Title

Multilevel Factor Analysis by Model Segregation: Comparing the Performance of Maximum Likelihood and Robust Test Statistics

Author

Schweig, Jonathan David

Publication Date

2014

Peer reviewed|Thesis/dissertation

University of California

Los Angeles

Multilevel Factor Analysis by Model Segregation: Comparing the Performance of Maximum Likelihood and Robust Test Statistics

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Science in Statistics

by

Jonathan David Schweig

2014

Abstract of the Thesis

Multilevel Factor Analysis by Model Segregation: Comparing the Performance of Maximum Likelihood and Robust Test Statistics

by

Jonathan David Schweig

Master of Science in Statistics

University of California, Los Angeles, 2014

Professor Peter Bentler, Chair

Survey measures of classroom climate and instructional practice have become central to policy efforts that assess school and teacher quality. Measures of classroom climate are often formed by aggregating individual survey responses. This has sparked a wide interest in using multilevel factor analysis to test hypotheses about the psychometric properties of classroom climate variables. One approach to multilevel factor analysis is conducted in two steps. First, the total covariance matrix is partitioned into separate between-group and within-group covariance matrices. Second, conventional factor analysis is used to test models separately. This study shows that when using this approach, rescaled and residual-based test statistics provide better inferences about the between group-level measurement structure than Maximum Likelihood test statistics even when the number of groups is large and there is no excess kurtosis in the observed variables. This study presents an empirical example and a simulation study to demonstrate how item intraclass correlations and within group sample sizes influence test statistic performance.

The thesis of Jonathan David Schweig is approved.

Rick Paik Schoenberg

Mahtash Esfandiari

Felipe Martinez

Peter Bentler, Committee Chair

University of California, Los Angeles

2014

Table of Contents

List of Tables

Acknowledgments

I gratefully acknowledge the support and guidance of Peter Bentler, who provided tremendous support and guidance throughout this process. I would also like to acknowledge Felipe Martinez, whose input greatly improved the clarity, organization, and precision of my language and thinking.

I am also grateful to Joan Herman, Jia Wang, and Noelle Griffin at the National Center for Research on Evaluation, Standards and Student Testing for their support.

I would also like to give thanks to Jenn for her patience and support, and to Fiona, for her uncanny ability to debug code.

CHAPTER 1

Introduction

Survey-based measures of classroom quality have become a staple of many teacher performance portfolios. Seventeen states and many local education agencies including Chicago and Memphis, Tennessee include student surveys as measures of teacher quality or professional practice (Doherty, & Jacobs, 2013). Measures of teacher quality and professional practice are constructed based on aggregated student survey responses. There is an increased attention in applied literature toward using measurement models that account for the hierarchical structure of these surveys, and the fact that individual students are associated with specific classrooms. There is a long tradition of literature (e.g., Cronbach, 1976; Harnqvist, 1978; Julian, 2001; Longford & Muthén, 1992; Zyphur, Kaplan & Christian, 2008) suggesting that single-level analytic methods that do not account for hierarchical data structures are problematic and can be substantively misleading (Reise, Ventura, Nuechterlein & Kim, 2005, p. 130).

Multilevel factor analysis (e.g.,Goldstein, 2003; Lee, 1990; Longford & Muthén, 1992; McDonald & Goldstein, 1989; Muthén, 1991; Muthén, 1994; Rabe-Hesketh, Skrondal & Zheng, 2007) provides a method to analyze multivariate data that is hierarchically structured. One widely used framework (Muthén, 1994) partitions the total covariance matrix into independent between-groups (or group level) and within-groups (or individual level) covariance matrices. As in conventional single level factor analysis, it is often of interest to researchers to test measurement hypotheses in multilevel factor analysis by using test statistics. There are sev-

eral different approaches that can be used to assess the adequacy of measurement models (e.g., Hox, 2010; Ryu & West, 2009) in multilevel factor analysis. These include simultaneously modeling both within-groups and between groups covariance structures, saturating (i.e. estimating all item covariances) the model at one level and fitting a factor model at the other level, and segregating the between and within covariance matrices and conducing factor analysis one level at a time. In conventional factor analysis, the commonly used Maximum-Likelihood (ML) test statistic is derived under the assumption that the observed data is continuous and multivariate normal (e.g. Bollen, 1989). Asymptotically, when this assumption holds, the ML test statistic will be appropriately distributed and inferences drawn from the model will be valid. In fact, it has been shown that normal theory estimators generally remain consistent and test statistics are correctly distributed unless kurtosis in the observed variables is excessive (Browne, 1984; Muthén & Kaplan, 1985, 1992).

Because the segregating method proceeds by conducting two conventional factor analyses, it is often assumed that if sample sizes are sufficiently large, there is no excess kurtosis and the measurement model is correctly specified, inferences about the between-group covariance structure based on the ML test statistic will be valid (e.g. Goldstein, 2003; Hox & Maas, 2004; Ryu & West, 2009). However, there are situations where this is not the case. While the statistical basis for this phenomenon has been developed elsewhere (Yuan & Bentler, 2002, 2006, 2007), the poor performance of ML test statistic is not widely known, and is rarely mentioned in the multilevel factor analysis literature. In fact, the poor performance of the ML test statistic is frequently mischaracterized as evidence of model misspecification in applied literature (Kaplan & Elliott, 1997; Mathiesen, Torsheim & Einarsen, 2006; van Horn, 2003).

This paper is organized as follows. First, the multilevel factor analysis framework is briefly described, along with the rationale for model testing at the between level.

Second, the three major approaches to testing multilevel factor models are summarized. Third, four test statistics are presented, including the conventional Maximum Likelihood (ML) test statistic, the Satorra-Bentler (1988) rescaled test statistic, Browne's (1982,1984) residual-based test statistic, and Yuan and Bentler's (1998) adjusted residual-based test statistic. Fourth, an empirical example from a classroom environment survey illustrates how these statistics may influence inferences about the measurement model in multilevel contexts. Finally, a simulation study is presented to demonstrate the specific conditions under which test statistics may yield valid inferences. The final section discusses the implications of the results for the use of the segregating method to investigate the between-classroom factorial structure of classroom environment surveys, and other surveys that have a group or a cluster as the primary unit of analysis.

CHAPTER 2

Theoretical Background

2.1 Multilevel Factor Analysis

The multilevel factor analysis framework used in this study (e.g. Goldstein, 2003; Lee, 1990; Longford & Muthén, 1992; McDonald & Goldstein, 1989; Muthén, 1991; Muthén, 1994) is based on a score decomposition model articulated by Cronbach and Webb (1975):

$$y_{ij} = y_j + (y_{ij} - y_j) \tag{2.1}$$

where the vector of p observed scores for individual i in group j (y_{ij}) can be decomposed into independent between groups (y_j), and within groups ($y_{ij} - y_j$) components. The associated covariance matrix of the observed scores can be expressed:

$$\Sigma_T = \Sigma_B + \Sigma_W \tag{2.2}$$

where Σ_T, Σ_B and Σ_B are symmetric $p \times p$ covariance matrices. The covariance matrices can be expressed in two separate factor models (e.g., Bollen, 1989), one for the between-groups level:

$$\Sigma_B = \Lambda_B \Phi_B \Lambda_B^T + \Psi_B \tag{2.3}$$

and another for the within-groups level

$$\Sigma_W = \Lambda_W \Phi_W \Lambda_W^T + \Psi_W \tag{2.4}$$

where Λ_B is a $p \times k$ matrix of factor loadings for p items on k factors, and Λ_W is a $p \times r$ matrix of factor loadings for p items on r factors. Note that while it is possible for $k = r$ and for $\Lambda_B = \Lambda_W$, this is not necessary. Φ_B and Φ_W are $k \times k$ and $r \times r$ matrices of factor covariances, respectively, and Ψ_B and Ψ_W are $p \times p$ diagonal matrices containing unique (residual) variances. It follows that Φ_B need not equal Φ_W, and Ψ_B need not equal Ψ_W.

2.2 The rationale for between-level model testing with student surveys

Surveys of classroom environments often assume a specific measurement model where students are are treated as objective raters of the classrooms in which they study (e.g., Follman, 1992; Ferguson, 2012; Worrell & Kuterbach, 2001). Variance between students within the same classroom is attributable to sampling error and represents noise. Averaging over individual students, variance between classrooms represents true variance in classroom quality. In this way, these surveys are often designed to measure climate variables (Marsh et al, 2012), and the primary unit of analysis is the classroom. Accordingly, understanding the between-classroom factor structure is critical for developing and testing theories about how the classroom climate relates to other variables of substantive interest, such as student achievement and persistence in school. There is a long tradition of research suggesting that multilevel factor analysis is the appropriate tool for testing the between-level measurement models (Cronbach, 1976; Harnqvist, 1978; Julian, 2001; Longford & Muthén, 1992; Marsh et al, 2012; Reise, Ventura, & Nuechterlein & Kim, 2005;

Zyphur, Kaplan & Christian, 2008).

2.3 Three approaches to multilevel fit testing

Though multilevel factor analysis provides a framework to test between-classroom measurement models, there is little consensus on the best approach to evaluate models within that framework. There are three primary approaches described in the methodological literature on multilevel factor analysis. a) Simultaneously modeling the within-level and between-level structures (Muthén, 1994) b) Fitting an unrestricted (saturated) model at the within level, and testing a measurement model at the between level (Muthén, 1994; Hox, 2010; Ryu & West, 2009), referred to as the partially saturated model method (Ryu & West, 2009, p. 589) and c) Segregating the between and within covariance matrices and conducting separate factor analyses, referred to as the segregating method (Ryu & West, 2009, p. 592).

