{CV}(n) = v_n/\bar{n}_1$ is small, MUML leads to valid inferences. When $\mathrm{CV}(n)$ is substantial, standard errors of $\hat{\boldsymbol{\theta}}_b$ and the overall model evaluation based on T_{MUML} are biased. This suggests that, if feasible, one should try to avoid level-1 units with small n_js in the data collection process, because small n_js not only contribute to a small $\bar{n}_1$ but also to a large v_n.

The results in Sections 2 and 3 can be used to correct the biases in MUML. When adding the term $Ja\mathbf{A}^{-1}\boldsymbol{\Delta}\mathbf{A}^{-1}$ to $\mathbf{A}^{-1}$ in estimating the covariance matrix of $\sqrt{J}\hat{\boldsymbol{\theta}}$, according to (19b), the corresponding standard error estimates become consistent for unbalanced data. Similarly, one may correct the bias in T_{MUML}. Let $\boldsymbol{\Gamma} = \mathrm{Var}(\mathbf{s} - \boldsymbol{\xi}_0)$ as given in (33) and

$$\mathbf{U} = \mathbf{W}^{1/2}\mathbf{Q}\mathbf{W}^{1/2} = \mathbf{W} - \mathbf{W}\dot{\boldsymbol{\xi}}(\dot{\boldsymbol{\xi}}'\mathbf{W}\dot{\boldsymbol{\xi}})^{-1}\dot{\boldsymbol{\xi}}'\mathbf{W}.$$

Parallel to the corrected statistic of Satorra and Bentler (1994), we can get a corrected statistic

$$T_{\mathrm{CMUML}} = \frac{(2p^* - q)T_{\mathrm{MUML}}}{\mathrm{tr}(\hat{\mathbf{U}}\hat{\boldsymbol{\Gamma}})}.$$

It follows from (32) that the asymptotic distribution of T_{CMUML} has a mean of $2p^* - q$. Although T_{CMUML} does not asymptotically follow the reference distribution $\chi^2_{2p^*-q}$ in general, it may behave like the Satorra and Bentler's (1994) rescaled statistic that works well in practice. Further study is needed to actually evaluate the performance of T_{CMUML}. Note that all the elements for evaluating $\boldsymbol{\Delta}$, $\mathbf{U}$ and $\boldsymbol{\Gamma}$ have already been computed in obtaining $\hat{\boldsymbol{\theta}}$ and its standard errors. Including these corrections by a software that already contains the MUML procedure is straightforward. The above corrections can be extended to non-normal data. Then the matrix $\mathbf{B}$ in (6) and $\boldsymbol{\Gamma} = \mathrm{Var}(\mathbf{s} - \boldsymbol{\xi}_0)$ need to be estimated using the fourth-order moments of the observed data. The details are complicated and can be pursued parallel to the development in Yuan and Bentler (2002, 2003a).

Note that the paper only studies the biases in MUML when models at both level-1 and level-2 are correctly specified. Because $E(\mathbf{S}_b) = \boldsymbol{\Sigma}_w + c\boldsymbol{\Sigma}_b$, both $\hat{\boldsymbol{\theta}}_b$ and $\hat{\boldsymbol{\theta}}_w$ depend on the data matrix $\mathbf{S}_b$ when they are obtained by minimizing the $F_{\mathrm{MUML}}(\boldsymbol{\theta})$ in (2). When a model at one level is misspecified, both parameter estimates $\hat{\boldsymbol{\theta}}_b$ and $\hat{\boldsymbol{\theta}}_w$ may contain biases even for balanced data (see Yuan et al., 2003). Of course, biases due to model misspecification apply not only to the MUML procedure but also the ML procedure. To avoid the biases in $\hat{\boldsymbol{\theta}}_w$ caused by a misspecified $\boldsymbol{\Sigma}_b(\boldsymbol{\theta})$, one may just fit $\mathbf{S}_w$ by $\boldsymbol{\Sigma}_w(\boldsymbol{\theta})$ using the conventional SEM software with sample size $N - J$. Within the MUML setup, $\mathbf{S}_b$ also contains the information of $\boldsymbol{\Sigma}_w$, it is not clear how to avoid the biases in $\hat{\boldsymbol{\theta}}_b$ caused by a misspecified $\boldsymbol{\Sigma}_w(\boldsymbol{\theta})$. Within the context of maximum likelihood, Yuan and Bentler (2003b) proposed a stepwise approach to avoid possible biases of parameter estimates at one level caused by misspecifications at different levels as well as other related complications of model evaluation.

