CRESST REPORT 829

TESTING THE ASSUMPTION OF CROSS-LEVEL MEASUREMENT INVARIANCE IN MULTILEVEL MODELS: EVIDENCE FROM SCHOOL AND CLASSROOM ENVIRONMENT SURVEYS

JUNE, 2013

Jonathan Schweig

National Center for Research
on Evaluation, Standards, & Student Testing
UCLA | Graduate School of Education & Information Studies

Testing the Assumption of Cross-Level Measurement Invariance in Multilevel Models: Evidence from School and Classroom Environment Surveys

CRESST Report 829

Jonathan Schweig
CRESST/University of California, Los Angeles

June 2013

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
300 Charles E. Young Drive North
GSE&IS Bldg., Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

Copyright © 2013 The Regents of the University of California.

The work reported herein was supported by grant number 52306 from the Bill and Melinda Gates Foundation with funding to the National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Part of this research is made possible by a pre-doctoral advanced quantitative methodology training grant (#R305B080016) awarded to UCLA by the Institute of Education Sciences of the US Department of Education.

The findings and opinions expressed in this report are those of the authors and do not necessarily reflect the positions or policies of the Bill and Melinda Gates Foundation or the US Department of Education.

I would like to thank Joan Herman, Jia Wang, and Noelle Griffin for their support of this study; and thank Felipe Martinez, Li Cai, and Peter Bentler, for reviewing the earlier version of this report and for their thoughtful feedback. The author is grateful to the North Carolina Education Research Data Center for data. My special thanks go to Laquita Stewart and Fred Moss. Without their assistance, this project would not have been possible.

To cite from this report, please use the following as your APA reference: Schweig, J. (2013). *Testing the assumption of cross-level measurement invariance in multilevel models: Evidence from school and classroom environment surveys* (CRESST Report 829). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

TABLE OF CONTENTS

Abstract 1
Introduction 1
Theoretical Framework 2
  Statistical Background 3
  Cross-Level Invariance Implied by Single-Level Factor Analyses 5
  Implications for Assuming Cross-Level Invariance in Policy and Practice 7
Methods 8
  Sample and Data Sources 8
  Analytic Approach 9
Results 10
  What is the Factorial Structure of the WCS Survey and the Tripod Survey? Is There Empirical Evidence to Support the Assumption of Cross-Level Measurement Invariance in Either Case? 10
  What Are the Consequences of Ignoring the Multilevel Structure and Conducting a Factor Analysis on the Disaggregated Data? 17
  What Are the Consequences of Ignoring the Multilevel Structure and Conducting a Factor Analysis on the Aggregated Data (Group Means)? 19
Summary 21
  There can be Significant Differences in Factorial Structure Across Levels 21
  Analysis of the Total covariance Matrix can Distort Perception of the Between-level Factorial Structure 21
  Analysis of the Group Mean Covariance Matrix can Distort Perception of Both the Between-level and Within-level Factorial Structures 22
Other Issues and Additional Questions 23
  Assumption of Reflective Aggregation 23
  Use of Linear Composites as Proxies for Latent Variables 23
  Omitted Levels in the Analysis 24
Factor Extraction and Model Fit 24
References 26

TESTING THE ASSUMPTION OF CROSS-LEVEL MEASUREMENT INVARIANCE IN MULTILEVEL MODELS: EVIDENCE FROM SCHOOL AND CLASSROOM ENVIRONMENT SURVEYS

Jonathan D. Schweig
CRESST/University of California, Los Angeles

Abstract

Measures of classroom and school environments are a central component of policy efforts that assess school and teacher quality. These measures are often formed by aggregating individual survey responses to form group-level measures, and assume an invariant measurement model holds at both the individual and group level. This paper explores the tenability of this assumption by applying multilevel factor analysis to two well-known surveys: the Working Conditions Survey, which assesses school environments, and the Tripod Classroom Environment Survey. The examples illustrate the consequences of using common factor analytic methods that assume cross-level invariance. Importantly, distorted perceptions of factorial structure can obscure the assessment of intervention effectiveness on key classroom outcomes, or the role of classrooms as mediators of educational interventions.

Introduction

As school districts strive to create comprehensive programs to appraise teaching quality and teacher performance, measures of the school and classroom environments have become increasingly important. Information about classroom and school environments is used in a variety of policy contexts and for a variety of purposes. Firstly, it can be used for teacher and school evaluation. Memphis bases 5% of a teacher evaluation on student surveys. By 2013, 10% of teacher evaluation in Chicago public schools will be based on student surveys (Butrymowicz, 2012). In New York City, teacher and parent surveys about the school environment can account for up to 15% of a school's score on its annual Progress Report ("NYC School Survey", n.d.).

Secondly, information about school and classroom environment can be used to predict important outcomes, such as student achievement and teacher retention. Preliminary results from the Measuring Effective Teaching (MET) project, for example, demonstrated relationships between students' perceptions of classroom environment and estimates of teachers' value added (VAM) scores (Bill & Melinda Gates Foundation, 2010). Ladd (2011) discussed how links between teacher mobility and working conditions can be used to develop and test teacher retention policies. Better understanding how targeted improvements in working conditions may improve retention is particularly critical for schools serving high-poverty, low-achieving student populations, where teacher turnover rates may be as high as 50% (Ingersoll, 2001).

Data about school and classroom environments are often collected by administering surveys to teachers and students who function as "raters" of the environments in which they work and study. These individual ratings are aggregated to form group-level variables, and inferences are then made about group qualities (Chan, 1998; Kozlowski & Klein, 2000).

While the use of this type of aggregated group-level variable is intuitively appealing, the validity of these aggregated variables entails a complex and nuanced set of assumptions. Substantively, it is assumed that the aggregates refer to the same constructs as the individual responses. Statistically, it is assumed that there is cross-level invariance in the measurement model (Bliese, 2000); that is, there is invariance in the measurement structure across the individual (within-group) level and the between-groups level.

Cross-level measurement invariance imposes strict constraints on the measurement model that may not be met in empirical data (Zyphur, Kaplan, & Christian, 2008). In much of the educational research and policy literature, however, cross-level invariance is assumed rather than explored. Glick (1985) cautions against relying on assumed composition rules to create environmental variables, and Reise, Ventura, Nuechterlein, and Kim (2005) noted that assuming cross-level measurement invariance is potentially "substantively misleading." (p. 130). Cronbach (1976) cautioned that a researcher might need "one set of factors for his between-groups theory and another set of factors for his within-groups theory. To be sure, he may find that the two sets of constructs coincide, but that is a possibility to be evaluated, not assumed" (p. 203). Reflecting on the same issue from a different research tradition, Raudenbush and Bryk (2002) cautioned that using both individual level and aggregated predictors in multilevel regression models imposes the assumption that the variables refer to the same construct.

This paper presents two empirical examples that illustrate that the consequences for policy and practice that may arise from assuming cross-level measurement invariance. In doing so, this paper expands on recent work (Marsh et al., 2009; Zyphur, Kaplan, & Christian, 2008) that calls attention to the importance of finding empirical evidence to support the cross-level invariance assumption.

Theoretical Framework

Many widely used surveys of school and classroom environments assume a specific underlying measurement model. At the student or teacher level, there is assumed to be measurement error among the survey items, so that variance among the items is caused by unobserved (latent) differences among individual teachers or students.

At the between-groups level, it is assumed that students or teachers are objective raters of the environments in which they study, and that variance between raters within the same school or

classroom is attributable to sampling error and represents "noise." On the other hand, averaging over individual raters, variance between schools represents actual variance in the quality of working conditions, or variance between classrooms represents true variance in the quality of classrooms.

Under these assumptions, it is appropriate to use a latent trait model where the group qualities themselves are conceived of as effects-indicated latent variables (Bollen & Lennox, 1991; Marsh et al., 2009). In an effects-indicated model, it is assumed that a latent variable causes variance in the indicators. In the case of school or classroom environment surveys, it is assumed that unobserved, latent aspects of the school or classroom environment cause variance between individual schools or classrooms. This is sometimes referred to as reflective aggregation (Kozlowski & Klein, 2000; Marsh et al., 2009).

It is important to distinguish an effects-indicated measurement model from another possible model, a so-called composite model where indicators are formed by making linear combinations of indicator variables (e.g., Bollen & Bauldry, 2011). An example of this sort of composite indicator would be socioeconomic status (SES). A set of indicators of SES can be used as a weighted combination to describe an individual student's SES. It is not a claim that an individual student has a latent SES that causes variance among the indicators. Individual student SES indices can then be aggregated to form a school-level variable (for example, Raudenbush & Bryk, 2002). Composite variables of this type implicitly impose cross-level measurement invariance because there is only a single indicator for each individual (Marsh et al., 2009). Survey-based indicators of school and classroom environments are rarely conceived of as composite variables in this way, and so this type of model and this type of cross-level measurement invariance are not the focus of this paper.