It has been shown in several studies (e.g. Hox, 2010; Ryu & West, 2009; Yuan & Bentler, 2007) that simultaneously modeling the within and between level structures does not produce meaningful diagnostic information about the between-level factor structure. Thus, the simultaneous modeling of between and within factor structures makes model or theory revision difficult (Yuan & Bentler, 2007), and this approach is not recommended in the literature. The partially saturated model method, on the other hand, does provide level-specific diagnostic information, but was not meant to provide parameter estimates or standard errors (Ryu & West, p. 599; Yuan & Bentler, 2007). A practical issue with this method is that estimates of fit indices such as the Root Mean Square Error of Approximation (RMSEA) (Steiger & Lind, 1980), and the Comparative Fit Index (CFI) (Bentler, 1990) provided by software programs will spuriously show good fit (Hox, 2010, p. 307), and so may be misinterpreted (e.g. Rosenberg, 2009).

The segregating method (Yuan & Bentler, 2007), which is the focus of this paper, is operationalized in two steps. First, the total covariance matrix is partitioned and Maximum-Likelihood estimates (MLEs) of Σ_B and Σ_W in Equation (2.2) are obtained. For balanced data, the MLEs of these two matrices are unbiased estimates of the population matrices, even when the data is not normally distributed (Muthén, 1994). Once the matrices have been separated, conventional single-level factor analyses can be conducted. Similar approaches are described in Goldstein (2003, p. 189) and Hox (2010). This approach potentially allows for a wide variety of test statistics and fit indices (Yuan & Bentler, 2007), since the model testing proceeds as two separate conventional single level analyses. It also allows for parameter estimates and standard errors (Ryu & West, 2009, p. 599) to be obtained.

Because the segregating method is a two-step procedure, parameter estimates may be less efficient than those obtained from the partially saturated model method (Goldstein, 2003). However, Yuan and Bentler (2007) suggested that, in small to medium sized samples, particularly with larger models, the segregating method may actually be more efficient than the partially saturated model method, because parameter estimates based on a smaller model will have more numerical stability (the segregating method will, in general, have far fewer parameters than partially saturated model method) (Yuan & Bentler, 2007, p. 56).

2.4 Four test statistics

Test statistics used in conjunction with the segregating method can be considered from a conventional, single-level framework, since it is operationalized by performing a series of conventional factor analyses. Before defining the test statistics used in this analysis, some general notation will be presented. Given a symmetric matrix A, let $vech(A)$ be the row half-vectorization of A. If the dimension of

A is $p \times p$, it has $p^* = \frac{(p+1)p}{2}$ unique elements, and $vech(A)$ is a $p^* \times 1$ vector. The matrix D_p is a $p^2 \times p^*$ duplication matrix (e.g., Magnus & Neudecker, 1988). Additionally, let a function with a dot on top denote a derivative. For example, let $\dot{\sigma}(\theta)$ denote the derivative of $\sigma(\theta)$ with respect to θ. For a total sample size N, let $n = N - 1$.

Given a $p \times p$ population covariance matrix Σ, a q-vector of free parameters θ, a testable null hypothesis can be expressed $\Sigma(\theta) = \Sigma$. In other words, the population covariance matrix, Σ, can be expressed as a function of the model parameters, θ (Bollen, 1989). This null hypothesis can be tested using a test statistic obtained from minimizing a discrepancy function, $F[S, \Sigma(\theta)]$, which indicates the discrepancy between the sample covariance matrix, S, and the model-implied covariance matrix $\Sigma(\theta)$. Optimal estimates of model parameters, $\hat{\theta}$, are found at the minimum of F.

Bentler and Dudgeon (1996) note that all discrepancy functions are associated with a weight matrix, W, and an asymptotic covariance matrix Γ, which is given by the distribution of $\sqrt{n}(s - \sigma(\theta))$:

$$\sqrt{n}(s - \sigma(\theta)) \xrightarrow{d} N(0, \Gamma) \tag{2.5}$$

Where $s = vech(S)$ and $\sigma(\theta) = vech(\Sigma(\theta))$. Γ is a symmetric positive definite $p^* \times p^*$ matrix.

Following Browne (1984) (see also Bentler & Dudgeon, 1996; Foldnes, Foss & Olsson, 2012), a discrepancy function is correctly specified for W if

$$W \xrightarrow{p} \Gamma^{-1} \tag{2.6}$$

When the model is correct and the discrepancy function is correctly specified:

$$n\hat{F} \xrightarrow{d} \chi^2_d \tag{2.7}$$

Where $\hat{F} = F[S, \Sigma(\hat{\theta})]$, the minimized value of the discrepancy function. The degrees of freedom, d, is given by $d = p^* - q$ (e.g., Bollen, 1989).

2.4.1 The Maximum Likelihood test statistic $\mathbf{T_{ML}}$

The Maximum Likelihood (ML) discrepancy function (Jöreskog, 1967) is derived from the normal-theory log-likelihood (e.g. Yuan & Bentler, 1999). Optimal estimates of model parameters, $\hat{\theta}$, are found by minimizing

$$F_{ML} = log|\Sigma(\theta)| + tr[S\Sigma(\theta)^{-1}] - log|S| - p \tag{2.8}$$

where $|\cdot|$ denotes the determinant, and tr denotes the trace of a matrix. In conventional factor analysis, S is the typical sample covariance matrix. In using the segregating method to investigate between-level covariance structre, S is given by $\hat{\Sigma}_B$. F_{ML} can be understood as asymptotically equivalent to a special member of a class of generalized least squares estimators (Browne, 1974) with a weight matrix given by:

$$W_{ML} = .5D_p^T[\Sigma(\hat{\theta})^{-1} \otimes \Sigma(\hat{\theta})^{-1}]D_p \tag{2.9}$$

Corresponding to this discrepancy function, the test statistic T_{ML} can be defined as $T_{ML} = n\hat{F}_{ML}$. When the model is correct, under the assumption of multivariate normality, W_{ML} satisfies (2.6), F_{ML} is asymptotically optimal (Browne, 1974; Foldnes, Foss & Olsson, 2012, p. 373) and T_{ML} will be asymptotically distributed as a central chi-square variate. In fact, Browne (1984) suggests that under some conditions, the weight matrix given in Equation (2.9) may still be correctly specified provided there is no excess multivariate kurtosis in the observed variables.

2.4.2 The residual-based test statistics $\mathbf{T_{RADF}}$ and $\mathbf{T_{CRADF}}$

Browne (1982, 1984) described a class of residual-based test statistics based on arbitrary distributional assumptions. A thorough discussion of these statistics can be found in Foldnes, Foss and Olsson (2012). Yuan and Bentler (2007) adapt Browne's (1984) residual based ADF statistic for use in conjunction with the segregating method. The residual based test statistic, T_{RADF}, is given by

$$T_{RADF} = n\hat{e}^T\{\dot{\sigma}_c(\hat{\theta})[\dot{\sigma}_c(\hat{\theta})^T\hat{\Gamma}\dot{\sigma}_c(\hat{\theta})]^{-1}\dot{\sigma}_c(\hat{\theta})^T\}\hat{e} \tag{2.10}$$

Where. $\hat{e} = s - \sigma(\hat{\theta})$, $\dot{\sigma}_c(\hat{\theta})$ is a $p^* \times (p^* - q)$ full rank orthogonal complement of $\dot{\sigma}(\hat{\theta})$, and $\hat{\Gamma}$ is a sample estimate of Γ. In conventional factor analysis, $\hat{\Gamma}$ is often obtained by calculating the fourth-order central sample moments (e.g. Bentler, 2006). In using the segregating method, Yuan and Bentler (2002, 2006, 2007) proposed using generalized estimating equations (Liang & Zeger, 1986; Yuan & Jennrich, 1998) to obtain $\hat{\Gamma}$.

Yuan and Bentler (1998, 2007) suggested a small sample corrected version to T_{RADF}:

$$T_{CRADF} = \frac{T_{RADF}}{1 + \frac{T_{RADF}}{n}} \tag{2.11}$$

Neither T_{RADF} nor T_{CRADF} will be defined unless $[\dot{\sigma}_c(\hat{\theta})^T\hat{\Gamma}\dot{\sigma}_c(\hat{\theta})]^{-1}$ in Equation (2.10) is invertible.

2.4.3 The rescaled test statistic $\mathbf{T_{RML}}$

T_{RML} was designed to rescale T_{ML} based on excess skew and kurtosis in the observed variables (Satorra & Bentler, 1988). Let

$$\hat{U} = W_{ML} - W_{ML}\dot{\sigma}(\hat{\theta})\left(\dot{\sigma}(\hat{\theta})^T W_{ML}\dot{\sigma}(\hat{\theta})\right)^{-1}\dot{\sigma}(\hat{\theta})^T W_{ML} \tag{2.12}$$

Also let $k = \frac{tr(\hat{U}\hat{\Gamma})}{d}$. Then:

$$T_{RML} = \frac{T_{ML}}{k} \tag{2.13}$$

While T_{RML} is not generally chi-square distributed, its first moment is asymptotically equal to the first moment of χ^2_d (e.g. Bentler & Yuan, 1999).

2.5 Behavior of T_{ML} in the segregating methodology

In using the segregating method, T_{ML} is often expected to converge to a central chi-square distribution with d degrees of freedom if the model is correct and there is no excess skew or kurtosis in the observed variables. Several sources (Goldstein, 2003; Ryu & West, 2009; Hox, 2010; Hox & Maas, 2004) suggested that T_{ML} will behave in this way and can be used to evaluate between-level measurement models. In practice, however, and contrary to the advice given in these sources, T_{ML} may be inflated, and may not have the correct asymptotic distribution, even when the data is normally distributed and the model is correctly specified. The extent of the inflation will be related to a) the proportion of total observed variance attributable to group membership (i.e., the ICCs of the observed variables) and b) within group sample size.