Finally, the purpose of MUML was to provide a procedure for analyzing hierarchical data using the conventional SEM software (see Muthén, 1990). For unbalanced data, MUML is computationally easier than ML. Recently, several important developments were made towards overcoming the computation hurdles with the ML procedure for unbalanced data (du Toit & du Toit, in press; Liang & Bentler, 2004). The demand for MUML may be lessened due to these developments.

Appendix A

This appendix provides the expressions for $\mathrm{Var}(\mathbf{s}_b)$, $\mathrm{Var}(\mathbf{s}_w)$ and $\mathrm{Cov}(\mathbf{s}_b, \mathbf{s}_w)$ as given in (9), (10) and (11). After some algebraic operation, we have

$$\mathbf{s}_b = \frac{1}{J-1}\sum_{j=1}^{J} n_j \mathbf{t}_{jj} - \frac{1}{N(J-1)}\sum_{i=1}^{J}\sum_{j=1}^{J} n_i n_j \mathbf{t}_{ij}, \tag{A1}$$

where $\mathbf{t}_{jj} = \mathrm{vech}[(\bar{\mathbf{y}}_{.j} - \boldsymbol{\mu})(\bar{\mathbf{y}}_{.j} - \boldsymbol{\mu})']$ and $\mathbf{t}_{ij} = \mathrm{vech}[(\bar{\mathbf{y}}_{.i} - \boldsymbol{\mu})(\bar{\mathbf{y}}_{.j} - \boldsymbol{\mu})']$. Because $\mathrm{vech}(\mathbf{A}) = \mathbf{D}_p^+\mathrm{vec}(\mathbf{A})$, and for vectors $\mathbf{a}$ and $\mathbf{b}$ there exists $\mathrm{vec}(\mathbf{ab}') = \mathbf{b}\otimes\mathbf{a}$,

$$\begin{aligned}\mathbf{t}_{ij}\mathbf{t}_{kl}' &= \mathbf{D}_p^+\mathrm{vec}[(\bar{\mathbf{y}}_{.i} - \boldsymbol{\mu})(\bar{\mathbf{y}}_{.j} - \boldsymbol{\mu})']\mathrm{vec}'[(\bar{\mathbf{y}}_{.k} - \boldsymbol{\mu})(\bar{\mathbf{y}}_{.l} - \boldsymbol{\mu})']\mathbf{D}_p^{+\prime}\\ &= \mathbf{D}_p^+\{[(\bar{\mathbf{y}}_{.j} - \boldsymbol{\mu})\otimes(\bar{\mathbf{y}}_{.i} - \boldsymbol{\mu})][(\bar{\mathbf{y}}_{.l} - \boldsymbol{\mu})\otimes(\bar{\mathbf{y}}_{.k} - \boldsymbol{\mu})]'\mathbf{D}_p^{+\prime}\}\\ &= \mathbf{D}_p^+\{[(\bar{\mathbf{y}}_{.j} - \boldsymbol{\mu})(\bar{\mathbf{y}}_{.l} - \boldsymbol{\mu})']\otimes[(\bar{\mathbf{y}}_{.i} - \boldsymbol{\mu})(\bar{\mathbf{y}}_{.k} - \boldsymbol{\mu})']\mathbf{D}_p^{+\prime}\}.\end{aligned} \tag{A2}$$

When $i = j = k = l$, denote

$$E(\mathbf{t}_{jj}\mathbf{t}_{jj}') = \mathbf{C}_{jjjj} \quad \text{and} \quad \mathrm{Var}(\mathbf{t}_{jj}) = \mathbf{C}_{j22}. \tag{A3a}$$

When $i = j$ and $k = l$ but $i \neq k$, it is easy to see

$$E(\mathbf{t}_{jj}\mathbf{t}_{kk}') = \left(\boldsymbol{\sigma}_b + \frac{1}{n_j}\boldsymbol{\sigma}_w\right)\left(\boldsymbol{\sigma}_b + \frac{1}{n_k}\boldsymbol{\sigma}_w\right)', \tag{A3b}$$