Statistical Background

In the reflective aggregation model typically underlying school and classroom environment surveys, the assumption that group means refer to the same constructs as individual responses implies a two-level measurement model with cross-level factorial invariance (Marsh et al., 2009). Specifically, the factor structure is assumed to be configurally and metrically invariant (Meredith, 1993)—meaning that at both levels the same number of factors are found, the same items load onto the same factors, and the strength of association between these items and the underlying factors is the same. To understand this assumption, it is necessary to briefly review the basic factor analysis model, and its two-level counterpart, multilevel factor analysis (Bentler & Liang, 2003; Muthén, 1994). The basic factor model (Bollen, 1989) can be expressed

$$y = \Lambda\eta + \varepsilon \quad (1)$$

where y is a p-variate vector of observed scores measuring η. η is an $m \times 1$ vector of latent variable scores on m factors, assumed to be normally distributed with 0 expectation. Λ is an $p \times m$ matrix of factor loadings. ε represents an $p \times 1$ vector of residuals, which are assumed to be identically and independently distributed. The covariance structure implied by this model can then be expressed

$$\Sigma(\theta) = \Lambda\Phi\Lambda^T + \Theta \quad (2)$$

where $\Sigma(\theta)$ is a $p \times p$ matrix of model-implied covariances. Φ is an $m \times m$ matrix of factor variances and covariances, and Θ is an $p \times p$ diagonal matrix containing error variances.

Factor analytic procedures based on Equation 2 assume that the observations are independent. When individuals are associated with groups (teachers with schools, students with classrooms) this independence assumption is likely to be violated. There are several equivalent models that account for the fact that observations are nested in groups (e.g., Goldstein, 2003; Muthén, 1994; Rabe-Hesketh, Skrondal, & Zheng, 2007). Many of these formulations are based on a score decomposition model articulated by Cronbach and Webb (1975):

$$y_{ij} = y_j + \left(y_{ij} - y_j\right) \quad (3)$$

where y_{ij} is a p-variate vector of observed scores for individual i in group j. y_{ij} can be decomposed into orthogonal between groups (y_j), and within groups $\left(y_{ij} - y_j\right)$ components. Because the between and within components are orthogonal, it is possible to express the covariance matrix of the observed scores as a sum of between and within covariance matrices:

$$\Sigma_{\mathrm{T}} = \Sigma_{\mathrm{B}} + \Sigma_{\mathrm{W}} \quad (4)$$

The covariance structure model presented in Equation 2 can then be extended to express a multilevel covariance structure model:

$$\Sigma_{\mathrm{T}} = \Lambda_B\Phi_B\Lambda_B^T + \Lambda_W\Phi_W\Lambda_W^T + \Theta_B + \Theta_W \quad (5)$$

The corresponding multilevel measurement model is:

$$y_{ij} = \Lambda_B\eta_B + \Lambda_W\eta_W + \mathrm{u_j} + \varepsilon_{\mathrm{ij}} \quad (6)$$

There are two random effects here—a between-groups random effect $\mathrm{u_j}$, and a within-group random effect $\varepsilon_{\mathrm{ij}}$. There are also two sets of factor loadings (Λ_{B} and Λ_{W}) and two latent variables (η_B and η_W). If there is measurement invariance and the factor loadings are invariant across levels (i.e., $\Lambda_{\mathrm{B}} = \Lambda_{\mathrm{W}} = \Lambda$), Equation 8 can be rewritten:

$$y_{ij} = \Lambda(\eta_B + \eta_W) + \mathrm{u_j} + \varepsilon_{\mathrm{ij}} \quad (7)$$

$$= \Lambda\eta_{ij} + \mathrm{u_j} + \varepsilon_{\mathrm{ij}}$$

In this way, the latent trait of individual i in group j can be expressed as a sum of orthogonal between and within latent components: $\eta_{ij} = \eta_B + \eta_W$. This means that with cross-level measurement invariance, the latent variable can be conceived of as a decomposition, just as the observed variable in Equation 3 (Marsh et al., 2009). On the other hand, in the more general case where $\Lambda_{\mathrm{B}} \neq \Lambda_{\mathrm{W}}$, these simplifications are not possible.

Cross-Level Invariance Implied by Single-Level Factor Analyses

One way in which cross-level measurement invariance is assumed in the policy and research literature on school and classroom environments is to conduct single-level factor analyses on data that contains both within-groups and between-groups variance. Factor analyses that do not model both the between and within factor structures of hierarchically structured data *de facto* impose invariance constraints on the factor structure (Zyphur, Kaplan, & Christian, 2008). This is because when a single-level factor analysis is conducted, either on the disaggregated responses or on the group means, only one Λ matrix is estimated, and the η matrix can take only one form.

Perhaps the most commonly used single-level approach is to conduct factor analysis on the total disaggregated covariance matrix (e.g., Ladd, 2011; Ryan & Patrick, 2001):

$$S_T = \frac{\sum_{j=1}^{J}\sum_{i=1}^{N_j}(y_{ij} - y_{..})(y_{ij} - y_{..})'}{N - 1} = \hat{\Sigma}_{\mathrm{T}} \quad (8)$$

where y_{ij} is the *p*-variate vector of observed scores for individual i in group j, $y_{..}$ is a vector of item grand means, and N is the total sample size. No group information is included here; however, S_T yields consistent and unbiased estimates of the population matrix Σ_{T} (Muthén, 1994). Equation 4 implies that for situations where individuals are nested in groups, conducting factor analyses on $\hat{\Sigma}_{\mathrm{T}}$ conflates within and between sources of variance, unless either $\Sigma_{\mathrm{B}} = 0$ or $\Sigma_{\mathrm{W}} = 0$. This conflation of variance sources can bias parameter estimates (Preacher, Zyphur , & Zhang, 2010), and can lead to substantively misleading inferences about relationships between indicators, or about relationships with external variables (Reise et al., 2005).

Another single-level approach is to conduct a single-level factor analysis on the unweighted group means (e.g., Hoy & Clover, 1986; Klinger, Rogers, Anderson, Poth, & Kalman, 2006). Analyses that are conducted on the unweighted group-means employ the covariance matrix:

$$S_B = \frac{\sum_{j=1}^{J}(y_j - y_{..})(y_j - y_{..})'}{J-1} \quad (9)$$

where y_j is a *p*-variate vector of means for group, $y_{..}$ is a vector of grand means, and J is the number of groups. There are two issues with this formulation of S_B. Each group is given the same weight in the estimation of the group-covariance matrix, regardless of the number of individuals in that group. Additionally, S_B is not a consistent and unbiased estimator of the population matrix Σ_B. In addition to the bias introduced by weighting each group equally, S_B itself contains both between and pooled-within variance sources (Muthén, 1994).

In addition to the concerns of bias that arise from the conflation of variance from within-group and between-groups sources, the cross-level measurement invariance assumption that $\Lambda_B = \Lambda_W = \Lambda$ is a particularly strong one. Cronbach (1976) and Harnqvist (1978) showed that factorial structures could vary at group and individual levels, and that often times, fewer factors may be found at the between-groups level than at the within-groups level. Recent empirical examples (Holfve-Sabel & Gustaffson, 2005; Hox, 2010) have provided additional evidence to support this idea.

Even if the number of factors at the within-group level and the between-groups level are the same, there is research (Yuan & Bentler, 2007; Zyphur, Kaplan, & Christian, 2008) showing item loadings may not be the same across levels, and items may load on different factors across levels of analysis.

Another issue concerns the patterns of covariance among the latent factors. When factor analyses are performed on either S_B or S_T, only one Φ matrix is estimated, the relationships among factors as assumed to be the same across levels. Muthén (1997, p. 455), stated, "A frequent shortcoming when ignoring the multilevel structure of the data is not what is misestimated, but what is not learned" (as cited in Zyphur, Kaplan, & Christian, 2008, p. 127). Estimating only one Φ matrix means that you lose the opportunity to learn if the relationships encoded in Φ_W differ from those encoded in Φ_B in substantively meaningful ways. For example, if factors are more strongly associated at one level than the other.

There are other statistical issues that arise in using either the disaggregated data or the unweighted group-mean covariance matrix as the basis for a factor analysis on clustered data. Research shows, for example, that using the disaggregated data covariance matrix can overestimate factor variances and covariances, underestimate standard errors, and inflate chi-square statistics (Julian, 2001), leading to high Type-I error rates.

Implications for Assuming Cross-Level Invariance in Policy and Practice

One of the most pervasive uses of factor analysis in policy research is to justify the formation of linear composites. This practice is sometimes called rank reduction, and is described in many sources (e.g., Alwin, 1973; Bollen & Lennox, 1991; Cronbach, 1976). Note that the linear composite that results from rank reduction is distinct from a composite of the sort described in Bollen and Bauldry (2011) and referenced earlier. In the case of rank reduction, a linear composite is used as a proxy for a latent variable. It is still an underlying assumption that the composite has "conceptual unity" (Bollen & Bauldry, 2011, p. 4), and that variance in the indicators is caused by a common underlying latent variable.

In studies of school and classroom environments, the rank reduction process takes three steps. First, a factor analysis is performed and evidence is collected that a set of items measures a common latent variable. Second, scores on those items are averaged together to form a single score for an individual teacher or student. Third, individual scores are averaged together to form a school or classroom level variable.

Two recent examples from policy literature where factor analysis is used to justify the formation of linear composites in this way include Ryan and Patrick (2001), which investigated the relationship between classroom environment and student motivation and engagement, and Ladd (2011), which considers the relationship between teacher working conditions and teacher retention. In both cases, linear composites of aggregated variables are justified based on the results of exploratory factor analyses conducted on the disaggregated covariance matrix. By assuming cross-level measurement invariance in this way, there is a strong possibility that this approach could result in the formation of unsupported linear composites, and could result in obscured or spurious information about prediction and correlation among policy relevant constructs.