For clarity of presentation, we will assume that the groups are balanced (i.e., that $n_1 = n_2 = \cdots = n_j = n$) and that the observed variables—in this case, y_j, $(y_{ij} - y_j)$, and y_{ij}—are multivariate normal in distribution. Then, let

$$W_W = .5D_p^T[\Sigma_W^{-1} \otimes \Sigma_W^{-1}]D_p \tag{2.14}$$

where Σ_W is the within-groups covariance matrix as defined in Equation (2.2). Also let

$$W_J = .5D_p^T[\Sigma_J^{-1} \otimes \Sigma_J^{-1}]D_p \tag{2.15}$$

where $\Sigma_J = \Sigma_B + \frac{1}{n}\Sigma_W$. The intraclass correlation (ICC) represents the proportion of observed variance attributable to group membership, and can be obtained from the diagonal elements of Σ_B and Σ_W. For any given item p, the intraclass correlation for that item can be expressed:

$$ICC_p = \frac{\Sigma_{Bpp}}{\Sigma_{Bpp} + \Sigma_{Wpp}} \tag{2.16}$$

where Σ_{Bpp} and Σ_{Wpp} are diagonal elements of Σ_B and Σ_W respectively. ICC values range between 0 and 1, and for fixed Σ_B, ICCs will increase as the elements of $\Sigma_W \to 0$.

Recall that $\sigma^2 = vech(\Sigma)$. Under normal theory, the asymptotic covariance matrix of $\sqrt{J}(\hat{\sigma}_B^2 - \sigma_B^2)$ is given by the inverse of the Fisher Information (Yuan & Bentler, 2006):

$$\Upsilon = W_J^{-1} + \frac{1}{n^2}\left((n-1)W_W\right)^{-1} \tag{2.17}$$

Υ depends not only on information from Σ_B, but on information from Σ_W as well, through W_J^{-1} and through $\frac{1}{n^2}\left((n-1)W_W\right)^{-1}$.

When a factor analysis is performed on $\hat{\Sigma}_B$ using Maximum Likelihood estimation in conventional software, it is associated with a weight matrix given only by W_B, the weight matrix in Equation (2.9) evaluated at $\Sigma(\hat{\theta})_B$, the estimated model-implied between-group covariance matrix. However, Equation (2.6) implies that in order for W_B to be correctly specified for F_{ML} in using the segregating method, it must converge to Υ^{-1}. In order for W_B to converge to Υ^{-1}, the $\frac{1}{n^2}\left((n-1)W_W\right)^{-1}$ and $\frac{1}{n}\Sigma_W$ terms need to be ignorable.

The ignorability of $\frac{1}{n^2}\left((n-1)W_W\right)^{-1}$ and $\frac{1}{n}\Sigma_W$ is directly related to the ICCs of the observed variables and within-group sample size. Keeping Σ_B fixed, as the ICC increases, the elements of Σ_W approach zero, and both $\frac{1}{n^2}\left((n-1)W_W\right)^{-1}$ and $\frac{1}{n}\Sigma_W$ become ignorable, For low ICCs, where the diagonal elements of Σ_W are relatively large, $\frac{1}{n^2}\left((n-1)W_W\right)^{-1}$ and $\frac{1}{n}\Sigma_W$ will not be ignorable. Alternatively,

keeping ICC fixed, as n increases $\frac{1}{n^2}\left((n-1)W_W\right)^{-1}$ and $\frac{1}{n}\Sigma_W$ become ignorable. For small within group sample sizes, $\frac{1}{n^2}\left((n-1)W_W\right)^{-1}$ and $\frac{1}{n}\Sigma_W$ will not be ignorable.

This implies that W_B is particularly likely to be misspecified for F_{ML} when ICCs are low or within-group sample sizes are small. Under those conditions, T_{ML} will not converge in distribution to a centrally distributed chi-square variate, even when the model is correct and the number of groups is sufficiently large. As a result, inferences about model structure based on T_{ML} may not be valid for the segregated analysis of $\hat{\Sigma}_B$ even when the data is normally distributed. It should be noted that, while the above argument assumed that the groups were balanced, this assumption was made only to simplify the presentation. Results in Yuan and Bentler (2002, 2006) suggest that similar results would hold for the case of unbalanced groups.

2.6 Behavior of T_{RADF}, T_{CRADF} and T_{RML}

Unlike T_{ML}, the residual-based test statistics T_{RADF} and T_{CRADF} use information from both between and within covariance sources through $\hat{\Gamma}$. Thus these test statistics are expected to converge to the correct distribution regardless of ICC or within group sample size. T_{RML} is expected to converge to a distribution with the correct first moment regardless of ICC and within group sample size. The scaling constant, k will be greater than 1. Bentler (2006) explained that $tr(\hat{U}\hat{\Gamma})$ can be thought of as a way to determine the discrepancy between the hypothesized model and data distribution (carried by $\hat{U}$) and the true data distribution (carried by $\hat{\Gamma}$). In analyzing $\hat{\Sigma}_B$, the discrepancy between between $\hat{U}$ and $\hat{\Gamma}$ occurs because $\hat{\Gamma}$ is based on information from both $\hat{\Sigma}_B$ and $\hat{\Sigma}_W$, and $\hat{U}$ is based on information from $\hat{\Sigma}_B$ alone.

2.7 Issues with T_{ML} in the study of classroom climate

The relationship between the discrepancy function, the weight matrix, item ICCs and within-group sample size is rarely made explicit in methodological literature on multilevel factor analysis. Even when the poor performance of T_{ML} is noted (Hox, 2010; Muthén, 1994, p. 389; Yuan & Bentler, 2007), the possible role of either item ICC or within-group sample size in the misspecification of F_{ML} for W_B is not described. In fact, several sources (Goldstein, 2003; Hox, 2010; Hox & Maas, 2004; Ryu & West, 2009) suggested that the segregating method is a "viable method" (Hox & Maas, 2004, p. 145) that can be "implemented within the preexisting ML SEM framework" (Ryu & West, 2009, p. 600).

As a result, there is confusion in the applied literature on the interpretation of T_{ML}. There are many cases in the applied literature where an inflated valued of T_{ML} is assumed to suggest model misspecification, and often the theorized between-groups model is then modified, either by removing items, adding additional factors, or modifying paths (e.g. Kaplan & Elliott, 1997; Mathiesen, Torsheim & Einarsen, 2006; van Horn, 2003). The possibility that T_{ML} may also reflect the fact that F_{ML} is misspecified for W_B is unexplored and untested.

The advice to use T_{ML} for model fit assessment is particularly problematic when the segregating method is used to assess the factor structure of classroom climate surveys, because the two conditions most likely to cause issues with the performance of T_{ML}—low item ICCs and relatively small within-group sample sizes—are particularly common in this field. Generally speaking, item ICCs for climate variables are often less than .1 and rarely greater than .3 (den Brok, Bergen, Stahl, & Brekelmans, 2004; Marsh et al, p. 115; Toland & Ayala, 2005). Class sizes typically range between 12 and 25 students per class (e.g., Holfve-Sabel & Gustaffson, 2005; Kunter et al.,2008). Under these conditions, the inflation of T_{ML} is likely to be severe. Relatedly, Type I error rates are likely to be far higher. It is unlikely that inferences about the between-classroom measurement models based on T_{ML}

would be valid.

Because T_{ML} is expected to perform poorly in the evaluation of between-level measurement models for classroom climate surveys, it may seem reasonable to recommend the use of alternative test statistics, such as the residual-based and rescaled test statistics, since the theory outlined above suggests these statistics should perform well asymptotically. In fact, Yuan and Bentler (2007) recommended the use of T_{RML} and T_{CRADF} for model evaluation in conjunction with the segregating methodology. However, there is only limited simulation work with the residual-based and rescaled test statistics in a multilevel context, and there are many known issues with statistics like T_{RADF}, T_{CRADF} and T_{RML} in conventional factor analysis (e.g., Bentler & Yuan, 1999; Curran, West & Finch, 1996; Hu, Bentler & Kano, 1992; Kaplan & Muthén, 1985; Kaplan & Muthén, 1992, Powell & Schaefer, 2001; Yuan & Bentler, 1998), particularly with small sample sizes and large models (large models are often defined as those containing more than 50 *df* (e.g. Kaplan & Muthén, 1992)). It may be expected that these conditions would also present problems in multilevel investigations. In conventional factor analysis, when models are large and sample sizes are small, T_{RADF} and T_{RML} tend to over-reject correct models, and T_{CRADF} tends to under-reject correct models (Yuan & Bentler, 1999).

In fact, as it turns out, these specific conditions (small sample sizes and large models) are also likely to occur with student surveys of classroom climate. In the literature on student surveys of classroom climate, the number of classrooms (i.e. the group level sample sizes) is typically between 50-500 (e.g. Fauth, Decristan, Riser, Klieme & Buttner, 2014; Holfve-Sabel & Gustaffson, 2005; Kunter et al., 2008; Toland & Ayala, 2005). Measurement models range from 25 degrees of freedom to well over 150 degrees of freedom (e.g., den Brok, Bergen, Stahl, & Brekelmans, 2004; Fauth, Decristan, Riser, Klieme & Buttner, 2014; Holfve-Sabel & Gustaffson, 2005; Kunter et al. 2008; Toland & Ayala, 2005). It is not clear

whether, under these conditions, the residual-based test statistics or the rescaled test statistics would continue to perform well. It is also unclear whether Yuan and Bentler's (2007) recommendations to use to use T_{RML} and T_{CRADF}, which were based on a simulation study using high item ICCs, large within-group sample sizes, relatively small measurement models, and a large number of groups, would be supported under a wider range of conditions, particularly those typically found in survey-based research on classroom climate.

The present study uses an illustrative example and a simulation in order to a) Illustrate the extent to which T_{ML} will be inflated b) Demonstrate how item ICC and within group sample size influence the distribution of T_{ML} c) Investigate the performance of several alternative test statistics—specifically T_{RML}, T_{RADF} and T_{CRADF}—under a broader range of conditions, particularly those that are frequently encountered in survey-based research on classroom climate. The empirical example comes from the Tripod Classroom Environment Survey (Ferguson, 2010), which is administered to measure aspects of classroom environment. Using the illustrative example and the simulation study, the following three research questions were addressed:

Research Question 1: To what extent can inferences about the measurement structure of the Tripod Classroom Environment Survey based on T_{ML} differ from those based on residual-based and rescaled test statistics?