When $i = k$ and $j = l$ but $i \neq j$, it follows from (A2) that

$$E(\mathbf{t}_{kj}\mathbf{t}_{kj}') = \mathbf{D}_p^+\left[\left(\boldsymbol{\Sigma}_b + \frac{1}{n_j}\boldsymbol{\Sigma}_w\right)\otimes\left(\boldsymbol{\Sigma}_b + \frac{1}{n_k}\boldsymbol{\Sigma}_w\right)\right]\mathbf{D}_p^{+\prime}. \tag{A3c}$$

Let $\mathbf{K}_p$ be the commutation matrix (see Magnus & Neudecker, 1999, p. 47) such that $\mathbf{K}_p\mathrm{vec}(\mathbf{A}) = \mathrm{vec}(\mathbf{A}')$. It follows from (A2) that

$$\begin{aligned}\mathbf{t}_{ij}\mathbf{t}_{kl}' &= \mathbf{D}_p^+\mathbf{K}_p\mathrm{vec}[(\bar{\mathbf{y}}_{.j} - \boldsymbol{\mu})(\bar{\mathbf{y}}_{.i} - \boldsymbol{\mu})']\mathrm{vec}'[(\bar{\mathbf{y}}_{.k} - \boldsymbol{\mu})(\bar{\mathbf{y}}_{.l} - \boldsymbol{\mu})']\mathbf{D}_p^{+\prime}\\ &= \mathbf{D}_p^+\mathbf{K}_p\{[(\bar{\mathbf{y}}_{.i} - \boldsymbol{\mu})(\bar{\mathbf{y}}_{.l} - \boldsymbol{\mu})']\otimes[(\bar{\mathbf{y}}_{.j} - \boldsymbol{\mu})(\bar{\mathbf{y}}_{.k} - \boldsymbol{\mu})']\}\mathbf{D}_p^{+\prime}.\end{aligned} \tag{A4}$$

When $i = l$ and $j = k$ but $i \neq j$, it follows from (A4) that

$$\begin{aligned}E(\mathbf{t}_{ij}\mathbf{t}_{ji}') &= \mathbf{D}_p^+\mathbf{K}_p\left[\left(\boldsymbol{\Sigma}_b + \frac{1}{n_i}\boldsymbol{\Sigma}_w\right)\otimes(\boldsymbol{\Sigma}_b + \frac{1}{n_j}\boldsymbol{\Sigma}_w)\right]\mathbf{D}_p^{+\prime}\\ &= \mathbf{D}_p^+\left[\left(\boldsymbol{\Sigma}_b + \frac{1}{n_j}\boldsymbol{\Sigma}_w\right)\otimes\left(\boldsymbol{\Sigma}_b + \frac{1}{n_i}\boldsymbol{\Sigma}_w\right)\right]\mathbf{D}_p^{+\prime}.\end{aligned} \tag{A3d}$$

Let

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j,\\ 0 & \text{if } i \neq j,\end{cases}$$

$$\mathbf{h}_{j2} = \mathbf{h}_{jj} = \mathbf{t}_{jj} - \left(\boldsymbol{\sigma}_b + \frac{1}{n_j}\boldsymbol{\sigma}_w\right), \quad \text{and} \quad \mathbf{h}_{ij} = \mathbf{t}_{ij} - \delta_{ij}\left(\boldsymbol{\sigma}_b + \frac{1}{n_j}\boldsymbol{\sigma}_w\right). \tag{A5}$$

Then it is easy to see $E(\mathbf{h}_{jj}) = E(\mathbf{h}_{ij}) = \mathbf{0}$. It follows from (A1)

$$\begin{aligned}\mathrm{Var}(\mathbf{s}_b) = &\frac{1}{(J-1)^2}\sum_{j=1}^{J}\sum_{k=1}^{J} n_j n_k E(\mathbf{h}_{jj}\mathbf{h}'_{kk})\\ &+\frac{1}{N^2(J-1)^2}\sum_{i=1}^{J}\sum_{j=1}^{J}\sum_{k=1}^{J}\sum_{l=1}^{J} n_i n_j n_k n_l E(\mathbf{h}_{ij}\mathbf{h}'_{kl})\\ &-\frac{1}{N(J-1)^2}\sum_{i=1}^{J}\sum_{j=1}^{J}\sum_{k=1}^{J} n_i n_j n_k E(\mathbf{h}_{ii}\mathbf{h}'_{jk})\\ &-\frac{1}{N(J-1)^2}\sum_{i=1}^{J}\sum_{j=1}^{J}\sum_{k=1}^{J} n_i n_j n_k E(\mathbf{h}_{ij}\mathbf{h}'_{kk}).\end{aligned} \tag{A6}$$