For example, this approach could result in identifying the wrong number of factors, or in associating items with the wrong factors altogether. In the context of school and classroom environments, it may be, for example, that items in a survey distinguish two psychological latent variables, such as the quality of academic support and instructional rigor at the individual level. But it is also conceivable that the two factors are indistinguishable at the group level and collapse into one broader academic factor. Researchers and policy makers who assume cross level invariance thus risk assuming they are working with two distinct dimensions of classroom quality, when in fact, they are not.

While the potential statistical, substantive, and policy consequences of assuming cross-level measurement invariance are well documented, there are few existing studies that explore

how these assumptions influence the analysis of empirical data. Marsh et al. (2009) called attention to the importance of testing the cross-level invariance assumption empirically. However, in their illustrative example, the invariance assumption held true. Zyphur et al. (2008) present a case where there is evidence for factorial non-invariance (different patterns of loadings), but that study did not present any cases where the number of factors differs across levels.

The purpose of the present study is to explore cross-level measurement invariance using two empirical examples, and to demonstrate the possible consequences that may arise for policy and practice when invariance is assumed. The first example comes from the Working Conditions Survey (New Teacher Center, 2009), which is a survey administered to measure aspects of school climate. The second comes from the Tripod Classroom Environment Survey (Ferguson, 2010), which is administered to measure aspects of classroom environment. These two surveys provide particularly salient examples for several reasons. First, both surveys are widely used to inform school policy decisions in the United States. Second, both surveys have an aggregated unit-of-analysis. For the Working Conditions Survey, the unit-of-analysis is the school; for the Tripod, the unit-of-analysis is the classroom. Lastly, in both surveys, it is an explicit measurement claim that variance between raters (teachers or students, respectively) constitutes error variance, and that variance between schools or classrooms represents true variance in environmental qualities. Using these two surveys, the following research questions were addressed:

1. What is the factorial structure of the Working Conditions Survey? What is the factorial structure of the Tripod Classroom Environment Survey? Is there empirical evidence to support the assumption of cross-level measurement invariance in either case?
2. What are the consequences of ignoring the multilevel structure and conducting a factor analysis on the disaggregated data? Are there differences between the factorial structure that arises as a result of this single-level analysis and the between-school or between-classroom factorial structure that is found using a multilevel analysis?
3. What are the consequences of ignoring the multilevel structure and conducting a factor analysis on the unweighted group means? Are there differences between the factorial structure that arises as a result of this single-level analysis and the between-school or between-classroom factorial structure that is found using a multilevel analysis?

Methods

Sample and Data Sources

The Working Conditions Survey. This survey was designed to assess teaching conditions at the school level. The sample data comes from the 2008 survey, administered to both teachers

and principals at schools in K-12 public and charter schools across the state of North Carolina. For this analysis, only surveys completed by teachers were considered, resulting in a data set with 88,936 individual teacher cases in 2,423 schools. Though the average school size is approximately 37 teachers, schools in this analysis range from 5 teachers to 146 teachers. The survey contains 36 items (Table 1) that measure five theoretical dimensions: Time, Decision Making, Leadership, Professional Development, and Facilities & Resources. There are two scales used in the survey. One has 5 points (1=*strongly disagree*, and 5=*strongly agree*) and is used for every item in the Time, Leadership, Professional Development and Facilities & Resources dimension. The other items also have 5 points (1=*no role at all*, and 5=*the primary role*) and are used in the Decision Making dimension.

The Tripod Classroom Environment Survey. The Tripod Survey assessment is designed to assess seven dimensions of teaching practice, often referred to as the "Seven C's": Caring, Captivating, Conferring, Clarifying, Challenging, Controlling, Consolidating. This version of the Tripod Survey contains 36 items (Table 1), and was administered in an urban school district in California in 2010. All items have 5-point scales (1=*totally untrue* and 5=*totally true*). The sample used in this analysis contained 6,386 students in 350 classrooms. The average classroom size was approximately 18 students, and the range was from 5 to 33 students.

Analytic Approach

In order to address the first research question, this paper follows the multilevel exploratory factor anlaysis (MEFA) procedure outlined by van de Vijver and Poortinga (2002) and Reise, Ventura, Nuechterlein and Kim (2005), which is based on a procedure first outlined by Muthén (1994). 1) The item Intraclass correlations (ICCs) were inspected in order to determine the amount of variance at the between-groups level to assess whether a multilevel factor analysis is warranted. 2) Maximum Likelihood estimates of the pooled-within level correlation matrix and between-groups level correlation matrix were obtained using Mplus version 6.11 (Muthén and Muthén, 2010). 3) Exploratory factor analysis (EFA) was then conducted on these two matrices separately. Factors were extracted using unweighted least squares (minres) factor analysis. Oblique (oblimin) rotation was used so that the factors were free to correlate.

In EFA, there is a long and rich literature on methods for determining the number of factors to retain (e.g., Fabrigar, Wegener, MacCallum, & Strahan, 1999; Ford, MacCallum, & Tait, 1986; Floyd & Widaman, 1995). Most of this literature is focused on the single-level context. There is relatively little research on these issues in multilevel EFA. Some studies, however, have suggested that features of the between-groups correlation matrix may result in the extraction of too many factors if maximum likelihood based approaches to factor selection (such as the

likelihood-ratio test statistic) are used (e.g., Briggs & MacCallum, 2003; Browne, MacCallum, Kim, Anderson, & Glaser, 2002; Preacher & MacCallum, 2002; Schmitt, 2011). Thus, this paper uses parallel analysis (Horn, 1965) in order to determine the number of factors to retain. Several studies (e.g., Crawford & Koopman, 1973; Hubbard & Allan, 1987; Humphreys & Montanelli, 1975; Schmitt, 2011) have shown that parallel analysis provides trustworthy estimates of the number of factors to retain in an exploratory factor anlaysis. D'Haenens, van Damme, & Onghena (2010a), used parallel analysis on the pooled-within and between-groups correlation matrices.

For the multilevel analyses, once an exploratory factor structure had been identified, two confirmatory factor analysis (CFA) models were fit to the data. The first constrained the configuration to be equal across levels. In other words, items loaded onto the same factors in the within-group and between-groups models. The second modeled the factor structures suggested by the exploratory analysis. The CFA included all item loadings greater than .3 (even cross-loading items), and restricted all other loadings to 0. In this way, CFA was used to confirm whether a model with different factorial configurations across levels fit the data better than a model with the same factorial configuration. The two models were subsequently compared by examining Δ_i, the difference in Akaike Information Criterion (AIC), which allows for the comparison of non-nested models (Burnham & Anderson, 2002). In the current study, $\Delta_i = AIC_{INVARIANCE} - AIC_{NONINVARIANCE}$. $\Delta_i$values larger than 10 indicate that the invariance model has almost no support in the data.

To address the second and third research questions, two additional exploratory factor analyses were conducted. These include an analysis on the total (disaggregated) covariance matrix, S_T (Equation 10), and an analysis on the unweighted group-mean covariance matrix, S_B (Equation 11). Both of these matrices were rescaled to be correlation matrices in the EFA analyses. These analyses investigate whether different factor structures would be extracted in those commonly misspecified cases. In order to apply a consistent and objective criterion to all of the analyses, decisions about how many factors to extract were based on parallel analysis.

Results

What is the Factorial Structure of the WCS Survey and the Tripod Survey? Is There Empirical Evidence to Support the Assumption of Cross-Level Measurement Invariance in Either Case?

ICCs (Table 1) range from around .10 to around .24 for the Working Conditions Survey, and from around .05 to .24 for the Tripod Survey. Though this shows that individual responses within clusters share a non-trivial amount of similarity, there is also variability in terms of how

much variance of each item is accounted for at the group level. Some items function better as indicators of group-level phenomenon than others. Overall, in both surveys, there is sufficient evidence that there is a non-trivial amount of between-level variance and that a multilevel factor analysis is warranted.

Table 1

Descriptive Statistics for the Working Conditions Survey and Tripod Survey

Working Conditions Survey				Tripod Survey			
Item	Mean	*SD*	ICC	Item	Mean	*SD*	ICC
TIME1[a]	3.36	1.36	0.16	CAPT1[f]	3.37	1.31	0.10
TIME2[a]	3.35	1.34	0.15	CAPT2[f]	3.61	1.27	0.18
TIME3[a]	3.18	1.35	0.12	CAPT3[f]	3.64	1.17	0.24
TIME4[a]	3.27	1.30	0.15	CAPT4[f]	3.75	1.22	0.15
TIME5[a]	3.07	1.36	0.16	CARE1[g]	3.62	1.13	0.14
FACR1[b]	3.81	1.20	0.12	CARE2[g]	3.17	1.25	0.11
FACR2[b]	3.85	1.26	0.16	CARE3[g]	3.69	1.11	0.15
FACR3[b]	3.96	1.18	0.17	CHAL1[h]	4.11	1.14	0.07
FACR4[b]	3.81	1.26	0.17	CHAL2[h]	3.78	1.02	0.10
FACR5[b]	3.97	1.16	0.13	CHAL3[h]	3.93	1.01	0.10
FACR6[b]	3.79	1.24	0.13	CHAL4[h]	3.93	1.08	0.12
FACR7[b]	3.94	1.21	0.24	CHAL5[h]	4.07	1.01	0.09
FACR8[b]	4.14	1.08	0.20	CHAL6[h]	3.92	1.01	0.11
LEAD1[c]	3.62	1.28	0.20	CHAL7[h]	3.91	1.03	0.15
LEAD2[c]	3.86	1.18	0.19	CHAL8[h]	3.96	1.01	0.12
LEAD3[c]	3.44	1.36	0.23	CLAR1[i]	4.00	1.03	0.14
LEAD4[c]	3.69	1.28	0.21	CLAR2[i]	3.74	1.01	0.11
LEAD5[c]	3.75	1.23	0.20	CLAR3[i]	3.34	1.18	0.06
LEAD6[c]	3.78	1.13	0.18	CLAR4[i]	3.89	1.04	0.17
LEAD7[c]	4.04	1.10	0.16	CLAR5[i]	3.81	1.05	0.14
LEAD8[c]	4.00	1.13	0.16	CONF1[j]	3.85	1.16	0.12
LEAD9[c]	3.98	1.12	0.14	CONF2[j]	2.64	1.16	0.11
LEAD10[c]	3.54	1.36	0.11	CONF3[j]	3.70	1.04	0.13
DECM1[d]	3.44	1.08	0.11	CONF4[j]	3.72	1.10	0.13
DECM2[d]	3.71	1.03	0.13	CONF5[j]	4.03	1.01	0.11