Research Question 2: How do item ICC and within group sample size influence the distribution of T_{ML} ? How do item ICC and within group sample size influence the differences between the two estimated asymptotic covariance matrices, $\hat{W}_B^{-1}$ and $\hat{\Gamma}$?

Research Question 3: How do T_{RML}, T_{RADF} and T_{CRADF} perform under a broader range of conditions, particularly those that are frequently encountered in survey-based research on classroom climate?

CHAPTER 3

Method

3.1 Data sources

The Tripod Classroom Environment Survey. The Tripod Survey (Ferguson, 2010) is designed to assess seven dimensions of teaching practice, often referred to as the Seven Cs: Caring, Captivating, Conferring, Clarifying, Challenging, Controlling, Consolidating. This version of the Tripod Survey was administered in an urban school district in California in 2010. This example uses 5 of the 8 items from the Challenging dimension that are rated on a 5-point scale (1 = totally untrue and 5 = totally true). Scores based on these five items correlate approximately .96 with the total score on the Challenging dimension, and show both good internal consistency (Cronbach's $\alpha = .87$) and aggregate-level reliability (average $ICC(2) = .76$). The sample used in this analysis contained 5,508 students in 285 classrooms. The average classroom size was approximately 17 students. Students are treated as nested within classrooms, and it is assumed that each student has rated only one classroom. Descriptive information about the survey items is summarized in Table 3.1.

Simulated Datasets. Data were generated from a population model with two within level factors and one between level factor. This population model was selected because several sources suggest that the between-level factor structure is likely to be simpler than the structure at the within-level (e.g. Holfve-Sabel & Gustaffson, 2005; Muthén & Asparouhov, 2011). Simulation conditions were selected in order to reflect the conditions commonly reported in survey-based re-

Table 3.1: Descriptive Statistics for Tripod Survey Variables

Item		Mean	St. Dev.	ICC
1)	My teacher asks questions to be sure we are following along when he or she is teaching.	4.11	.40	.07
2)	My teacher asks students to explain more about answers they give.	3.93	.39	.10
3)	In this class, my teacher accepts nothing less than our full effort.	3.93	.46	.13
4)	My teacher doesn't let people give up when the work gets hard.	4.07	.38	.10
5)	My teacher wants us to use our thinking skills, not just memorize things.	3.95	.42	.12

search on classroom climate. Four conditions were manipulated: a)Item ICCs (I$CC = .50$, $ICC = .26$, $ICC = .10$, $ICC = .05$), b)Level 2 sample size ($J = 100$, $J = 200$, $J = 500$), c)Group size ($n = 10$, $n = 30$, $n = 50$), d)The size of the measurement model ($df = 9$, $df = 54$, $df = 135$).

For the $ICC = .50$ condition with 6 observed variables, the generating model used the following parameters:

$$\Lambda_W = \begin{pmatrix} .7 & 0 \\ .7 & 0 \\ .7 & 0 \\ 0 & .7 \\ 0 & .7 \\ 0 & .7 \end{pmatrix}, \Lambda_B = \begin{pmatrix} .7 \\ .7 \\ .7 \\ .7 \\ .7 \\ .7 \end{pmatrix}, \Phi_W = \begin{pmatrix} 1 & .3 \\ .3 & 1 \end{pmatrix}, \Psi_W = \Psi_B = diag(.51) \tag{3.1}$$

Ψ_B and Ψ_W were 6×6 diagonal matrices so that all of the diagonal elements equal .51. For the other ICC conditions used in this simulation, Λ_B, Φ_W and Ψ_B were fixed, and Λ_W and Ψ_W were varied. For the $ICC = .26$ condition, the non-zero elements of Λ_W were set to 1.41, and the diagonal elements of Ψ_W were set to 2.00. For the $ICC = .10$ condition, the non-zero elements of Λ_W were set

to 2.1, and the diagonal elements of Ψ_W were set to 4.59. For the $ICC = .05$ condition, the non-zero elements of Λ_W were set to 3.08, and the diagonal elements of Ψ_W were set to 9.50. Model size was varied by adding additional items, but keeping the general pattern of factor loadings the same as given in Equation (3.1). In total, this simulation contained in $4 \times 3 \times 3 \times 3 = 108$ conditions. While certain constellations of conditions may be unlikely to occur in practice (i.e. many large classrooms, high ICCs and a small model), the inclusion of conditions across this range allows for a more comprehensive study of the behavior of the test statistics. 500 replications were conducted in each condition for a total of 54,000 replications. Simulations were conducted using MPlus's (Muthén & Muthén, 2010) Monte Carlo capabilities. For each of the replicated data sets, the MPlusAutomation package (Hallquist, 2012) in R (R Development Core Team, 2012) was used to obtain saturated estimates of Σ_B and Σ_W. T_{ML}, T_{RADF}, T_{CRADF} and T_{RML} were estimated in EQS (Bentler, 2006) using the REQS (Mair & Wu, 2012) package.

3.1.1 Analytic Approach

To address the first research question, the Tripod Survey data was used. ML estimates of Σ_B and Σ_W were obtained. $\hat{\Sigma}_B$ was then used as the input covariance matrix for a confirmatory factor analysis, where the hypothesized model was unidimensional (i.e., all 5 items loaded onto one factor). Then, the four test statistics T_{ML}, T_{RADF}, T_{CRADF} and T_{RML} were estimated.

Simulated data was used to answer the second research question. For each simulation condition, the mean and standard deviation of T_{ML} was estimated, and an empirical Type I error rate was calculated. For the purpose of this study, the Type I error rate was calculated at the nominal $\alpha = .05$ level. Because it is expected that the empirical error rates will differ somewhat from the nominal rate, an acceptable empirical error rate is taken as one that falls in the inter-

val [.028, .079], the estimated 2-sided 99% adjusted Wald confidence interval (e.g. Agresti & Coull, 1998). In addition to T_{ML}, the two asymptotic covariance matrices, W_B^{-1} and $\hat{\Gamma}$, were compared through their square distances from each other: $\|W_B^{-1} - \hat{\Gamma}\|$. Based on Equation (2.17), it is anticipated that the distance between the two covariance matrices should increase as ICCs and within group sample sizes decrease.

In order address the third research question, investigating the performance of T_{RADF}, T_{CRADF} and T_{RML} under a range of conditions similar to those encountered in survey-based research on classroom climate, means and standard deviations of these three test statistics were estimated for each simulation condition, and an empirical model rejection rate was calculated. As in the case of T_{ML}, the rejection rate was calculated at the nominal $\alpha = .05$ level and acceptable rates were those in the interval [.028, .079]. It should be noted that for several conditions (when $J = 100$ and $df = 135$), the residual-based test statistics are not estimable because $[\dot{\sigma}_c(\hat{\theta})^T \hat{\Gamma} \dot{\sigma}_c(\hat{\theta})]^{-1}$ is not invertible under these conditions, and so those statistics are not included in those specific analyses.

CHAPTER 4

Results

4.1 Research Question 1

To what extent can inferences about the measurement structure of the Tripod Classroom Environment Survey based on T_{ML} differ from those based on residual-based and rescaled test statistics?

The estimate of T_{ML} is 136.9. This can be referred to χ^2_5 and suggests strong evidence for rejecting the null hypothesis that the proposed model holds in the population ($p < .0001$). If T_{ML} were used as the basis for model evaluation, it would be concluded that these five items are not unidimensional.

However, based on the theoretical results presented above, there is reason to suspect that the T_{ML} test statistic should not be trusted in this particular case. Firstly, the item ICCs are fairly low, ranging from .07 to .13 (Table 3.1). Secondly, the average number of individuals in each classroom is fairly small. Even if all of the distributional assumptions were satisfied, with ICCs that are in this range, the correct specification of F_{ML} for W_B would require much larger classroom sizes in order for T_{ML} to have the correct distribution. Thus, it may be more appropriate to make model inferences based on rescaled or residual-based test statistics. Here, T_{RADF} (4.54, $p = .454$), T_{CRADF} (4.47, $p = .484$) and T_{RML} (5.50, $p = .358$) all suggest strong evidence for failing to reject the null hypothesis. In other words, these three test statistics all suggest that the items are indeed unidimensional, an inference that completely contradicts the inference based on T_{ML}.

It should be noted that while this example provides a clear illustration of how

low ICCs and small within-group sample sizes can distort inferences about the between-classroom model based on T_{ML}, it was also limited in some important ways. First, the data generating mechanism was unknown. While the inflation of T_{ML} relative to T_{RADF}, T_{CRADF} and T_{RML} is related to ICC and within-group sample size, it is possible that other factors, including multivariate kurtosis, play a role in model appraisal. Second, the model is relatively small, containing only 5 variables and 5 degrees of freedom, and so while the rescaled and residual based test statistics provide valid inferences in this case, these results may not generalize to larger models. This issues are addressed in the analyses that follow.

4.2 Research Question 2

How do item ICC and within group sample size influence the distribution of T_{ML}? How do item ICC and within group sample size influence the differences between the two asymptotic covariance matrices, W_B^{-1} and $\hat{\Gamma}$?

Tables 4.1–4.3 present the test statistic means, variances, and empirical Type I error rates across simulation conditions. As expected, as either ICC or within group sample size decrease, T_{ML} increases, and, relatedly, Type I error rates increase. T_{ML} is only well behaved with 500 groups, more than 30 individuals per group and an ICC of .50 (Table 4.1). This condition is most similar to the simulation conditions of Ryu and West (2009) and Hox and Maas (2004), and offers some insight into why those studies concluded that Maximum Likelihood methods were appropriate for use in conjunction with the segregating method.