When $j \neq k$, $E(\mathbf{h}_{jj}\mathbf{h}'_{kk}) = \mathbf{0}$,

$$\sum_{j=1}^{J}\sum_{k=1}^{J} n_j n_k E(\mathbf{h}_{jj}\mathbf{h}'_{kk}) = \sum_{j=1}^{J} n_j^2 E(\mathbf{h}_{jj}\mathbf{h}'_{jj}) = \sum_{j=1}^{J} n_j^2 \mathbf{C}_{j22}. \tag{A7a}$$

Note that

$$\sum_{i=1}^{J}\sum_{j=1}^{J}\sum_{k=1}^{J}\sum_{l=1}^{J} = \left(\sum_{i=j=k=l} + \sum_{i=j,k=l,i\neq k} + \sum_{i=k,j=l,i\neq j} + \sum_{i=l,j=k,i\neq j} + \sum_{\text{else}}\right),$$

it follows from (A3a) to (A3d) and (A5),

$$\begin{aligned}\sum_{i=1}^{J}\sum_{j=1}^{J}\sum_{k=1}^{J}\sum_{l=1}^{J} n_i n_j n_k n_l E(\mathbf{h}_{ij}\mathbf{h}'_{kl}) = &\sum_{j=1}^{J} n_j^4 \mathbf{C}_{j22} + 2\mathbf{D}_p^{+}\left\{\left[\sum_{j=1}^{J} n_j^2\left(\boldsymbol{\Sigma}_b + \frac{1}{n_j}\boldsymbol{\Sigma}_w\right)\right]\right.\\ &\left.\otimes\left[\sum_{j=1}^{J} n_j^2\left(\boldsymbol{\Sigma}_b + \frac{1}{n_j}\boldsymbol{\Sigma}_w\right)\right]\right\}\mathbf{D}_p^{+\prime}\\ &-2\sum_{j=1}^{J} n_j^4 \mathbf{D}_p^{+}\left[\left(\boldsymbol{\Sigma}_b + \frac{1}{n_j}\boldsymbol{\Sigma}_w\right)\otimes\left(\boldsymbol{\Sigma}_b + \frac{1}{n_j}\boldsymbol{\Sigma}_w\right)\right]\mathbf{D}_p^{+\prime}.\end{aligned} \tag{A7b}$$

It follows from (A3a), (A3b) and (A5)

$$\sum_{i=1}^{J}\sum_{j=1}^{J}\sum_{k=1}^{J} n_i n_j n_k E(\mathbf{h}_{ii}\mathbf{h}'_{jk}) = \left(\sum_{i=j=k} + \sum_{i\neq j=k} + \sum_{\text{else}}\right) n_i n_j n_k E\left(\mathbf{h}_{ii}\mathbf{h}'_{jk}\right) = \sum_{j=1}^{J} n_j^3 \mathbf{C}_{j22}, \tag{A7c}$$

and

$$\sum_{i=1}^{J}\sum_{j=1}^{J}\sum_{k=1}^{J} n_i n_j n_k E(\mathbf{h}_{ij}\mathbf{h}'_{kk}) = \sum_{j=1}^{J} n_j^3 \mathbf{C}_{j22}. \tag{A7d}$$