Working Conditions Survey				Tripod Survey			
Item	Mean	*SD*	ICC	Item	Mean	*SD*	ICC
DECM3 [d]	3.52	1.11	0.11	CONS1[k]	3.58	1.12	0.12
DECM4 [d]	2.05	1.15	0.14	CONS2[k]	4.07	0.98	0.13
DECM5 [d]	2.65	1.17	0.12	CONS3[k]	3.86	1.06	0.12
DECM6 [d]	2.09	1.06	0.15	CONS4[k]	3.80	1.09	0.13
DECM7 [d]	2.95	1.14	0.13	CONT1[l]	3.72	1.23	0.16
DECM8 [d]	2.64	1.11	0.11	CONT2[l]	3.55	1.27	0.16
PROF1 [e]	3.26	1.29	0.13	CONT3[l]	3.30	1.28	0.20
PROF2 [e]	3.63	1.17	0.10	CONT4[l]	3.62	1.26	0.21
PROF3 [e]	3.58	1.18	0.11	CONT5[l]	3.43	1.14	0.24
PROF4 [e]	3.49	1.21	0.11	CONT6[l]	3.85	1.11	0.26
PROF5 [e]	3.62	1.14	0.10	CONT7[l]	3.69	1.11	0.22

Note: [a]Time, [b]Facilities & Resources, [c]School Leadership, [d]Decision Making, [e]Professional Development, [f]Captivating, [g]Caring, [h]Challenging, [i]Clarifying, [j]Conferring, [k]Consolidating, [l]Controlling

For the Working Conditions Survey, parallel analysis suggested the extraction of 6 factors at the within level and 5 factors at the between level. In the within-school factor structure, items almost load cleanly into the dimensions that the Working Conditions Survey was designed to measure. Table 2 shows all rotated factor loadings. A sixth factor seems to arise from a bifurcation of the School Leadership factor into two sub-factors. One contains three items related specifically to performance evaluation; the other contains items about other aspects of school leadership, including administrative support for classroom discipline, clarity of communication with teachers and parents, and the fostering of a shared vision among members of the faculty and staff.

Table 2

Rotated Factor Loadings for the Working Conditions Survey: Multilevel Analysis

	Within-schools factor						Between-schools factor				
Item	1	2	3	4	5	6	1	2	3	4	5
TIME1[a]	0.02	0.00	0.12	-0.02	-0.03	**0.46**	0.27	0.25	**0.36**	-0.13	-0.09
TIME2[a]	-0.08	-0.02	0.00	0.01	0.11	**0.70**	**0.39**	0.01	**0.62**	0.02	**-0.31**
TIME3[a]	0.07	0.02	0.04	0.01	-0.05	**0.66**	**0.34**	0.24	**0.40**	0.10	-0.08
TIME4[a]	0.19	0.09	0.02	0.04	-0.08	**0.52**	0.05	0.25	**0.57**	0.13	0.14
TIME5[a]	-0.02	0.01	0.00	0.00	0.03	**0.74**	0.06	0.04	**0.91**	0.02	-0.19
FACR1[b]	0.01	0.07	**0.53**	0.01	0.07	0.05	**0.73**	0.07	-0.06	0.07	0.05
FACR2[b]	-0.02	-0.01	**0.73**	-0.01	0.05	-0.03	**0.82**	0.05	-0.13	-0.08	0.02
FACR3[b]	-0.01	0.01	**0.72**	0.02	0.02	-0.01	**0.66**	0.04	0.03	-0.10	0.07
FACR4[b]	0.06	0.04	**0.56**	0.01	0.02	0.05	**0.42**	0.04	0.20	0.05	0.15
FACR5[b]	0.00	0.00	**0.59**	0.01	0.01	0.02	**0.67**	0.02	-0.13	-0.05	0.03
FACR6[b]	0.01	0.01	**0.50**	0.02	0.03	0.13	**0.58**	0.07	0.07	-0.02	-0.10
FACR7[b]	0.19	-0.01	**0.38**	0.03	-0.02	0.07	**0.39**	0.25	-0.03	0.00	-0.07
FACR8[b]	0.27	-0.01	**0.36**	0.04	-0.05	0.08	**0.36**	**0.50**	-0.08	0.02	-0.01
LEAD1[c]	**0.49**	0.10	0.03	0.15	0.03	0.06	0.07	**0.32**	0.19	**0.33**	**0.36**
LEAD2[c]	**0.61**	0.02	0.03	0.13	0.06	0.00	0.09	**0.63**	0.00	0.27	0.12
LEAD3[c]	**0.86**	-0.01	0.01	-0.06	0.01	0.02	0.00	**1.02**	-0.04	0.01	-0.09
LEAD4[c]	**0.87**	-0.01	0.03	-0.03	-0.01	0.00	-0.07	**0.99**	0.08	0.01	-0.02
LEAD5[c]	**0.68**	0.05	0.00	0.14	0.02	0.04	0.00	**0.62**	0.12	0.24	0.22
LEAD6[c]	**0.43**	0.09	0.01	0.19	0.13	0.01	0.27	**0.41**	0.01	0.27	0.26
LEAD7[c]	-0.01	-0.01	0.01	**0.92**	-0.01	0.01	-0.02	-0.01	-0.01	**1.03**	-0.02
LEAD8[c]	-0.02	0.01	0.01	**0.92**	-0.01	0.01	0.00	0.00	-0.01	**1.01**	-0.02
LEAD9[c]	0.09	0.00	-0.01	**0.73**	0.08	0.00	0.07	0.1	-0.01	**0.86**	-0.02
LEAD10[c]	**0.43**	0.16	-0.02	0.14	0.17	0.05	0.19	**0.36**	0.14	**0.33**	0.27
DECM1[d]	-0.04	**0.62**	0.13	0.03	-0.02	-0.02	0.19	0.05	**0.48**	0.09	**0.34**
DECM2[d]	-0.02	**0.56**	0.11	0.08	-0.09	0.02	-0.06	0.02	**0.69**	0.12	**0.32**
DECM3[d]	-0.05	**0.53**	0.09	0.05	-0.06	0.02	-0.14	0.00	**0.83**	-0.04	0.19
DECM4[d]	0.00	**0.6**	-0.03	-0.02	0.15	0.02	**0.39**	0.09	**0.31**	0.12	0.25
DECM5[d]	-0.01	**0.58**	-0.07	-0.03	0.02	0.03	0.25	0.07	0.14	-0.07	**0.32**
DECM6[d]	0.18	**0.56**	-0.02	-0.04	0.00	0.01	**0.33**	**0.51**	-0.02	0.01	0.22
DECM7[d]	0.00	**0.64**	-0.06	-0.04	0.04	0.03	**0.38**	0.08	0.06	0.08	**0.36**

	Within-schools factor						Between-schools factor				
Item	1	2	3	4	5	6	1	2	3	4	5
DECM8 [d]	0.05	**0.61**	0.00	0.04	0.03	-0.01	**0.36**	0.16	0.05	0.16	**0.32**
PROF1 [e]	0.00	0.05	0.09	-0.03	**0.61**	-0.03	**0.82**	-0.1	0.02	0.00	0.05
PROF2 [e]	0.04	0.03	-0.02	0.04	**0.62**	0.09	**0.69**	0.01	0.18	0.17	-0.06
PROF3 [e]	-0.01	-0.01	-0.02	0.00	**0.80**	0.06	**0.78**	-0.04	0.10	0.14	-0.06
PROF4 [e]	0.02	0.01	0.12	0.00	**0.64**	-0.02	**0.75**	0.00	0.10	0.08	-0.03
PROF5 [e]	0.09	0.04	0.03	0.06	**0.63**	-0.03	**0.76**	0.10	-0.02	0.17	0.00

Note: [a]Time, [b]Facilities & Resources, [c]School Leadership, [d]Decision Making, [e]Professional Development. All loadings greater than .3 are shown in bold. Strongest loadings for each item are shaded.

The between-school structure of the Working Conditions Survey differs from the within-school structure (Table 2). There is considerably more cross-loading. In total, there are 11 items that load onto more than 1 factor. This indicates that the factor structure may be less well-defined at the school level than at the teacher level.

The factors are still interpretable, however. The Teacher Evaluation factor is similarly constituted at the school level as it is at the teacher level. However, the Professional Development factor is no longer distinguished as a factor. These items now load with 7 of the Facilities & Resources items and several Decision Making items onto a factor which can be interpreted as a more broadly defined resources factor, where quality professional development is conceived of as a school-wide resource. This is reasonable since PROF1 reads "Sufficient funds and resources are available to allow teachers to take advantage of professional development activities."