T_{ML} inflation can be quite severe. When ICCs are low and the within group sample sizes are small, the correct model is rejected 100% of the time, and the test statistic mean is about 20 times larger than expected, for all model sizes. This pattern of inflation suggests that T_{ML} will not provide valid inferences about

between-classroom measurement models in survey-based research on classroom climate. The results presented in Tables 4.1–4.3 also suggest little evidence that T_{ML} would ever converge to the correct distribution, regardless of the number of groups that are included in the sample. For example, for $ICC = .50$, with within-group sample sizes of 10, there is little evidence of convergence as the number of groups increases from 100 to 500.

Table 4.1: T_{ML} Means, Standard Deviations and Type I Error Rates, $df = 9$

	Group Sizes					
	10		30		50	
	Mean (sd)	Rej	Mean (sd)	Rej	Mean (sd)	Rej
$J = 500$						
$ICC = .50$	12.55 (6.15)	0.216	10.05 (4.65)	**0.072**	9.64 (4.29)	**0.054**
$ICC = .26$	18.37 (9.55)	0.490	11.68 (5.38)	0.142	10.58 (4.75)	0.102
$ICC = .10$	55.57 (35.89)	0.942	19.34 (9.21)	0.524	14.74 (6.93)	0.318
$ICC = .05$	176.35 (143.71)	1.00	35.87 (19.34)	0.896	22.6 (11.55)	0.644
$J = 200$						
$ICC = .50$	12.34 (6.10)	0.198	10.11 (4.73)	0.088	9.77 (4.52)	**0.074**
$ICC = .26$	18.32 (9.55)	0.492	11.71 (5.56)	0.186	10.75 (4.99)	0.104
$ICC = .10$	63.01(50.91)	0.962	19.72 (10.07)	0.544	15.13 (7.40)	0.336
$ICC = .05$	193.74 (120.24)	1.00	39.36 (26.93)	0.872	24.25 (14.80)	0.690
$J = 100$						
$ICC = .50$	12.52 (6.22)	0.190	10.27 (4.81)	0.092	10.20 (5.09)	0.102
$ICC = .10$	80.96 (61.30)	0.966	21.14 (11.76)	0.592	15.93 (8.35)	0.382
$ICC = .26$	19.55 (11.10)	0.514	12.01 (5.73)	0.180	11.22 (5.62)	0.158
$ICC = .05$	192.25 (96.32)	1.00	47.54 (41.23)	0.906	26.96 (19.26)	0.690

Note: Empirical Type I error rates in the interval [.028,.079] shown in bold.

Tables 4.4–4.6 present the square distances between W_B^{-1} and $\hat{\Gamma}$. As anticipated by theory, the two asymptotic covariance matrices diverge as either ICC decreases, or within group sample size decreases. At $ICC = .50$, with group sizes of 50, the distance between the covariance matrices is relatively small, implying a small amount of misspecification of the discrepancy function for W_B. Relatedly, T_{ML} is relatively well behaved. For $ICC = .05$, however, the distances between the covariance matrices are quite large, the misspecification of F_{ML} for W_B is more severe and T_{ML} is more inflated.

Table 4.2: T_{ML} Means, Standard Deviations and Type I Error Rates, $df = 54$

	Group Sizes					
	10		30		50	
	Mean (sd)	Rej	Mean (sd)	Rej	Mean (sd)	Rej
$J = 500$						
$ICC = .50$	73.08 (14.27)	0.502	58.87 (11.55)	0.124	58.33 (11.42)	0.110
$ICC = .26$	107.62 (24.04)	0.964	67.79 (13.55)	0.356	63.70 (12.80)	0.248
$ICC = .10$	355.59 (158.71)	1.00	111.12 (24.62)	0.968	87.38 (18.67)	0.758
$ICC = .05$	1054.79 (336.47)	1.00	211.24 (66.09)	1.00	133.36 (33.78)	0.994
$J = 200$						
$ICC = .50$	73.46 (14.97)	0.508	61.18 (12.24)	0.192	58.67 (11.50)	0.110
$ICC = .26$	113.20 (32.08)	0.962	70.76 (14.57)	0.432	64.35 (12.70)	0.248
$ICC = .10$	445.88 (163.05)	1.00	121.21 (30.51)	0.976	91.14 (19.63)	0.84
$ICC = .05$	1233.88 (299.86)	1.00	262.93 (109.39)	1.00	150.01 (50.77)	0.994
$J = 100$						
$ICC = .50$	78.61 (16.36)	0.630	64.67 (12.76)	0.272	61.09 (11.66)	0.172
$ICC = .26$	136.81 (54.23)	0.982	76.33 (16.40)	0.564	67.57 (13.38)	0.350
$ICC = .10$	553.58 (153.69)	1.00	151.84 (63.91)	0.992	100.86 (29.09)	0.894
$ICC = .05$	1144.92 (205.67)	1.00	341.59 (110.41)	1.00	190.39 (78.90)	1.00

Note: Empirical Type I error rates in the interval [.028,.079] shown in bold.

Table 4.3: T_{ML} Means, Standard Deviations and Type I Error Rates, $df = 135$

	Group Sizes					
	10		30		50	
	Mean (sd)	Rej	Mean (sd)	Rej	Mean (sd)	Rej
$J = 500$						
$ICC = .50$	177.93 (22.55)	0.728	151.27 (18.48)	0.248	144.41 (17.36)	0.122
$ICC = .26$	263.08 (48.18)	0.998	174.22 (22.50)	0.668	157.51 (19.23)	0.384
$ICC = .10$	899.07 (270.84)	1.00	289.69 (46.98)	1.00	218.16 (29.11)	0.972
$ICC = .05$	2552.47 (512.77)	1.00	574.96 (170.26)	1.00	342.00 (64.83)	1.00
$J = 200$						
$ICC = .50$	185.74 (23.90)	0.814	154.93 (18.90)	0.310	148.12 (17.44)	0.180
$ICC = .26$	300.61 (79.66)	1.00	180.36 (23.26)	0.770	162.35 (19.74)	0.466
$ICC = .10$	1125.38 (221.85)	1.00	324.90 (71.50)	1.00	232.20 (35.11)	0.994
$ICC = .05$	3009.19 (437.05)	1.00	660.18 (202.01)	1.00	408.67 (123.65)	1.00
$J = 100$						
$ICC = .50$	204.02 (30.38)	0.926	163.11 (19.66)	0.470	156.3 (19.01)	0.340
$ICC = .26$	378.52 (106.55)	1.00	195.50 (29.51)	0.910	173.29 (22.64)	0.668
$ICC = .10$	1424.66 (227.68)	1.00	423.72 (107.89)	1.00	273.63 (66.51)	1.00
$ICC = .05$	2485.48 (248.38)	1.00	911.03 (175.26)	1.00	520.02 (123.77)	1.00

Note: Empirical Type I error rates in the interval [.028,.079] shown in bold.

Table 4.4: Square Distances Between Asymptotic Covariance Matrices, $df = 9$

	Group Sizes		
	10	30	50
$J = 500$			
$ICC = .50$	13.37	6.93	5.89
$ICC = .26$	79.20	13.59	8.45
$ICC = .10$	930.55	74.89	28.98
$ICC = .05$	7371.88	388.32	121.33
$J = 200$			
$ICC = .50$	24.97	16.36	14.40
$ICC = .26$	98.54	25.16	17.67
$ICC = .10$	1010.64	96.7	41.33
$ICC = .05$	7864.97	440.61	141.81
$J = 100$			
$ICC = .50$	43.98	27.98	28.42
$ICC = .26$	130.40	13.59	33.41
$ICC = .10$	1144.99	117.3	63.28
$ICC = .05$	8502.18	486.50	178.86

Table 4.5: Square Distances Between Asymptotic Covariance Matrices, $df = 54$

	Group Sizes		
	10	30	50
$J = 500$			
$ICC = .50$	132.70	77.79	72.13
$ICC = .26$	437.04	112.50	86.69
$ICC = .10$	7547.11	667.49	286.63
$ICC = .05$	58907.71	3317.32	1097.86
$J = 200$			
$ICC = .50$	261.67	190.56	176.47
$ICC = .26$	959.62	282.88	212.88
$ICC = .10$	8429.14	912.93	430.09
$ICC = .05$	63017.13	3871.44	1301.13
$J = 100$			
$ICC = .50$	458.82	364.11	324.45
$ICC = .26$	1293.77	507.18	373.91
$ICC = .10$	9697.79	1350.5	647.05
$ICC = .05$	69335.72	4908.45	1669.69

Table 4.6: Square Distances Between Asymptotic Covariance Matrices, $df = 135$

	Group Sizes		
	10	30	50
$J = 500$			
$ICC = .50$	572.75	335.9	343.99
$ICC = .26$	3099.04	617.33	468.70
$ICC = .10$	30286.52	2679.83	1233.39
$ICC = .05$	234646.43	13045.81	4553.31
$J = 500$			
$ICC = .50$	1158.64	759.65	756.52
$ICC = .26$	4092.00	1101.68	922.85
$ICC = .1033777.94$	3520.95	1849.29	
$ICC = .05$	248861.52	15157.23	5553.22
$J = 500$			
$ICC = .50$	2036.26	1361.54	1508.60
$ICC = .26$	5746.05	1829.12	1772.68
$ICC = .1040686.82$	4864.94	3033.12	
$ICC = .05$	282624.76	18259.65	7540.48

4.3 Research Question 3

How do T_{RML}, T_{RADF} and T_{CRADF} perform under a broader range of conditions, particularly those that are frequently encountered in survey-based research on classroom climate?