By combining (A6) and (A7) we obtain

$$\begin{aligned}\mathrm{Var}(\mathbf{s}_b) &= \frac{1}{(J-1)^2}\sum_{j=1}^{J}\left(n_j^2+\frac{n_j^4}{N^2}-\frac{2n_j^3}{N}\right)\mathbf{C}_{j22}\\ &+\frac{2}{N^2(J-1)^2}\mathbf{D}_p^+\left\{\left[\sum_{j=1}^{J}n_j^2\left(\boldsymbol{\Sigma}_b+\frac{1}{n_j}\boldsymbol{\Sigma}_w\right)\right]\otimes\left[\sum_{j=1}^{J}n_j^2\left(\boldsymbol{\Sigma}_b+\frac{1}{n_j}\boldsymbol{\Sigma}_w\right)\right]\right\}\mathbf{D}_p^{+\prime}\\ &-\frac{2}{N^2(J-1)^2}\sum_{j=1}^{J}n_j^4\mathbf{D}_p^+\left[\left(\boldsymbol{\Sigma}_b+\frac{1}{n_j}\boldsymbol{\Sigma}_w\right)\otimes\left(\boldsymbol{\Sigma}_b+\frac{1}{n_j}\boldsymbol{\Sigma}_w\right)\right]\mathbf{D}_p^{+\prime}. \end{aligned}\tag{A8}$$

Let

$$\mathbf{h}_{j3}=\mathrm{vech}\left[\sum_{i=1}^{n_j}(\mathbf{y}_{ij}-\bar{\mathbf{y}}_{.j})(\mathbf{y}_{ij}-\bar{\mathbf{y}}_{.j})'\right]-(n_j-1)\boldsymbol{\sigma}_w,$$

and $E(\mathbf{h}_{j3}\mathbf{h}_{j3}')=\mathbf{C}_{j33}$. Then

$$\begin{aligned}\mathrm{Var}(\mathbf{s}_w) &= \frac{1}{(N-J)^2}\sum_{j=1}^{J}\sum_{k=1}^{J}E(\mathbf{h}_{j3}\mathbf{h}_{k3}')=\frac{1}{(N-J)^2}\left(\sum_{j=k}+\sum_{j\neq k}\right)E(\mathbf{h}_{j3}\mathbf{h}_{k3}')\\ &= \frac{1}{(N-J)^2}\sum_{j=1}^{J}\mathbf{C}_{j33}. \end{aligned}\tag{A9}$$

Let $E(\mathbf{h}_{jj}\mathbf{h}_{j3}')=\mathbf{C}_{j23}$. Then

$$\begin{aligned}\mathrm{Cov}(\mathbf{s}_b,\mathbf{s}_w) &= \frac{1}{(J-1)(N-J)}\sum_{j=1}^{J}\sum_{k=1}^{J}n_jE(\mathbf{h}_{jj}\mathbf{h}_{k3}')\\ &\quad-\frac{1}{N(J-1)(N-J)}\sum_{i=1}^{J}\sum_{j=1}^{J}\sum_{k=1}^{J}n_in_jE(\mathbf{h}_{ij}\mathbf{h}_{k3}')\\ &= \frac{1}{(J-1)(N-J)}\left(\sum_{j=k}+\sum_{j\neq k}\right)n_jE(\mathbf{h}_{jj}\mathbf{h}_{k3}')\\ &\quad-\frac{1}{N(J-1)(N-J)}\left(\sum_{i=j=k}+\sum_{\text{else}}\right)n_in_jE(\mathbf{h}_{ij}\mathbf{h}_{k3}')\\ &= \frac{1}{(J-1)(N-J)}\sum_{j=1}^{J}n_j\mathbf{C}_{j23}-\frac{1}{N(J-1)(N-J)}\sum_{j=1}^{J}n_j^2\mathbf{C}_{j23}\\ &= \frac{1}{(J-1)(N-J)}\sum_{j=1}^{J}\left(n_j-\frac{n_j^2}{N}\right)\mathbf{C}_{j23}. \end{aligned}\tag{A10}$$

Appendix B

This appendix outlines the steps leading to (20). When $n_j \sim U[n_a+1, n_a+n]$, then

$$\bar{n}_1 = n_a + \frac{1}{n}\sum_{j=1}^{n} j = n_a + \frac{n+1}{2}, \tag{B1}$$

$$\bar{n}_2 = \frac{1}{n}\sum_{j=1}^{n}(n_a+j)^2 = \frac{1}{n}\sum_{j=1}^{n}(n_a^2 + 2n_a j + j^2) = n_a^2 + n_a(n+1) + \frac{(n+1)(2n+1)}{6}, \tag{B2}$$

$$\begin{aligned}\bar{n}_3 &= \frac{1}{n}\sum_{j=1}^{n}(n_a+j)^3 = \frac{1}{n}\sum_{j=1}^{n}(n_a^3 + 3n_a^2 j + 3n_a j^2 + j^3)\\ &= n_a^3 + \frac{3n_a^2(n+1)}{2} + \frac{n_a(n+1)(2n+1)}{2} + \frac{n(n+1)^2}{4}.\end{aligned} \tag{B3}$$

Combining (14b) and (B1) to (B3) leads to the a in (20).