Another factor that concerns leadership and school safety seems to emerge at the group level. The two strongest loading items on this factor are LEAD3, ("The school leadership consistently enforces rules for student conduct.") and LEAD4 ("The school leadership support teachers' efforts to maintain discipline in the classroom.") Other items that deal with school discipline and safety (DECM6, which asks about the role teachers have in "Establishing and implementing policies and student discipline", and FACR8 which reads, "Teachers and staff work in a school environment that is safe.") also associate with this factor.

The Time items now load with several of the Decision Making items into a factor that concerns the relationship of school leadership to classroom-specific issues. The three strongest loading items on this factor are TIME5 ("The non-instructional time provided for teachers in my school is sufficient."), DECM2 (how large a role do teachers have in "Devising teaching

techniques") and DECM3 (how large a role do teachers have in "Setting grading and student assessment practices").

When the model with the cross-level configural non-invariance was compared to a model that imposed the same factor structure across levels, the change in AIC ($\Delta_i = 538.1$) suggests that there is little evidence to support the model with cross-level invariance. Overall, it is possible to conclude that there is strong evidence showing non-invariance across levels in the WCS.

For the Tripod Survey, parallel analysis suggests 5 factors at the within level, and 2 factors at the between level. For the within-classroom factorial structure, 20 of the first 29 items load onto a single factor (Table 3). These items deal with a broad range of the academic and emotional dimensions of classroom environment, but the strongest loading items are about understanding: CARE3, "My teacher really tries to understand how students feel about things", CONS2, "My teacher checks to make sure we understand what s/he is teaching us." and CLAR1, "If you don't understand something, my teacher explains it another way."

The items from the Controlling dimension load distinctly onto two separate factors. One of those factors deals with positive aspects of classroom discipline "Students in this class treat the teacher with respect." (CONT6). The other, with negative aspects: "Student behavior in this class is a problem" (CONT4). Two other items load onto the factor dealing with negative aspects. CAPT1: "This class does not keep my attention—I get bored." And CLAR3: "When s/he is teaching us, my teacher thinks we understand even when we don't." These items also deal with negative dimensions of the classroom environment.

Table 3

Rotated Factor Loadings for the Tripod Survey: Multilevel Analysis

	Within-classrooms factor					Between-classrooms factor	
Item	1	2	3	4	5	1	2
CAPT1[f]	**0.32**	0.18	**0.38**	-0.16	-0.18	**0.74**	0.24
CAPT2[f]	0.01	**0.89**	-0.01	0.05	-0.15	**0.87**	0.07
CAPT3[f]	**0.43**	**0.33**	-0.02	0.18	-0.18	**0.87**	0.07
CAPT4[f]	0.01	**0.84**	0.03	0.03	-0.02	**0.84**	0.14
CARE1[g]	**0.73**	0.05	-0.03	0.01	-0.16	**1.05**	-0.19
CARE2[g]	**0.58**	0.04	-0.09	0.08	-0.16	**1.06**	**-0.38**
CARE3[g]	**0.80**	-0.01	-0.04	0.00	-0.08	**1.04**	-0.18
CHAL1[h]	0.02	**0.68**	0.00	-0.08	0.28	**0.76**	0.14

	Within-classrooms factor					Between-classrooms factor	
Item	1	2	3	4	5	1	2
CHAL2[h]	0.28	0.19	-0.02	0.06	**0.38**	**0.58**	0.27
CHAL3[h]	**0.66**	-0.06	0.00	0.04	0.08	**0.84**	0.12
CHAL4[h]	**0.57**	0.03	0.00	0.11	0.14	**0.83**	0.19
CHAL5[h]	**0.30**	0.28	0.01	0.01	**0.33**	**0.73**	**0.31**
CHAL6[h]	**0.54**	-0.04	-0.04	0.15	0.28	**0.65**	0.28
CHAL7[h]	**0.40**	0.14	0.00	0.20	0.09	**0.68**	0.26
CHAL8[h]	**0.44**	0.11	-0.01	0.16	0.19	**0.78**	0.19
CLAR1[i]	**0.84**	-0.03	0.05	-0.08	0.02	**0.93**	0.05
CLAR2[i]	**0.69**	0.01	-0.03	0.02	-0.02	**0.97**	0.01
CLAR3[i]	0.29	0.02	**0.36**	-0.28	-0.03	**0.71**	**0.30**
CLAR4[i]	**0.57**	0.13	0.02	0.18	-0.02	**0.82**	0.24
CLAR5[i]	**0.67**	0.03	0.04	0.04	-0.07	**0.91**	0.10
CONF1[j]	0.02	**0.60**	0.03	-0.04	0.27	**0.69**	0.18
CONF2[j]	0.13	0.12	-0.23	**0.32**	-0.15	**0.91**	**-0.34**
CONF3[j]	**0.47**	0.09	-0.03	0.18	0.14	**0.94**	0.02
CONF4[j]	0.24	0.23	0.03	0.13	0.22	**0.81**	0.11
CONF5[j]	**0.66**	0.04	0.04	0.01	0.05	**0.88**	0.06
CONS1[k]	**0.55**	0.04	-0.06	0.16	0.03	**0.96**	-0.1
CONS2[k]	**0.81**	-0.03	0.05	-0.06	0.06	**0.95**	0.02
CONS3[k]	**0.51**	0.10	0.00	0.09	0.17	**0.82**	0.17
CONS4[k]	**0.70**	0.04	-0.02	0.01	0.00	**0.89**	0.09
CONT1[l]	-0.05	0.19	0.14	**0.34**	0.16	0.28	**0.78**
CONT2[l]	-0.07	-0.02	**0.67**	0.06	0.02	-0.09	**1.01**
CONT3[l]	-0.01	0.04	**0.66**	0.04	-0.04	0.21	**0.79**
CONT4[l]	-0.03	-0.01	**0.80**	0.07	0.02	0.06	**0.94**
CONT5[l]	0.08	0.00	0.13	**0.67**	-0.03	**0.38**	**0.70**
CONT6[l]	0.22	-0.02	0.26	**0.43**	-0.04	**0.45**	**0.64**
CONT7[l]	0.17	0.07	0.13	**0.49**	0.07	**0.40**	**0.67**

Note: [f]Captivating, [g]Caring, [h]Challenging, [i]Clarifying, [j]Conferring, [k]Consolidating, [l]Controlling. All loadings greater than .3 are shown in bold. Strongest loadings for each item are shaded.

The between-classroom level analysis (Table 3) shows two factors—one of which is dominated by items relating to the academic and emotional support of a classroom, and one of which is dominated by items related to classroom management (Control). There is substantial

cross-loading at the between level, with 8 items loading onto both factors. The two between level factors correlate roughly .55.

When the model with the cross-level configural non-invariance was compared to a model that imposed the same factor structure across levels, the change in AIC ($\Delta_i = 307.0$) suggests that there is little evidence to support the model with cross-level invariance. Overall, it is possible to conclude that there is strong evidence showing non-invariance across levels in the Tripod Survey.

What Are the Consequences of Ignoring the Multilevel Structure and Conducting a Factor Analysis on the Disaggregated Data?

For the Working Conditions Survey, parallel analysis suggested extracting 6 factors from the total, disaggregated correlation matrix. The pattern of factor loadings, and their relative magnitude, is consistent with the within factor structure that was suggested by the multilevel factor analysis (Table 4).

For the Tripod Survey, parallel analysis suggested extracting 5 factors (Table 4). The factor structure is also similar to the within-structure in the multilevel analysis, both in terms of the pattern of loadings and their relative magnitude.

Table 4

Rotated Factor Loadings: Disaggregated Analysis

	Working Conditions Survey factor							Tripod Survey factor				
Item	1	2	3	4	5	6	Item	1	2	3	4	5
TIME1[a]	0.06	0.01	0.14	-0.04	-0.01	**0.43**	CAPT1	**0.32**	0.20	**0.4**	-0.18	-0.20
TIME2[a]	-0.06	-0.03	0.01	-0.01	0.14	**0.70**	CAPT2	0.01	**0.9**	0.00	0.05	-0.15
TIME3[a]	0.11	0.02	0.07	0.02	-0.01	**0.61**	CAPT3	**0.43**	**0.37**	0.00	0.17	-0.21
TIME4[a]	0.21	0.11	0.02	0.05	-0.07	**0.51**	CAPT4	0.01	**0.84**	0.04	0.03	0.00
TIME5[a]	-0.02	0.02	-0.01	0.00	0.03	**0.76**	CARE1	**0.77**	0.05	-0.03	0.00	-0.14
FACR1[b]	0.02	0.08	**0.55**	0.02	0.08	0.03	CARE2	**0.63**	0.04	-0.12	0.05	-0.16
FACR2[b]	-0.01	-0.01	**0.76**	-0.01	0.06	-0.04	CARE3	**0.83**	-0.02	-0.03	-0.01	-0.08
FACR3[b]	-0.02	0.02	**0.74**	0.02	0.00	-0.01	CHAL1	0.04	**0.67**	0.00	-0.08	**0.32**
FACR4[b]	0.05	0.08	**0.55**	0.02	0.00	0.07	CHAL2	**0.32**	0.18	-0.02	0.07	**0.4**
FACR5[b]	-0.02	-0.01	**0.62**	0.02	0.02	0.01	CHAL3	**0.69**	-0.07	0.01	0.05	0.09
FACR6[b]	0.00	-0.01	**0.53**	0.03	0.03	0.14	CHAL4	**0.62**	0.03	0.02	0.11	0.13
FACR7[b]	0.20	-0.04	**0.42**	0.03	-0.02	0.05	CHAL5	**0.33**	0.29	0.03	0.02	**0.33**
FACR8[b]	0.33	-0.01	**0.39**	0.04	-0.04	0.05	CHAL6	**0.6**	-0.07	-0.03	0.14	0.29