4.3.1 Performance of $\mathbf{T_{RADF}}$

Consistent with theoretical expectation, there is evidence that T_{RADF} converges to the correct distribution as the number of groups increases regardless of ICC or within-group sample size. This pattern of convergence is most apparent in Table 4.7 as the number of groups increases from 100 to 500. With 100 groups, T_{RADF} over rejects the correct model for nearly all ICC and within group sample size conditions. Contrary to this pattern, T_{RADF} has a mean that is too low at $ICC = .05$ and $n = 10$, which may reflect some of the instability of the estimates at low ICCs and small sample sizes. With 500 groups, T_{RADF} is much better behaved. However, when the model is sufficiently large, the number of groups would have to be enormous in order for T_{RADF} to provide correct inferences. In Table 4.9,

when the model is large and the number of groups is small, T_{RADF} rejects the correct model 100% of the time. Even with 500 groups the empirical Type I error rates approach 90%. This is consistent with results from both conventional and multilevel factor analysis, where it has been shown that T_{RADF} and other similar statistics converge slowly to the appropriate distribution (e.g. Curran, West & Finch, 1996; Hu, Bentler & Kano, 1992; Kaplan & Muthén, 1985; Kaplan & Muthén, 1992, Powell & Schaefer, 2001; Yuan & Bentler,1998, 2003, 2007).

4.3.2 Performance of $\mathbf{T_{CRADF}}$

T_{CRADF} shows well-behaved means, standard deviations and empirical Type I error rates across a wide variety of simulation conditions. Table 4.7 which displays results for the small models ($df = 9$) shows that T_{CRADF} performs well when the number of groups is sufficiently large, relative to the size of the model. This pattern continues in Table 4.8, with the medium sized models ($df = 54$), provided that the number of groups is sufficiently large ($J = 200$ or $J = 500$). However, when the number of groups is small relative to the size of the model (for example, in the condition with $J = 100$ and $df = 54$), the multilevel version of T_{CRADF} performs similarly to the conventional version (Yuan & Bentler, 1998; Bentler & Yuan, 1999). That is, the statistic accepts more correct models than would be expected by chance.

4.3.3 Performance of $\mathbf{T_{RML}}$

Consistent with theory, in all ICC and group size conditions, the scaling constant for T_{RML}, k, is larger than 1. The amount of rescaling changes as a function of within group sample size and item ICC. At $ICC = .50$, there is virtually no rescaling at all. At $ICC = .05$, the T_{ML} value is scaled by almost 90%. There is also some evidence that the mean of T_{RML} converges appropriately as the number

of groups increases. For small models, relatively large sample sizes and high ICCs, T_{RML} behaves well. These conditions are most similar to the conditions that lead Yuan and Bentler (2007) to recommend T_{RML} for model testing. However, when the full range of simulation conditions are considered, it becomes clear that T_{RML}, cannot adequately control Type I errors when group sizes are small or when ICCs are low. In the condition where $[\dot{\sigma}_c(\hat{\theta})^T \hat{\Gamma} \dot{\sigma}_c(\hat{\theta})]^{-1}$ is not invertible and neither T_{RADF} nor T_{CRADF} are estimable, T_{RML} is unable to control Type I errors under any of the simulation conditions. The current study suggests that, contrary to the recommendation of Yuan and Bentler (2007), and even though T_{RML} always performs better than T_{ML}, T_{RML} should not be used to make inferences about model fit in conjunction with the segregating method.

Table 4.7: T_{RML}, T_{RADF} and T_{CRADF} Performance, $df = 9$

		Group Sizes					
		10		30		50	
		Mean (sd)	Rej	Mean (sd)	Rej	Mean (sd)	Rej
$J = 500$							
T_{RML}	$ICC = .50$	9.59 (4.69)	**0.078**	9.19 (4.23)	**0.062**	9.16 (4.06)	**0.046**
	$ICC = .26$	9.56 (4.97)	0.082	9.26 (4.23)	**0.054**	9.21 (4.12)	**0.048**
	$ICC = .10$	9.59 (4.69)	**0.078**	9.19 (4.23)	**0.062**	9.16 (4.06)	**0.046**
	$ICC = .05$	10.31 (7.33)	0.128	9.60 (5.02)	0.098	9.43 (4.80)	0.084
T_{RADF}	$ICC = .50$	9.79 (4.79)	**0.072**	9.41 (4.41)	**0.060**	9.42 (4.28)	**0.056**
	$ICC = .26$	9.69 (4.84)	**0.070**	9.49 (4.45)	**0.070**	9.44 (4.33)	**0.058**
	$ICC = .10$	9.79 (4.79)	**0.072**	9.41 (4.41)	**0.06**	9.42 (4.28)	**0.056**
	$ICC = .05$	9.02 (4.27)	**0.046**	9.60 (4.63)	0.086	9.53 (4.70)	0.094
T_{CRADF}	$ICC = .50$	9.56 (4.57)	**0.064**	9.20 (4.21)	**0.050**	9.21 (4.10)	**0.040**
	$ICC = .26$	9.46 (4.61)	**0.062**	9.28 (4.24)	**0.060**	9.23 (4.14)	**0.052**
	$ICC = .10$	9.56 (4.57)	**0.064**	9.20 (4.21)	**0.050**	9.21 (4.1)	**0.040**
	$ICC = .05$	8.82 (4.10)	**0.034**	9.38 (4.41)	**0.07**4	9.31 (4.48)	**0.078**
$J = 200$							
T_{RML}	$ICC = .50$	9.42 (4.61)	**0.074**	9.25 (4.3)	**0.064**	9.30 (4.32)	**0.062**
	$ICC = .26$	9.43 (4.84)	**0.058**	9.28 (4.38)	**0.052**	9.35 (4.35)	**0.056**
	$ICC = .10$	9.42 (4.61)	**0.074**	9.25 (4.30)	**0.064**	9.30 (4.32)	**0.062**
	$ICC = .05$	9.91 (6.40)	0.130	10.02 (6.26)	0.102	9.88 (5.75)	0.088
T_{RADF}	$ICC = .50$	9.96 (5.04)	0.096	9.81 (4.73)	0.088	9.87 (4.78)	**0.078**
	$ICC = .26$	9.76 (4.88)	**0.076**	9.77 (4.66)	**0.072**	9.90 (4.83)	0.086
	$ICC = .10$	9.96 (5.04)	0.096	9.81 (4.73)	0.088	9.87 (4.78)	**0.078**
	$ICC = .05$	8.21 (3.69)	**0.026**	9.70 (4.56)	**0.072**	9.97 (4.86)	**0.076**
T_{CRADF}	$ICC = .50$	9.38 (4.45)	**0.064**	9.26 (4.22)	**0.05**	9.30 (4.25)	**0.054**
	$ICC = .26$	9.20 (4.32)	**0.052**	9.22 (4.16)	**0.044**	9.33 (4.29)	**0.054**
	$ICC = .10$	9.38 (4.45)	**0.064**	9.26 (4.22)	**0.05**	9.30 (4.25)	**0.054**
	$ICC = .05$	8.83 (3.97)	**0.038**	9.18 (4.12)	**0.05**	9.36 (4.28)	**0.052**
$J = 100$							
T_{RML}	$ICC = .50$	9.59 (4.76)	**0.072**	9.42 (4.39)	**0.050**	9.72 (4.85)	0.086
	$ICC = .26$	9.91 (5.49)	0.098	9.49 (4.49)	**0.064**	9.75 (4.90)	**0.074**
	$ICC = .10$	9.59 (4.76)	**0.072**	9.42 (4.39)	**0.050**	9.72 (4.85)	0.086
	$ICC = .05$	8.99 (5.75)	0.096	11.33 (9.48)	0.124	10.63 (7.25)	0.128
T_{RADF}	$ICC = .50$	10.86 (5.60)	0.122	10.56 (5.29)	0.118	10.89 (5.75)	0.152
	$ICC = .26$	10.70 (5.51)	0.134	10.55 (5.25)	0.122	10.89 (5.85)	0.150
	$ICC = .10$	10.86 (5.60)	0.122	10.56 (5.29)	0.118	10.89 (5.75)	0.152
	$ICC = .05$	7.45 (3.51)	0.020	10.26 (4.81)	0.086	10.88 (5.81)	0.150
T_{CRADF}	$ICC = .50$	9.55 (4.32)	**0.064**	9.33 (4.16)	**0.054**	9.57 (4.45)	**0.070**
	$ICC = .26$	9.43 (4.28)	**0.048**	9.33 (4.13)	**0.048**	9.56 (4.50)	**0.076**
	$ICC = .10$	9.55 (4.32)	**0.064**	9.33 (4.16)	**0.054**	9.57 (4.45)	**0.070**
	$ICC = .05$	6.82 (2.94)	0.006	9.12 (3.84)	**0.034**	9.55 (4.46)	**0.078**

Note: Empirical Type I error rates in the interval [.028,.079] shown in bold.