References

Anderson, J.C., & Gerbing, D.W. (1984). The effects of sampling error on convergence, improper solutions and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika, 49*, 155–173.

Bentler, P.M., & Liang, J. (2003). Two-level mean and covariance structures: Maximum likelihood via an EM algorithm. In S. Reise & N. Duan (Eds.), *Multilevel Modeling: Methodological Advances, Issues, and Applications* (pp. 53–70). Mahwah, NJ: Erlbaum.

Boomsma, A. (1982). The robustness of LISREL against small sample sizes in factor analysis models. In K.G. Jöreskog & H. Wold (Eds.), *Systems Under Indirect Observation: Causality, Structure, Prediction* (Part I, pp. 149–173). Amsterdam: North-Holland.

Browne, M.W. (1984). Asymptotic distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 37*, 62–83.

Curran, P.J. (1994). *The Robustness of Confirmatory Factor Analysis to Model Misspecification and Violations of Normality*. Ph.D. thesis, Arizona State University.

Curran, P.J., Bollen, K.A., Paxton, P., Kirby, J., & Chen, F. (2002). The noncentral chi-square distribution in misspecified structural equation models: Finite sample results from a Monte Carlo simulation. *Multivariate Behavioral Research, 37*, 1–36.

Duncan, T.E., Duncan, S.C., Strycker, L.A., Li, F., & Alpert, A. (1999). *An Introduction to Latent Variable Growth Curve Modeling: Concepts, Issues, and Applications*. Mahwah, NJ: Erlbaum.

du Toit, S. & M. du Toit (in press) Multilevel structural equation modeling. In J. de Leeuw & I. Kreft (Eds.), *Handbook of Quantitative Multilevel Analysis*. New York: Kluwer.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika, 73*, 43–56.

Goldstein, H. (1995). *Multilevel Statistical Models* (2nd edn). London: Edward Arnold.

Goldstein, H., & McDonald, R.P. (1988). A general model for the analysis of multilevel data. *Psychometrika, 53*, 435–467.

Heck, R.H., & Thomas, S.L. (2000). *An Introduction of Multilevel Modeling Techniques*. Mahwah, NJ: Erlbaum.

Hox, J.J. (1993). Factor analysis of multilevel data: Gauging the Muthén model. In J.H.L. Oud & R.A.W. van Blokland-Vogelesang (Eds.), *Advances in Longitudinal and Multivariate Analysis in the Behavioral Sciences* (pp. 141–156). Nijmegen: ITS.

Hox, J.J. (2002). *Multilevel Analysis: Techniques and Applications*. Mahwah, NJ: Erlbaum.

Hox, J.J., & Maas, C.J.M. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Structural Equation Modeling, 8*, 157–174.

Hox, J.J., & Maas, C.J.M. (2002). Sample sizes for multilevel modeling. In: J. Blasius, J. Hox, E. de Leeuw, & P. Schmidt (Eds.), *Social science methodology in the new millennium. Proceedings of the Fifth International Conference on Logic and Methodology* (2nd expanded edn). Opladen, RG: Leske + Budrich Verlag (CD-ROM).

Kano, Y., & Miura, A. (2002). *Graphical multivariate analysis with AMOS, EQS, and CALIS: A visual approach to covariance structure analysis* (revised edn). Kyoto, Japan: Gendai-Sugakusha.

Kreft, I., & de Leeuw, J. (1998). *Introducing multilevel modeling*. London: Sage.

Lee, S.-Y. (1990). Multilevel analysis of structural equation models. *Biometrika, 77*, 763–772.

Lee, S.-Y., & Poon, W.-Y. (1998). Analysis of two-level structural equation models via EM type algorithms. *Statistica Sinica, 8*, 749–766.

Liang, J., & Bentler, P. M. (2004). An EM algorithm for fitting two-level structural equation models. *Psychometrika*, *69*, 101–122

Liang, K.Y., & Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73*, 13–22.