	Working Conditions Survey factor							Tripod Survey factor				
Item	1	2	3	4	5	6	Item	1	2	3	4	5
LEAD1[c]	**0.48**	0.13	0.02	0.18	0.03	0.08	CHAL7	**0.43**	0.16	0.01	0.19	0.08
LEAD2[c]	**0.63**	0.03	0.02	0.15	0.07	0.01	CHAL8	**0.49**	0.12	0.00	0.13	0.18
LEAD3[c]	**0.89**	-0.02	0.03	-0.05	0.02	0.01	CLAR1	**0.85**	-0.02	0.07	-0.07	0.00
LEAD4[c]	**0.88**	0.00	0.03	-0.02	-0.02	0.02	CLAR2	**0.72**	0.01	-0.02	0.03	-0.03
LEAD5[c]	**0.67**	0.07	-0.01	0.16	0.02	0.06	CLAR3	**0.30**	0.04	**0.4**	-0.28	-0.03
LEAD6[c]	**0.44**	0.10	0.02	0.20	0.14	0.01	CLAR4	**0.58**	0.15	0.05	0.18	-0.04
LEAD7[c]	-0.01	-0.01	0.01	**0.93**	-0.01	0.00	CLAR5	**0.69**	0.04	0.04	0.05	-0.07
LEAD8[c]	-0.01	0.00	0.01	**0.93**	-0.01	0.00	CONF1	0.05	**0.58**	0.04	-0.02	0.29
LEAD9[c]	0.09	0.00	-0.01	**0.75**	0.09	0.00	CONF2	0.20	0.15	-0.25	**0.31**	-0.19
LEAD10[c]	**0.44**	0.18	-0.03	0.17	0.17	0.06	CONF3	**0.55**	0.1	-0.02	0.14	0.11
DECM1[d]	-0.05	**0.65**	0.11	0.04	-0.04	0.03	CONF4	0.29	0.25	0.03	0.12	0.18
DECM2[d]	-0.05	**0.6**	0.07	0.09	-0.13	0.12	CONF5	**0.70**	0.03	0.05	-0.01	0.05
DECM3[d]	-0.08	**0.55**	0.03	0.04	-0.09	0.13	CONS1	**0.61**	0.04	-0.08	0.13	0.01
DECM4[d]	0.03	**0.59**	-0.03	-0.02	0.19	0.01	CONS2	**0.84**	-0.03	0.06	-0.06	0.05
DECM5[d]	0.02	**0.57**	-0.04	-0.05	0.05	-0.02	CONS3	**0.56**	0.11	0.01	0.07	0.15
DECM6[d]	0.24	**0.52**	0.00	-0.04	0.05	-0.03	CONS4	**0.72**	0.04	0.00	0.02	0.01
DECM7[d]	0.04	**0.61**	-0.01	-0.04	0.09	-0.04	CONT1	0.04	0.21	0.19	**0.39**	0.14
DECM8[d]	0.09	**0.58**	0.02	0.04	0.08	-0.04	CONT2	0.09	-0.01	**0.73**	0.07	0.02
PROF1[e]	-0.01	0.06	0.11	-0.02	**0.61**	-0.02	CONT3	0.01	0.04	**0.7**	0.06	-0.05
PROF2[e]	0.03	0.04	-0.01	0.05	**0.61**	0.12	CONT4	0.02	-0.01	**0.84**	0.07	0.03
PROF3[e]	-0.01	-0.01	-0.02	0.02	**0.79**	0.09	CONT5	0.10	0.02	0.17	**0.68**	-0.03
PROF4[e]	0.01	0.01	0.16	0.01	**0.61**	0.00	CONT6	0.23	0.01	0.300	**0.46**	-0.05
PROF5[e]	0.09	0.05	0.04	0.07	**0.63**	-0.02	CONT7	0.19	0.08	0.18	**0.5**	0.06

Note: [a]Time, [b]Facilities & Resources, [c]School Leadership, [d]Decision Making, [e]Professional Development, [f]Captivating, [g]Caring, [h]Challenging, [i]Clarifying, [j]Conferring, [k]Consolidating, [l]Controlling. All loadings greater than .3 are shown in bold. Strongest loadings for each item are shaded.

Importantly, in both the Working Conditions Survey and the Tripod Survey, analysis of S_T results in a factorial structure that is inconsistent with either the between-classroom level or the between-school level of the corresponding multilevel analysis. This is consistent with other findings (e.g., Hox, 2010; Holfve-Sabel & Gustaffsen, 2005; D'Haenens, van Damme, & Onghena, 2010a; Reise, Ventura, Nuechterlein, & Kim, 2005) and provides a clear illustration of the methodological consequences of assuming cross-level invariance (e.g., Julian, 2001; Marsh et al., 2012).

What Are the Consequences of Ignoring the Multilevel Structure and Conducting a Factor Analysis on the Aggregated Data (Group Means)?

For the Working Conditions Survey, parallel analysis suggested 5 factors. The patterns of association (Table 5) between the items are different than at the between level of the multilevel analysis (Table 2). In particular, Leadership factor is similar to that of the within-level of the multilevel analysis, and the Professional Development items now associate separately with the Time items and the Facilities and Resources items. Additionally, the Decision Making items still largely form a distinct group. Overall, there is less cross-loading than in the between level of the multilevel analysis. The analysis of the unweighted group means distorts the factor structure and leads to a false sense of factorial structure.

Table 5

Rotated Factor Loadings: Group-Means Analysis

	Working Conditions Survey factor						Tripod Survey factor	
Item	1	2	3	4	5	Item	1	2
TIME1[a]	0.21	0.14	-0.15	0.19	**0.35**	CAPT1[f]	**0.41**	**0.45**
TIME2[a]	0.01	0.06	0	0.04	**0.76**	CAPT2[f]	**0.78**	0.08
TIME3[a]	0.29	0.09	0.06	0.1	**0.49**	CAPT3[f]	**0.76**	0.16
TIME4[a]	**0.35**	-0.1	0.05	**0.34**	**0.39**	CAPT4[f]	**0.76**	0.10
TIME5[a]	0.05	-0.12	-0.04	0.29	**0.73**	CARE1[g]	**0.97**	-0.17
FACR1[b]	0.05	**0.7**	0.12	0.07	-0.01	CARE2[g]	**0.88**	-0.3
FACR2[b]	0.03	**0.82**	-0.02	0.03	-0.01	CARE3[g]	**0.93**	-0.12
FACR3[b]	0.03	**0.72**	-0.06	0.17	-0.03	CHAL1[h]	**0.66**	0.03
FACR4[b]	0.04	**0.46**	0.06	0.35	-0.01	CHAL2[h]	**0.59**	0.10
FACR5[b]	0.01	**0.66**	-0.01	0.01	-0.01	CHAL3[h]	**0.81**	0.04
FACR6[b]	0.09	**0.53**	-0.02	-0.02	0.14	CHAL4[h]	**0.61**	0.26
FACR7[b]	0.27	**0.41**	-0.01	-0.05	0.01	CHAL5[h]	**0.72**	0.16
FACR8[b]	**0.5**	**0.42**	-0.02	-0.01	-0.07	CHAL6[h]	**0.74**	0.06
LEAD1[c]	**0.52**	-0.02	0.29	0.21	0.04	CHAL7[h]	**0.75**	0.07
LEAD2[c]	**0.73**	0.02	0.24	-0.04	0.04	CHAL8[h]	**0.77**	0.05
LEAD3[c]	**1**	0.02	-0.03	-0.09	0.01	CLAR1[i]	**0.92**	-0.02
LEAD4[c]	**0.97**	-0.02	-0.04	0.03	0.02	CLAR2[i]	**0.92**	-0.06
LEAD5[c]	**0.74**	-0.05	0.21	0.12	0.04	CLAR3[i]	0.27	**0.42**
LEAD6[c]	**0.56**	0.13	0.28	0.06	0.04	CLAR4[i]	**0.85**	0.13
LEAD7[c]	0	-0.02	**0.98**	0.03	-0.02	CLAR5[i]	**0.9**	-0.01