Table 4.8: T_{RML}, T_{RADF} and T_{CRADF} Performance, $df = 54$

		Group Sizes					
		10		30		50	
		Mean (sd)	Rej	Mean (sd)	Rej	Mean (sd)	Rej
$J = 500$							
T_{RML}	$ICC = .50$	56.31 (11.02)	0.084	53.96 (10.59)	**0.052**	55.38 (10.84)	**0.074**
	$ICC = .26$	57.11 (12.64)	0.122	54.09 (10.82)	**0.062**	55.52 (11.16)	**0.068**
	$ICC = .10$	56.31 (11.02)	0.084	53.96 (10.59)	**0.052**	55.38 (10.84)	**0.074**
	$ICC = .05$	68.80 (21.62)	0.380	58.39 (17.60)	0.164	56.86 (14.24)	0.124
T_{RADF}	$ICC = .50$	64.28 (13.3)	0.282	61.62 (12.72)	0.192	63.48 (12.97)	0.242
	$ICC = .26$	64.68 (13.34)	0.278	61.74 (12.88)	0.200	63.47 (13.31)	0.248
	$ICC = .10$	64.28 (13.30)	0.282	61.62 (12.72)	0.192	63.48 (12.97)	0.242
	$ICC = .05$	60.24 (11.88)	0.168	61.66 (12.95)	0.192	63.22 (13.91)	0.260
T_{CRADF}	$ICC = .50$	56.68 (10.37)	**0.072**	54.61 (9.98)	**0.06**	56.07 (10.12)	**0.062**
	$ICC = .26$	57.00 (10.38)	**0.074**	54.69 (10.10)	**0.052**	56.05 (10.38)	**0.054**
	$ICC = .10$	56.68 (10.37)	**0.072**	54.61 (9.98)	**0.060**	56.07 (10.12)	**0.062**
	$ICC = .05$	53.54 (9.42)	**0.032**	54.63 (10.16)	**0.054**	55.83 (10.83)	**0.064**
$J = 200$							
T_{RML}	$ICC = .50$	56.47 (11.50)	0.082	56.11 (11.11)	0.088	55.79 (10.88)	0.084
	$ICC = .26$	59.55 (16.50)	0.172	56.48 (11.55)	0.102	56.19 (11.05)	0.092
	$ICC = .10$	56.47 (11.50)	0.082	56.11 (11.11)	0.088	55.79 (10.88)	0.084
	$ICC = .05$	76.74 (21.21)	0.540	70.37 (28.15)	0.354	62.96 (20.48)	0.224
T_{RADF}	$ICC = .50$	81.20 (19.29)	0.648	81.26 (19.61)	0.662	81.2 (18.75)	0.668
	$ICC = .26$	81.76 (20.1)	0.660	81.02 (19.26)	0.656	81.55 (18.98)	0.660
	$ICC = .10$	81.20 (19.29)	0.648	81.26 (19.61)	0.662	81.20 (18.75)	0.668
	$ICC = .05$	65.97 (13.13)	0.312	80.43 (19.01)	0.658	81.86 (19.36)	0.668
T_{CRADF}	$ICC = .50$	56.93 (9.56)	**0.054**	56.94 (9.77)	**0.054**	56.97 (9.18)	**0.052**
	$ICC = .26$	57.16 (9.85)	**0.068**	56.84 (9.61)	**0.058**	57.14 (9.28)	**0.056**
	$ICC = .10$	56.93 (9.56)	**0.054**	56.94 (9.77)	**0.054**	56.97 (9.18)	**0.052**
	$ICC = .05$	49.12 (7.32)	0.004	56.56 (9.44)	0**.048**	57.27 (9.47)	**0.064**
$J = 100$							
T_{RML}	$ICC = .50$	60.24 (12.45)	0.170	59.41 (11.58)	0.152	58.42 (11.22)	0.108
	$ICC = .26$	70.61 (27.19)	0.370	60.77 (12.93)	0.194	59.17 (11.77)	0.136
	$ICC = .10$	60.24 (12.45)	0.170	59.41 (11.58)	0.152	58.42 (11.22)	0.108
	$ICC = .05$	73.09 (20.97)	0.464	89.20 (29.85)	0.658	77.78 (31.76)	0.450
T_{RADF}	$ICC = .50$	143.50 (45.63)	0.988	144.83 (40.46)	0.990	142.43 (42.71)	0.986
	$ICC = .26$	144.59 (46.84)	0.986	144.95 (40.85)	0.986	142.84 (42.32)	0.984
	$ICC = .10$	143.50 (45.63)	0.988	144.83 (40.46)	0.990	142.43 (42.71)	0.986
	$ICC = .05$	86.07 (23.38)	0.682	137.14 (40.21)	0.980	143.45 (41.21)	0.982
T_{CRADF}	$ICC = .50$	56.91 (7.23)	0.008	57.41 (6.45)	0.006	56.84 (7.05)	0.006
	$ICC = .26$	57.04 (7.33)	0.012	57.40 (6.55)	0.008	56.94 (6.96)	0.006
	$ICC = .10$	56.91 (7.23)	0.008	57.41 (6.45)	0.006	56.84 (7.05)	0.006
	$ICC = .05$	45.03 (6.33)	0.00	56.04 (6.75)	0.004	57.12 (6.66)	0.006

Note: Empirical Type I error rates in the interval [.028,.079] shown in bold.

Table 4.9: T_{RML}, T_{RADF} and T_{CRADF} performance, $df = 135$

		Group Sizes					
		10		30		50	
		Mean (sd)	Rej	Mean (sd)	Rej	Mean (sd)	Rej
$J = 500$							
T_{RML}	$ICC = .50$	137.54 (17.49)	0.090	138.66 (16.87)	0.080	137.05 (16.46)	**0.062**
	$ICC = .26$	141.22 (25.47)	0.154	139.46 (17.94)	0.100	137.56 (16.78)	**0.064**
	$ICC = .10$	167.56 (48.68)	0.400	143.89 (22.88)	0.174	140.01 (18.57)	0.110
	$ICC = .05$	179.15 (35.42)	0.670	160.21 (45.33)	0.350	146.69 (26.88)	0.220
T_{RADF}	$ICC = .50$	195.17 (29.99)	0.866	198.46 (28.92)	0.904	197.07 (29.26)	0.886
	$ICC = .26$	194.25 (29.28)	0.860	198.83 (29.28)	0.898	197.13 (28.95)	0.886
	$ICC = .10$	190.42 (27.82)	0.864	198.70 (29.68)	0.892	197.55 (28.90)	0.890
	$ICC = .05$	180.78 (25.25)	0.758	197.38 (29.59)	0.898	197.89 (29.44)	0.892
T_{CRADF}	$ICC = .50$	139.55 (15.39)	**0.056**	141.30 (14.68)	**0.078**	140.57 (15.00)	**0.064**
	$ICC = .26$	139.10 (15.14)	**0.064**	141.48 (14.87)	**0.078**	140.61 (14.87)	**0.060**
	$ICC = .10$	137.17 (14.36)	**0.038**	141.39 (15.03)	0.086	140.83 (14.80)	**0.060**
	$ICC = .05$	132.13 (13.57)	0.008	140.72 (14.97)	**0.078**	140.99 (14.99)	**0.070**
$J = 200$							
T_{RML}	$ICC = .50$	143.36 (18.4)	0.158	142.22 (17.3)	0.124	140.87 (16.61)	0.1
	$ICC = .26$	159.57 (41.28)	0.344	144.19 (18.54)	0.14	141.91 (17.3)	0.112
	$ICC = .10$	207.24 (40.85)	0.852	159.36 (34.52)	0.368	148.33 (22.15)	0.216
	$ICC = .05$	220.39 (38.75)	0.946	180.52 (56.55)	0.59	172.92 (51.39)	0.46
T_{RADF}	$ICC = .50$	483.47 (113.75)	1.00	491.95 (111.34)	1.00	485.81 (98.10)	1.00
	$ICC = .26$	487.13 (110.77)	1.00	491.28 (108.80)	1.00	487.08 (101.72)	1.00
	$ICC = .10$	463.15 (97.98)	1.00	484.40 (103.82)	1.00	490.88 (117.35)	1.00
	$ICC = .05$	353.32 (69.62)	1.00	436.21 (127.34)	0.998	491.01 (119.39)	1.00
T_{CRADF}	$ICC = .50$	138.98 (9.26)	0.004	139.80 (8.87)	0.002	139.51 (8.27)	0.00
	$ICC = .26$	139.35 (9.05)	0.00	139.79 (8.74)	0.002	139.53 (8.59)	0.00
	$ICC = .10$	137.43 (8.79)	0.00	139.23 (8.87)	0	139.5 (9.59)	0.002
	$ICC = .05$	125.79 (8.89)	0.00	133.48 (14.03)	0.00	139.51 (9.53)	0.004
$J = 100$							
T_{RML}	$ICC = .50$	157.21 (23.39)	0.372	150.08 (18.19)	0.232	149.30 (18.01)	0.208
	$ICC = .26$	200.01 (57.33)	0.656	155.98 (23.37)	0.330	151.80 (19.63)	0.258
	$ICC = .10$	267.82 (46.44)	0.998	205.56 (53.38)	0.762	173.39 (41.35)	0.506
	$ICC = .05$	209.04 (39.39)	0.886	250.46 (49.47)	0.976	218.84 (53.80)	0.844

Note: empirical Type I error rates in the interval [.028,.079] shown in bold.

CHAPTER 5

Summary and Conclusion

As surveys of the classroom environment have gained traction as components of teacher evaluation portfolios, there has been an increased amount of attention paid to using multilevel factor analyses to explore hypotheses about the measurement structure of between-classroom phenomena. The segregating method has many theoretical benefits. It allows for the separate testing and identification of measurement models at the between level and within level. This is a key advantage over approaches that simultaneously fit models to Σ_B and Σ_W, since many studies have found that the simultaneous testing of between and within models can make diagnosing sources of model misfit difficult (e.g. Hox, 2010; Ryu & West, 2009; Yuan & Bentler, 2007).

The current study, however, clarifies an important characteristic of the segregating method for applied research. Namely, at ICC and sample size configurations likely to be encountered when data about the classroom environment is collected by surveying students, the commonly used Maximum Likelihood test statistic obtained from the segregating method is likely not asymptotically distributed as central chi-square variates under the null hypothesis. This suggests that the reliance on the ML test statistic can result in unwarranted model modifications or revisions. The current study used an illustrative example and a simulation study to investigate the performance of test statistics under conditions likely to be encountered in using the segregating method in applied research on classroom climate. The results reflect some general patterns that are worth noting here. As with any simulation

study, caution should be used in generalizing these results to other conditions not included in the study. More work would be needed to investigate how other conditions, such as differences in factor loadings, alternative (non-normal) distributions, and imbalanced group-sizes, would influence test statistic performance.