Little, T.D., Schnabel, K.U., & Baumert, J. (Eds.) (2000). *Modeling Longitudinal and Multilevel Data: Practical Issues, Applied Approaches and Specific Examples*. Mahwah, NJ: Erlbaum.

Longford, N.T. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika, 74*, 817–827.

Longford, N.T. (1993). Regression analysis of multilevel data with measurement error. *British Journal of Mathematical and Statistical Psychology, 46*, 301–311.

Magnus, J.R., & Neudecker, H. (1999). *Matrix Differential Calculus with Applications in Statistics and Econometrics* (revised edn). New York: Wiley.

McArdle, J.J., & Hamagami, F. (1996). Multilevel models from a multiple group structural equation perspective. In G.A. Marcoulides & R.E. Schumacker (Eds.), *Advanced Structural Equation Modeling Techniques* (pp. 89–124). Mahwah, NJ: Erlbaum

McDonald, R.P., & Goldstein, H. (1989). Balanced versus unbalanced designs for linear structural relations in two-level data. *British Journal of Mathematical and Statistical Psychology*, 42, 215–232.

McDonald, R.P. (1994). The bilevel reticular action model for path analysis with latent variables. *Sociological Methods and Research, 22*, 399–413.

Muirhead, R.J. (1982). *Aspects of Multivariate Statistical Theory*. New York: Wiley.

Muthén, B. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54*, 557–585.

Muthén, B. (1990). Mean and covariance structure analysis of hierarchical data. Paper presented at the Psychometric Society meeting in Princeton, NJ, June 1990. UCLA Statistics Series 62.

Muthén, B. (1994). Multilevel covariance structure analysis. *Sociological Methods and Research, 22*, 376–398.

Muthén, B. (1997). Latent variable modeling of longitudinal and multilevel data. In A. Raftery (Ed.), *Sociological Methodology 1997* (pp. 453–480). Boston: Blackwell Publishers.

Muthén, B., & Satorra, A. (1995). Complex sample data in structural equation modeling. In P.V. Marsden (Ed.), *Sociological Methodology 1995* (pp. 267–316). Cambridge, MA: Blackwell Publishers.

Poon, W.-Y., & Lee, S.-Y. (1994). A distribution free approach for analysis of two-level structural equation model. *Computational Statistics and Data Analysis, 17*, 265–275.

Raudenbush, S.W., & Bryk, A.S. (2002). *Hierarchical linear models* (2nd edn). Newbury Park: Sage.

Reise, S., & Duan, N. (Eds.) (2003). *Multilevel Modeling: Methodological Advances, Issues, and Applications*. Mahwah, NJ: Erlbaum.

Satorra, A. & Bentler, P.M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C.C. Clogg (Eds.), *Latent Variables Analysis: Applications for Developmental Research* (pp. 399–419). Thousand Oaks, CA: Sage.

Snijders, T., & Bosker, R. (1999). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. Thousand Oaks, CA: Sage.

Yuan, K.-H. & Bentler, P.M. (1997). Improving parameter tests in covariance structure analysis. *Computational Statistics and Data Analysis*, *26*, 177–198.

Yuan, K.-H., & Bentler, P. M. (2002). On normal theory based inference for multilevel models with distributional violations. *Psychometrika, 67*, 539–561.

Yuan, K.-H., & Bentler, P.M. (2003a). Eight test statistics for multilevel structural equation models. *Computational Statistics and Data Analysis, 44*, 89–107.

Yuan, K.-H., & Bentler, P.M. (2003b). Stepwise analysis of multilevel covariance structure models (under review).

Yuan, K.-H., & Jennrich, R.I. (1998). Asymptotics of estimating equations under natural conditions. *Journal of Multivariate Analysis, 65*, 245–260.

Yuan, K.-H., Marshall, L.L., & Bentler, P.M. (2002). A unified approach to exploratory factor analysis with missing data, nonnormal data, and in the presence of outliers. *Psychometrika, 67*, 95–122.

Yuan, K.-H., Marshall, L.L., & Bentler, P.M. (2003). Assessing the effect of model misspecifications on parameter estimates in structural equation models. In R.M. Stolzenberg (Ed.), *Sociological Methodology 2003* (pp. 241–265). Oxford: Blackwell Publishing.

Manuscript received 19 MAR 2003
Final version received 10 OCT 2003