	Working Conditions Survey factor						Tripod Survey factor	
Item	1	2	3	4	5	Item	1	2
LEAD8 [c]	0.02	0	**0.96**	0.03	-0.02	CONF1[j]	**0.47**	0.29
LEAD9 [c]	0.12	0.02	**0.83**	-0.01	0.05	CONF2[j]	**0.63**	-0.19
LEAD10 [c]	**0.51**	0.05	**0.32**	0.14	0.11	CONF3[j]	**0.84**	-0.01
DECM1 [d]	0.01	0.21	0.09	**0.74**	-0.03	CONF4[j]	**0.7**	0.13
DECM2 [d]	0.02	-0.02	0.08	**0.82**	0.04	CONF5[j]	**0.78**	0.11
DECM3 [d]	-0.04	-0.11	-0.04	**0.74**	0.21	CONS1[k]	**0.88**	-0.13
DECM4 [d]	0.13	0.22	0.12	**0.45**	0.13	CONS2[k]	**0.92**	-0.04
DECM5 [d]	0.13	0.2	-0.02	**0.42**	-0.08	CONS3[k]	**0.58**	0.26
DECM6 [d]	**0.54**	0.2	0	0.22	-0.03	CONS4[k]	**0.85**	0.00
DECM7 [d]	0.14	0.29	0.12	**0.4**	-0.09	CONT1[l]	**0.44**	**0.5**
DECM8 [d]	0.21	0.25	0.23	**0.3**	-0.08	CONT2[l]	0.14	**0.93**
PROF1 [e]	-0.07	**0.54**	0.1	0	**0.31**	CONT3[l]	0.14	**0.79**
PROF2 [e]	0.07	**0.34**	0.21	-0.05	**0.49**	CONT4[l]	0.01	**0.94**
PROF3 [e]	0.02	**0.4**	0.2	-0.11	**0.52**	CONT5[l]	**0.58**	**0.44**
PROF4 [e]	0.01	**0.52**	0.15	-0.01	**0.36**	CONT6[l]	**0.50**	**0.53**
PROF5 [e]	0.15	**0.43**	0.22	-0.09	**0.36**	CONT7[l]	**0.53**	**0.46**

Note: [a]Time, [b]Facilities & Resources, [c]School Leadership, [d]Decision Making, [e]Professional Development, [f]Captivating, [g]Caring, [h]Challenging, [i]Clarifying, [j]Conferring, [k]Consolidating, [l]Controlling. All loadings greater than .3 are shown in bold. Strongest loadings for each item are shaded.

For the Tripod Survey, parallel analysis suggests the extraction of 2 factors (Table 5). The structure suggested by the analysis of the group-mean correlation matrix is fairly similar to that of the between level of the multilevel analysis. There is still one large factor; however, the control items no longer load as distinctively onto a separate factor. As was the case with the WCS, there is less crossloading in the group-mean analysis than in the between level of the multilevel analysis.

In summary, in both the WCS and Tripod Surveys, analysis of the unweighted group means has the effect of distorting the perceived factorial structure, and leads to inferences that are not consistent with either the within or between levels of analysis. Again, this is consistent with theoretical results discussed elsewhere (Preacher, Zyphur, Zhang, 2010). Conceptually, this distortion makes sense. There are at least two distinct sources of bias that are present in this group means analysis. First, differences in group size are not accounted for, and this may distort the covariance matrix. Second, the covariance matrix of group-means contains both between and within sources (Muthén, 1994), and to the extent that the between and within covariance matrices

have different structures, this will also have the effect of distorting inferences about the factorial configuration.

Summary

Even as the awareness of multilevel modeling has grown, analytic methods that assume cross-level invariance are still widely used in the educational policy and research literature. It is common to find studies that use single-level factor analyses that ignore the clustered, hierarchical structure of the data, and using linear composites to create individual scores. This article used two examples to investigate whether there is empirical evidence to support the assumption of cross-level measurement invariance, and whether using factor analytic techniques that assume cross-level invariance would influence the analysis of empirical data. The results reflect some general patterns that are worth noting here.

There can be Significant Differences in Factorial Structure Across Levels

In these two empirical examples, fewer factors were found at the between-groups level than the within-group level of analysis. In the case of the Working Conditions Survey, the multilevel analysis suggested 6 within-school factors, and 5 between-school factors. In the case of the Tripod Survey, the differences in factorial structure are even greater. While there is support for 5 factors at the within-classroom level, there is only support for 2 factors at the between-classroom level.

This exploratory analysis may, as Cronbach (1976) suggested, lead to the articulation of a specific (and independent) theory for constructs that exist and are distinguishable between groups (school or classroom). For example, in the Working Conditions Survey, there are five dimensions of school climate that are distinguishable based on aggregated survey responses. For the Tripod Survey, there are two dimensions of classroom environment that are distinguishable based on aggregated survey responses. This is consistent with other factor analyses conducted on the Tripod data, which found that the items from the "Five Support C's" (Conferring, Consolidating, Captivating, Caring, Clarifying) and Challenge load onto one factor as an "amorphous group" (Ferguson, 2010, p. 6).

Analysis of the Total covariance Matrix can Distort Perception of the Between-level Factorial Structure

The results of the factor analyses on the total covariance matrix did not predictably show concordance with the between-level structure for either survey. In both cases, the structure that was identified bore a strong resemblance to the within-structure identified in the multilevel analysis. Since there were fewer factors identified at the between-level, this can lead to an

individualistic fallacy (Alker, 1969), where phenomenon that occur between individuals are assumed to occur between groups.

Analysis of the Group Mean Covariance Matrix can Distort Perception of Both the Between-level and Within-level Factorial Structures

The factor analysis on the unweighted group mean covariance matrix yielded results that were not consistent in factorial structure with any of the other analyses. While this analysis did suggest five factors for the Working Conditions Survey, the patterns of loadings were different than in either the disaggregated analysis or the multilevel analysis. In the case of Tripod, two factors were identified, but again, the patterns of association were not consistent with the between-level of the multilevel analysis.

The results have direct implications and raise important questions for applied research and policy. Factor analysis is commonly used for rank reduction. Based on the results of a factor analysis, linear composites are created that act as proxies for factors and that may be interpreted directly or included in a range of predictive or inferential statistical analyses. In this kind of analysis, depending on which covariance matrix was analyzed, there may be evidence for completely different linear composites. These composites differ not only in the number of included items, but in the way they would be defined and articulated. This means that, depending on which factor analysis was conducted, different qualities of school or classroom environment would be defined, and entirely different sets of relationships would be explored.

Taking a scenario posed by Ladd (2011) to show these consequences in a practical way, imagine that interventions to improve teacher retention are based on relationships between dimensions the school environment and teacher mobility. Would the same set of policy recommendations be derived if we identified Time and Professional Development as two differential predictors of teacher mobility than if you identified them as a single construct? What about the situation where you identified Decision Making as a single predictor, rather than as two differentially predictive dimensions of school environment?

Improperly constructed linear composites make appropriate theory testing difficult if not impossible, with important implications that are not only theoretical but also eminently practical. If an intervention targeted at improving retention is found not to have the desired effects, for example, it would be impossible to disentangle theory failure (i.e., the intervention is ill-conceived and will never improve retention because it is based on a faulty model of teacher mobility) from implementation failure (i.e., the theory is sound, but the intervention was implemented poorly. Had the intervention been implemented correctly, teacher retention would

have improved, and so policy should address proper implementation; Raudenbush & Sadoff, 2008).

Other Issues and Additional Questions

This study described how assumptions of cross-level measurement invariance can lead to the identification of factorial structures that are inconsistent with those found when conducting a multilevel analysis. There are, however, several limitations of this study, and these present areas for future research and additional questions.

Assumption of Reflective Aggregation

The assumption that variance between teachers or students within a school or a classroom is attributable to error is a complex issue in organizational climate research. Sirotnik's (1980) affective-descriptive continuum attempts to delineate items that are intended to measure individual, psychological constructs from items that are intended to measure organizational constructs. Affective items typically have an "I-form" structure and ask about a psychological construct. Descriptive items are typically have a "they-form" structure and position individuals as raters of a single organizational quality. Many items fall in the middle of the continuum. This raises important questions about what is being measured. Are the items measuring qualities of the classroom or school, qualities of the teachers or students, or something else?

The two surveys used in this article are based primarily on "they-form" items from the descriptive end of the continuum. However, that does not mean they do not reflect a certain amount of true variation in the psychological standing of the respondents. If school climate is to be measured by aggregating lower level responses to questions about climate, attention should be paid to whether items refer primarily to the psychological characteristics of individuals, or primarily to characteristics of organizations.

Use of Linear Composites as Proxies for Latent Variables

This study focused on whether or not there was empirical evidence to support the assumption of cross-level measurement invariance, and how perceptions of measurement structure may be distorted by conducting single level analysis. Because the two empirical examples in this study showed evidence for different structures at the within-group and between-groups levels, other types of consequences that can arise by using linear composites as proxies for individual latent variables were not explored. Even in the case where there is strong evidence for cross-level invariance, there have been many recent studies (Lüdtke, Marsh, Robitzsch, & Trautwein, 2011; Marsh et al., 2009; Preacher et al., 2010) that show that using linear composites as proxies for latent variables in multilevel analyses can give biased estimates of regression

parameters. In other words, if it is assumed that individual indicators are fallible indicators of an individual's latent standing, and if it is believed that the observed group mean contains sampling error, then using linear composites may introduce bias into the estimation of relationships with external variables. The extent of this bias will be a function of the amount of measurement error and sampling error. Preacher et al. (2010) provides some mathematical insight into how to quantify this bias.

Omitted Levels in the Analysis

This analysis, for the sake of simplicity, focused only on a two-level measurement model. With both the student surveys of classroom climate and the teacher surveys of schools, only one level of nesting is accounted for. For the Tripod, the assumption was that students were nested within classrooms. For the WCS, the assumption was that teachers were nested within schools. However, it is possible to imagine nested facets that are excluded here. For example, students nested within classrooms nested within teachers (nested within schools). Or teachers nested within departments nested within schools. Recent work by Wei and Haertel (2011) suggests that omitted levels of variance can bias the estimation of variance components.