Inferences about model fit based on T_{ML} can lead to invalid conclusions about the between-classroom factor structure

At ICC and sample size configurations likely to be encountered when data about classroom climate are collected by surveying students, T_{ML} is not asymptotically distributed as central chi-square variate under the null hypothesis. The Tripod Survey example demonstrated that inferences based on T_{ML} can lead to invalid conclusions about the between-classroom factor structure when ICCs are low and within-classroom sample sizes are small. The simulation study shows the extent to which T_{ML} can be inflated. In some conditions, the mean of the test statistic was nearly 20 times too large, and every model was rejected. Thus, T_{ML} is very poorly behaved in general and should not be used to make inferences about between-level measurement models. While beyond the scope of the current study, these results suggest that, beyond issues with assessing model fit, Maximum Likelihood estimation would result in biased standard errors. This is because, typically speaking, estimated standard errors are computed using $\dot{\sigma}(\hat{\theta})^T W_{ML} \dot{\sigma}(\hat{\theta})$ (e.g. Bentler, 2006). Based on results presented elsewhere (Hox, 2010; Yuan & Bentler, 2007), it is anticipated that the parameter estimates themselves will be unbiased. Further research is needed to address these issues.

T_{CRADF} can provide valid inferences, provided the number of groups is sufficiently large

While both the rescaled and residual based test statistics show evidence supporting the hypothesis that they would converge to the appropriate distribution regardless

of item ICC and within-group sample size, only T_{CRADF} showed adqueate performance over a wide range of conditions. T_{RADF} showed a tendency to over-reject correct models for all but the largest samples, consistent with findings in conventional factor analysis. T_{RML} , too, over-rejected correct models, particularly for small between-level sample sizes. Only T_{CRADF} is recommended for use in conjunction with the segregating methodology. However, caution should be used with small samples, where T_{CRADF} shows a tendency to under-reject correct models.

References

Agresti, A., & Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, *52*(2), 119–126.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238.

Bentler, P. M. (2006). EQS 6 structural equations program manual [Computer software manual]. Los Angeles: BMDP Statistic Software.

Bentler, P. M., & Dudgeon, P. (1996). Covariance structure analysis: Statistical practice, theory, and directions. *Annual Review of Psychology*, *47*(1), 563–592.

Bentler, P. M., & Yuan, K. H. (1999). Structural equation modeling with small samples: Test statistics. *Multivariate Behavioral Research*, *34*(2), 181–197.

Bollen, K. (1989). *Structural equations with latent variables.* New York: Wiley.

Browne, M. (1974). The analysis of patterned correlation matrices by generalized least squares. *British Journal of Mathematical and Statistical Psychology*, *30*(1), 113–124.

Browne, M. (1982). Covariance structures. In D. M. Hawkins (Ed.), *Topics in applied multivariate analysis* (pp. 72–141). Cambridge: Cambridge University Press.

Browne, M. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, *37*, 1–21.

Cronbach, L. (1976). *Research on classrooms and schools: Formulation of questions, design and analysis.* Stanford University: Stanford Evaluation Consortium.

Cronbach, L., & Webb, N. (1975). Between-class and within-class effects in a

reported aptitude x treatment interaction: Reanalysis of a study by g. l. anderson. *Journal of Educational Psychology*, *67*, 717–724.

Curran, P. S., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, *1*, 16–29.

Doherty, K. M., & Jacobs, S. (2013). *Connect the dots:using evaluations of teacher effectiveness to inform policy and practice*. Washington, DC: National Council on Teacher Quality.

Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014). Student ratings of teaching quality in primary school: Dimensions and prediction of student outcomes. *Learning and Instruction*, *29*, 1–9.

Ferguson, R. (2010). Student perceptions of teaching effectiveness. *Harvard University*.

Ferguson, R. (2012). Can student surveys measure teaching quality? *Phi Delta Kappan*, *94*(3), 24–28.

Foldnes, N., Foss, T., & Olsson, U. H. (2012). Residuals and the residual-based statistic for testing goodness of fit of structural equation models. *Journal of Educational and Behavioral Statistics*, *37*(3), 367–386.

Follman, J. (1992). Secondary school students' ratings of teacher effectiveness. *The High School Journal*, *14*(4).

Goldstein, H. (2003). *Multilevel statistical models*. New York: John Wiley & Sons.

Hallquist, M., & Wiley, J. (2013). MplusAutomation: Automating mplus model estimation and interpretation [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=MplusAutomation` (R package version 0.6-2)

Härnqvist, K. (1978). Primary mental abilities at collective and individual levels. *Journal of Educational Psychology*, *70*(5), 706.

Holfve-Sabel, M.-A., & Gustafsson, J.-E. (2005). Attitudes towards school, teacher, and classmates at classroom and individual levels: An application of two-level confirmatory factor analysis. *Scandinavian Journal of Educational Research*, *49*(2), 187–202.

Hox, J., & Maas, C. (2004). Multilevel structural equation models: The limited information approach and the multivariate multilevel approach. In *Recent developments on structural equation models* (pp. 135–149). Netherlands: Springer.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications*. Psychology Press.

Hu, L., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin*, *112*(2), 351–362.

Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika*, *32*(4), 443–482.

Julian, M. W. (2001). The consequences of ignoring multilevel data structures in nonhierarchical covariance modeling. *Structural Equation Modeling*, *8*(3), 325–352.

Kaplan, D., & Elliott, P. R. (1997). A didactic example of multilevel structural equation modeling applicable to the study of organizations. *Structural Equation Modeling: A Multidisciplinary Journal*, *4*(1), 1–24.

Kunter, M., Tsai, Y.-M., Klusmann, U., Brunner, M., Krauss, S., & Baumert, J. (2008). Students' and mathematics teachers' perceptions of teacher enthusiasm and instruction. *Learning and Instruction*, *18*(5), 468–482.

Lee, S.-Y. (1990). Multilevel analysis of structural equation models. *Biometrika*, *77*(4), 763–772.

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13–22.

Longford, N., & Muthén, B. (1992). Factor analysis for clustered observations.

Psychometrika, *57*(4), 581–597.

Magnus, J. R., & Neudecker, H. (1988). *Matrix differential calculus with applications in statistics and econometrics*. Wiley.

Mair, P., & Wu, E. (2012). REQS: R/EQS interface [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=REQS` (R package version 0.8-12)

Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J., Abduljabbar, A. S., & Köller, O. (2012). Classroom climate and contextual effects: conceptual and methodological issues in the evaluation of group-level effects. *Educational Psychologist*, *47*(2), 106–124.

Mathisen, G. E., Torsheim, T., & Einarsen, S. (2006). The team-level model of climate for innovation: A two-level confirmatory factor analysis. *Journal of Occupational and Organizational Psychology*, *79*(1), 23–35.

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*(2), 215–232.

Muthén, B., & Asparouhov, T. (2011). Beyond multilevel regression modeling: Multilevel analysis in a general latent variable framework. *Handbook of Advanced Multilevel Analysis*, 15–40.

Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal likert variables. *British Journal of Mathematical and Statistical Psychology*, *38*(2), 171–189.

Muthén, B., & Kaplan, D. (1992). A comparison of some methodologies for the factor analysis of non-normal likert variables: A note on the size of the model. *British Journal of Mathematical and Statistical Psychology*, *45*(1), 19–30.

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement*, *28*(4), 338–354.

Muthen, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, *22*(3), 376–398.

Muthén, L. K., & Muthén, B. O. (2010). Mplus: Statistical analysis with latent variables: User's guide [Computer software manual]. Los Angeles: Muthén & Muthén.

Powell, D. A., & Schafer, W. D. (2001). The robustness of the likelihood ratio chi-square test for structural equation models: A meta-analysis. *Journal of Educational and Behavioral Statistics*, *26*(1), 105–132.

R Core Team. (2013). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Rabe-Hesketh, S., Skrondal, A., & Zheng, X. (2007). *Multilevel structural equation modeling*. Elsevier.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Sage.

Reise, S. P., Ventura, J., Nuechterlein, K. H., & Kim, K. H. (2005). An illustration of multilevel factor analysis. *Journal of Personality Assessment*, *84*(2), 126–136.

Rosenberg, S. L. (2009). *Multilevel validity: Assessing the validity of school-level inferences from student achievement test data*. Unpublished doctoral dissertation, The University of North Carolina, Chapel Hill.

Ryu, E., & West, S. G. (2009). Level-specific evaluation of model fit in multilevel structural equation modeling. *Structural Equation Modeling*, *16*(4), 583–601.

Satorra, A., & Bentler, P. M. (1988). *Scaling corrections for chi-square statistics in covariance structure analysis* (Vol. 1).

Steiger, J. H., & Lind, J. C. (1980). *Statistically based tests for the number of common factors* (Vol. 758). Iowa City, IA.

Toland, M. D., & De Ayala, R. (2005). A multilevel factor analysis of students evaluations of teaching. *Educational and Psychological Measurement*, *65*(2), 272–296.

Van Horn, M. L. (2003). Assessing the unit of measurement for school climate through psychometric and outcome analyses of the school climate survey. *Educational and Psychological Measurement*, *63*(6), 1002–1019.

Worrell, F. C., & Kuterbach, L. D. (2001). The use of student ratings of teacher behaviors with academically talented high school students. *Prufrock Journal*, *12*(4), 236–247.

Yuan, K.-H., & Bentler, P. M. (1998). Normal theory based test statistics in structural equation modelling. *British Journal of Mathematical and Statistical Psychology*, *51*(2), 289–309.

Yuan, K.-H., & Bentler, P. M. (2002). On normal theory based inference for multilevel models with distributional violations. *Psychometrika*, *67*(4), 539–561.

Yuan, K.-H., & Bentler, P. M. (2006). Asymptotic robustness of standard errors in multilevel structural equation models. *Journal of Multivariate Analysis*, *97*(5), 1121–1141.

Yuan, K.-H., & Bentler, P. M. (2007). Multilevel covariance structure analysis by fitting multiple single-level models. *Sociological Methodology*, *37*(1), 53–82.

Yuan, K.-H., & Jennrich, R. I. (1998). Asymptotics of estimating equations under natural conditions. *Journal of Multivariate Analysis*, *65*(2), 245–260.

Zyphur, M. J., Kaplan, S. A., & Christian, M. S. (2008). Assumptions of cross-level measurement and structural invariance in the analysis of multilevel data: Problems and solutions. *Group Dynamics: Theory, Research, and Practice*, *12*(2), 127–140.