Factor Extraction and Model Fit

Since the determination of whether or not cross-level invariance is supported empirically may boil down to a comparison of two or more possible measurement models, understanding the limitations on determining the appropriate number of factors to retain in a multilevel context is particularly important. In order to ensure an objective and consistent criterion, this analysis used parallel analysis, and then extracted (and rotated) factors separately on the within and between levels. There are several other possible approaches to MEFA, including simultaneously fitting an exploratory model at the two levels of analysis. Additionally, decisions about how many factors to extract could be made based on other criteria, including scree plots (Cattell, 1966), or the Kaiser criterion (Kaiser, 1960). The use of Maximum Likelihood factor methods would allow for the use of likelihood ratio test statistics and fit indices to determine how many factors to retain. There has been limited simulation work (Yuan & Bentler, 2007; Ryu & West, 2009; Ryu, 2011; Hox, Maas, & Brinkhuis, 2010) on the performance of test statistics and popularly reported fit indices in empirical data, and most of this work is focused on confirmatory analyses. What's more, there are competing approaches to testing and assessing model fit in multilevel models (see Ryu & West, 2009 for a description of three of the most widely used methods), and which approach is "best" is an open issue. Whether the use of these other approaches to factor retention would yield decisions consistent with those found in this study is an issue that can be explored. There are also other paradigms for model comparison. For example, Lee and Song (2001)

approach model selection by using Bayes factors. These other methodologies were not explored here because they are relatively rare in applied literature. However, their utility in exploring cross-level measurement invariance is worth investigating further.

References

Alker, H. R. (1969). A typology of ecological fallacies. *Quantitative ecological analysis in the social sciences,* 69-86.

Alwin, D. F. (1973). The use of factor analysis in the construction of linear composites in social research. *Sociological Methods and Research, 2*:191-214.

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

Bill & Melinda Gates Foundation. (2010). *Teachers' perceptions and the MET project.* Retrieved from http://www.metproject.org/downloads/Teacher_Perceptions_092110.pdf

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analyses. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 349–381). San Francisco: Jossey-Bass.

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

Bollen, K. A., & Bauldry, S. (2011). Three Cs in measurement models: Causal indicators, composite indicators, and covariates. *Psychological Methods, 16,* 265–284.

Bollen, K. A., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin, 110,* 305–314.

Briggs, N. E., & MacCallum, R. C. (2003). Recovery of weak common factors by maximum likelihood and ordinary least squares estimation. *Multivariate Behavioral Research*, *38*(1), 25-56.

Browne, M. W., MacCallum, R. C., Kim, C. T., Andersen, B. L., & Glaser, R. (2002). When fit indices and residuals are incompatible. *Psychological methods*, *7*(4), 403.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and inference: a practical information-theoretic approach, 2nd edition*. New York: Springer.

Butrymowicz, S. A. (2012, May 13). Student surveys for children as young as 5 years old may help rate teachers. *The Washington Post*. Retrieved from http://www.washingtonpost.com/local/education/student-surveys-may-help-rate-teachers/2012/05/11/gIQAN78uMU_story.html

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral. Research, 1*, 245–276.

Chan, D. (1998). Functional relations among constructs in the same content domain at different levels of analysis: A typology of composition models. *Journal of Applied Psychology, 83,* 234-246.

Crawford, C. B., & Koopman, P. (1973). A note on Horn's test for the number of factors in factor analysis. *Multivariate Behavioral Research*, *8*(1), 117-125.

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

Cronbach, L. J., & 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.

D'haenens, E., Van Damme, J., & Onghena, P. (2010a). Multilevel exploratory factor analysis: Illustrating its surplus value in educational effectiveness research. *School Effectiveness and School Improvement, 21*(2), 209-235.

D'haenens, E., Van Damme, J., & Onghena, P. (2010b). Linking student outcome variables with school process variables: multilevel confirmatory factor analysis takes precedence. *Effective Education*, *2*(2), 117-142.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4,* 272–299.

Ferguson, R. (2010, October 14). *Student perceptions of teaching effectiveness. Retrieved from* http://www.gse.harvard.edu/ncte/news/Using_Student_Perceptions_Ferguson.pdf

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment, 7,* 286-299.

Ford, J. C., McCallum R. C., & Tait, M. (1986). The application of exploratory factor analysis in applied psychology: a critical review and analysis. *Personnel Psychology, 39*: 291–314.

Glick, W. H. (1985). Conceptualizing and measuring organizational and psychological climate: Pitfalls in multilevel research. *Academy of Management Review, 10,* 601-616.

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

Harnqvist, K. (1978). Primary mental abilities at collective and individual levels. *Journal of Educational Psychology, 70,* 706–716.

Holfve-Sabel, M., & Gustaffsson, J. (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.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrica, 30*: 179-185

Hox, J. J. (2010). *Multilevel analysis. Techniques and applications*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hox, J. J., Maas, C. J. M., & Brinkhuis, M. J. S. (2010). The effect of estimation method and sample size in multilevel structural equation modeling. *Statistica Neerlandica*, *64*, 157-170.

Hoy, W. K., & Clover, S. I. R. (1986). Elementary school climate: A revision of the OCDQ. *Educational Administration Quarterly*, *22*(1), 93-110.

Hubbard, R., & Allen, S. J. (1987). An empirical comparison of alternative methods for principal component extraction. *Journal of Business Research*,*15*(2), 173-190.

Humphreys, L. G., & Montanelli Jr, R. G. (1975). An investigation of the parallel analysis criterion for determining the number of common factors. *Multivariate Behavioral Research, 10*(2), 193-205.

Ingersoll, R. (2001). Teacher turnover and teacher shortages: An organizational analysis. *American Educational Research Journal, Vol. 38*, No. 3, pp. 499-534.

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

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*, 141-151.

Klinger, D. A., Rogers, W. T., Anderson, J. O., Poth, C., & Calman, R. (2006). Contextual and school factors associated with achievement on a high-stakes examination. *Canadian Journal of Education, 29*(3), 771–797.

Kozlowski, S. W. J., & Klein, K. J. (2000). A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations.* San Francisco: Jossey-Bass. 3–90.

Ladd, H. (2011). Teachers' perceptions of their working conditions: How predictive of planned and actual teacher movement? *Educational Evaluation and Policy Analysis, 33*(2), 235-261.

Lee, S. Y., & Song, X. Y. (2001). Hypothesis testing and model comparison in two-level structural equation models. *Multivariate Behavioral Research, 36*, 639-655.

Lüdtke, O., Marsh, H. W., Robitzsch, A., & Trautwein, U. (2011). A 2×2 taxonomy of multilevel latent contextual models: Accuracy and bias trade-offs in full and partial error-correction models. *Psychological Methods, 16* (4), 444-467.

Marsh, H. W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2009). Doubly-latent models of school contextual effects: Integrating multilevel and structural equation approaches to control measurement and sampling error. *Multivariate Behavioral Research, 44*, 764–802.

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.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*, 4:525-543.

Muthén, B. (1994). Multilevel covariance structure analysis. In J. Hox, & I. Kreft (Eds.), *Multilevel Modeling*, a special issue of *Sociological Methods & Research, 22*, 376-398.

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

Muthén, B. O., & Muthén, L. K. (2010). *Mplus* (version 6.11) [computer software] Los Angeles: Muthen & Muthen.

New Teacher Center. (2009). *Validity and reliability of the North Carolina teacher working conditions survey.* Retrieved from http://www.ncteachingconditions.org/sites/default/files/attachments/validityandreliability.pdf

NYC School Surveys. (n.d.). Retrieved from http://schools.nyc.gov/Accountability/tools/survey/default.htm

Preacher, K. J., & MacCallum, R. C. (2002). Exploratory factor analysis in behavior genetics research: Factor recovery with small sample sizes. *Behavior Genetics*, *32*(2), 153-161.

Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods*, *15*, 209–233.

Rabe-Hesketh, S., Skrondal, A., & Zheng, X. (2007). Multilevel structural equation modeling. In S.-Y. Lee (Ed.), *Handbook of latent variable and related models* (pp. 209–227). Amsterdam, The Netherlands: Elsevier.

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

Raudenbush, S., & Sadoff, S. (2008). Statistical inference when classroom quality is measured with error. *Journal of Research on Educational Effectiveness, 1,* 138–154.

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

Ryan, A. M., & Patrick, H. (2001). The classroom social environment and changes in adolescents' motivation and engagement during middle school. *American Educational Research Journal*, *28*, 437–460.

Ryu, E. (2011). Effects of skewness and kurtosis on normal-theory based maximum likelihood test statistic in multilevel structural equation modeling. *Behavior Research Methods*, *43* (4), 1066–1074.

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

Schmitt, T. A. (2011). Current methodological considerations in exploratory and confirmatory factor analysis. *Journal of Psychoeducational Assessment*, *29*(4), 304-321.

Sirotnik, K. A. (1980). Psychometric implications of the unit-of- analysis problem (with examples from the measurement of organizational climates). *Journal of Educational Measurement, 17*, 245–282.

Van de Vijver, F. J., & Poortinga, Y. H. (2002). Structural equivalence in multilevel research. *Journal of Cross-Cultural Psychology*, *33*(2), 141-156.

Wei, X., & Haertel, E. (2011). The effect of ignoring classroom-level variance in estimating the generalizability of school mean scores. *Educational Measurement: Issues and Practice*. *30* (1) 13-22.

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

Zyphur, M., Kaplan, S., & Christian, M. (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*, 127–140.